Imputation of injury data of uninsured patients using Estonian Health Insurance Fund database

Tervise Arengu Instituut
National Institute for Health Development

National Institute for Health Development

Imputation of injury data of uninsured patients using Estonian Health Insurance Fund database

Viktoria Kirpu

Tallinn 2019

The **mission** of the National Institute for Health Development is to create and share knowledge for influencing the attitudes, behaviour, policies and the environment with evidence-based information with an aim of improving the well-being of the people in Estonia.

I thank my colleagues and faculty members of the University of Tartu for advice and comments.

When using the information presented in this report, refer to the publication. Recommended reference:

Kirpu V. Imputation of injury data of uninsured patients using Estonian Health Insurance Fund database. Tallinn: National Institute for Health Development; 2019.

Contents

Abstract

Abstract. The objective of this bachelor's thesis is to supplement the data sent to The Estonian Health Insurance Fund's database using Health Information System's data as additional information. To achieve the objective, the data from both databases is linked and as a result a lot of ages of the uninsured patients are found to The Estonian Health Insurance Fund's database. This variable's values are only marked in the Health Information System's data and originally missing from The Estonian Health Insurance Fund's data. For those epicrisis, where the patient's age is still missing, the required variable is imputed with three different methods: general random *Hot-Deck* method, k-nearest neighbour method and random *Hot-Deck* imputation within classes combined with the k-nearest neighbour method. As the result of the data linking, ages were found to 5633 patients and 3515 epicrisis remained without this variable's value. Based on the results of further analysis, it was decided to use the data imputed with general random *Hot-Deck* method, because in the imputation simulation this method gave the most precise and stable results.

CERCS research specialisation: P160 Statistics, operation research, programming, financial and actuarial mathematics

Keywords: data processing, statistical data processing, missing data, observation errors, imputation, *Hot-Deck* method, k-nearest neighbour method.

Summary

Estonian health care providers submit documentation about treatment cases into many different systems. The decision has been made to adopt one data submission system for the production of statistics – the Health Information System, but unfortunately the data sent there is too incomplete. Health care providers submit the most complete documentation to the Health Insurance Fund because they pay to health care providers by the data sent. That is the reason why the data of the latter have been used in this analysis.

This work aimed to supplement the data from the Estonian Health Insurance Fund's database using the data sent to the Health Information System's database as additional source of information. The main problem of the analysis was that it is not possible to obtain complete demographical statistics on the basis of the Health Insurance Fund's data since the treatment invoices of uninsured patients have missing age.

To find the patients' age to the treatment case documentation, the injury data submitted to the Health Insurance Fund's database and the Health Information System were decided to be linked. The task was made difficult by the fact that the ID code generated to the same patient within the two datasets was different and therefore this variable could not be used in the linking process. The team of analysts at the National Institute for Health Development (NIHD) decided to use the other variables present in the datasets to carry out the linking. It was discovered that health care providers submit the data of the exact same patients differently into the two databases. To nevertheless carry out the data-linking, 187 variable combinations and rules for linking were selected by the analysts at NIHD. If the data contained duplicated i.e. treatment cases too similar to be linked, the analysts at NIHD decided to not link these in this analysis. The linking processes produced an age value for 5633 patients of treatment cases in the Health Insurance Fund's database. The patient's age was not found for 3515 treatment invoices.

Because after joining process there was still missing some data, then it was decided to use imputation on the basis of the existent data. Three methods were used for this: general random Hot-Deck method, the k-nearest neighbour method and donor group based random Hot-Deck imputation combined with the k-nearest neighbour method. The k-nearest neighbour method was used with the latter method only with observations that were in such donor groups where there were no treatment cases for the patients of which ages could be found. In total there were 9 missing values in the diagnosis code of external cause of injury, 37 in the gender variable and 3515 in age variable. The values were imputed starting from the variable which included the least amount of missing values and in the imputation of each subsequent variable, the results of the previous imputation were used.

Unfortunately, the ranking of the studied methods could not be verified after the imputation. Therefore, it was decided to carry out two tests, in the first of which, 70% of existent data was retained, and in the second, 50% of the values of the age variable. Initially one imputation on missing data was carried out with the studied methods. Then it was decided to carry out two different simulations in both cases where in the first, the data was imputed at each step of the simulation on the same data, and in the second, 70% or 50% of data was selected in the beginning of each step of the simulation. In the first simulation, the quality assessment of the k-nearest neighbour method was not possible. However, the second simulation revealed that the general random Hot-Deck method gives the most stable results in the imputation of age values. For that reason this bachelor's thesis used the data imputed with this method.

In the future, the National Institute for Health Development plans to conduct a similar analysis, with the difference that the data requested from the Health Insurance Fund and the Health Information System would have ID codes generated identically for patients across the different datasets. The next analysis also aims to link the duplicates presented in the datasets which was not done in this work due to a corresponding decision by the experts.

Introduction

The task of Estonian health care providers beside providing health care is to submit treatment case documentation and reports into three different systems: A-veeb (National Institute for Health Development), Health Information System and Estonian Health Insurance Fund. In order to reduce the work load of doctors, it has been decided to adopt one system to produce statistics – using only Health Information System. Unfortunately, this system contains many flaws, one of which is that doctors submit considerably less data there than to other systems.

At the moment, doctors submit the most treatment case data into the Health Insurance Fund's database. This is why the data from this database forms the basis for this bachelor's thesis. Only data of registrated injuries has been included in this thesis. Unfortunately, it is not possible to obtain complete demographical statistics on the basis of this data alone since the epicrises of uninsured patients have missing age variable. This work aims to supplement the data from the Estonian Health Insurance Fund's database using Health Information System's database as additional source of information.

For high-quality statistics, it is important to have as much available information as possible. For this purpose, the databases of the Health Insurance Fund and Health Information System were decided to be linked in order to obtain the ages of many uninsured patients to the epicrises. The treatment cases for which the patient's age was not revealed by the process of linking the databases are imputed either with the Hot-Deck method or k-nearest neighbour method.

This work is divided into four chapters. The first chapter introduces the process of imputation and used methods in more detail. In the second chapter, there are described the peculiarities of the data submission systems that were used. The third, fourth and fifth chapter constitute the practical part of the work, where the data-linking and imputation process are described. The third chapter provides all the various combinations used in linking the data. The fourth chapter includes the missing values for the imputed data used in the analysis. In the fifth chapter, imputation tests have been carried out, on the basis of which the best imputation method is chosen.

To conduct the practical part of the work, the application software STATA is used for data-linking and the software R is used for imputation.

1 Imputation

The problem for extensive studies is often incomplete data (1). Data with missing values arise when the subject in the sample does not provide complete answers to the questionnaire, i.e. **unit nonresponse**, or the questionnaire is partially left unanswered, i.e. **item nonresponse** (1, 7). Usually, to adjust for the loss on the level of the unit, weighting methods are used that assume the possibility to use background information (registers, previous similar surveys, etc.) (7). However, the method used most commonly to adjust for the loss on the level of the item is **imputation**. By using this method, missing values are given estimates to complete the data as the end result that can be analysed with traditional analysis methods. (1)

Sample surveys are usually conducted to find descriptive characteristics of the population, for example averages, correlations and regression coefficients. However, the values of certain objects are not of primary importance in the data. In short, the goal of imputation is not so much about getting the best prognoses for the missing values, but replacing them with sufficiently reliable values so that the complete data acquired would be as good as possible to find the best estimates for descriptive characteristics of the population. (1)

1.1 The importance of imputation

Lack of data does not only cause a loss of the necessary information and a reduction of the capacity of the study but it also causes biased estimates. It is important to minimise the volume of undiscovered lost observations i.e. number of treatment cases for which documentation was not submitted to the Health Information System and the lack of which was not discovered during checking. Otherwise, statistical conclusions, for example the confidence interval, are probably faulty. Not taking into account the loss will increase the bias of the estimates. It is necessary to have an unbiased estimate for high-quality statistics, or the bias should as little as possible. The smaller the bias, the better the statistical results reflect the actual situation. (5)

For example, the emergency type of a treatment case is largely not submitted to the Health Information System by the doctors. If a situation arises where upon emergency treatment cases doctors fail to note down the observable variable, then this gives the impression that there are few emergency treatment cases in our country. In such a situation we can be certain that the obtained statistics do not describe the reality and we have received biased estimates. On the same principle, biases also arise when some epicrises have not even been documented. (5)

There is a prevalent false understanding that if the rate of response is high, it is not important to take into account the loss of data. For example in sample surveys, the most active responders are older people and the rate of nonresponse is higher among younger people. By increasing the number of responders, most likely more older people will be drawn to the sample and this can lead to a biased estimates that is not descriptive of the general population. Therefore, in producing statistics, it is not suitable to focus on the response rate as an indicator that decreases the bias caused by loss. Unlike the dispersion of the estimate, the bias may not approach zero as the sample size increases. In order to reduce the bias caused by loss it is important to use the appropriate assessment methods. (5)

1.2 Donor-based imputation methods

Under investigation in this bachelor's thesis are widely used donor-based imputation methods, in which the missing values are replaced by a donor group's real values that have been obtained as the value of some other object. The advantage of this method is that the imputed value is possible in reality too. (9)

1.3 Hot-Deck imputation method

The Hot-Deck method is a very popular donor-based imputation method in which each missing value is substituted by an existing value of an object similar to the given object (1). With Hot-Deck imputation procedure, we signify the imputed value of k $\hat{y}_k = y_{l(k)}$, where $l(k)$ is a randomly selected donor from all donor elements $l \in r_i$, where r_i signifies the set of all possible donor elements. The method also has its disadvantage: although on visual inspection the distribution of the imputed variable seems to be rather natural, imputation bias may occur since responded objects may considerably differ from non-responded objects. (9).

1.4 Examples of Hot-Deck imputation methods

Generally the following Hot-Deck methods are distinguished.

1. *Random Hot-Deck imputation within class* is an imputation method where first donor groups are formed from data on the basis of an auxiliary variable, after which the missing variable value is replaced by an existing value taken from a corresponding donor group. The selection from the donor group is often random. Only such variables of the register are suitable to be the additional variable whose values are known for all objects of the sample (i.e. gender, residence, age group, etc.). (8)
2. In case of the *general random Hot-Deck imputation* method, the missing value is imputed with a value of an object selected randomly from all responders (8). In this method, objects are not divided into groups, and the result is more robustious compared to the previous case.
3. *Sequential Hot-Deck imputation* is an imputation method where all objects of the sample are sequenced according to a background variable. The missing value is imputed with the value of an object within the same class and previous in sequence. (8) Unlike the 1st and 2nd method, this is a deterministic[1], non-random imputation method.

1.5 Creating a donor group

For the abovementioned first method, non-coinciding imputation groups i.e. donor pools have to be created first (1, 9). These imputation groups are formed with the help of auxiliary variables the value of which is known for all objects in the sample (1). The same imputation method is often used to find missing values within each group, but exceptions can occur as well (9).

[1] deterministic – following the objective causal conditionality of all events and phenomenas (4)

Imputation is done in different groups, mainly for two reasons. First, within different subgroups of a single sample, the connections might be different and therefore, the imputation variable suitable in one group may not be that in the other. Determining suitable groups requires good ability to assess the situation and know the subject matter. (9)

The second reason is that the same auxiliary information is not always known for all variables. The variables required for a specific imputation method may not be known for the entire s sample. For example, let's suppose that there is a strongly related imputation vector $\boldsymbol{x}$, but only for one group of the sample. In this case, the regression or k-nearest neighbour method can be used with this subgroup. To impute the rest of the group, inferior imputation vectors have to be used. Upon insufficient auxiliary information, imputation with averages of responders or Hot-Deck procedure may also be used as an aid. (9)

1.6 The k-nearest neighbour imputation

The aim of the k-nearest neighbour imputation method is to find the most related object to the imputed value to reduce the error that may occur. The idea is that it is being assumed that two objects with similar x-values have also similar y-values. The donor element m is found with the distance minimisation method. (9)

Continuous variables

For continuous variables, the absolute distance is divided with the length of the entire observable range:

$$d_{i,j,m} = \frac{|x_{i,m} - x_{j,m}|}{r_m},$$

where $x_{i,m}$ is the value of m-th variable upon i-th observation and r_m is the range of m-th variable (6).

Discrete variables

Sequential variables are changed into numerical variables and then, the entire distance is divided by the length of the range calculated. Nominal variables are, however, treated as they are at the same distance. (6)

For nominal and binary variables, the simple $0/1$ distance is used:

$$d_{i,j,m} = \begin{cases} 0, \ if \ x_{i,m} = x_{j,m} \\ 1, if \ x_{i,m} \neq x_{j,m} \end{cases}. \text{(6)}$$

1.7 The advantages and disadvantages of Hot-Deck methods

Regardless of the wide use of Hot-Deck methods in practice, there are no clear theoretical results about them (1). Today, many imputation methods have been developed that are well studied from the theoretical aspect as well. Still, a big advantage of Hot-Deck methods is their simplicity and speed, which is why this method is used with particularly large datasets. Imputation with the k-nearest neighbour method takes considerably more time because each observation requires the calculation of the distance from missing values to find the k closest neighbours. (6)

2 The peculiarities of the databases connected to the Estonian health care services

At present, the health care providers have the obligation to submit documentation of treatment cases separately to different systems. The data of the same treatment cases are duplicated, however, this is not needed for producing statistics. One of the priorities of the National Institute for Health Development (NIHD) is to reduce the documentation work load of the health care providers. The Health Information System i.e. e-Health is regarded as one potential source of statistical data. This not only allows to reduce the work load of the health care providers but also submit more diverse and detailed statistics to consumers and increase the quality of health statistics. For this purpose, NIHD regularly evaluates the quality of the data in the Health Information System. (5)

Data is submitted to the Health Insurance Fund on a treatment case basis, however, in this database, documentation may be submitted multiple times per one treatment case (see Annex 3.a). In the Health Information System, medical care is documented on a treatment case basis as well, although that system contains much less epicrises (see Annex 3.b). Unlike the Health Information System, the Health Insurance Fund pays money to the health care providers for providing health care by the submitted data. Consequently, much more data has been submitted to the Health Insurance Fund's database than to the Health Information System, and this gives reasons to suspect that a considerable amount of data has not been submitted to the latter. If, however, data has been left undocumented, it is important to take this into account in the production of statistics and to use corresponding statistical measures. Otherwise, conclusions are drawn from incomplete data that do not correspond to the real situation. (5) Therefore, before adopting one system, it needs to be ascertained that all necessary data is submitted there.

Since more data is submitted to the Health Insurance Fund's database, it was chosen for this analysis and the production of statistics is justified on the basis of the data sent there. However, the data submitted there is not complete either. Therefore, NIHD wishes to improve the data sent to the Health Insurance Fund by using the Health Information System's data.

2.1 Database of the Estonian Health Insurance Fund

The most important task of the public agency, the Estonian Health Insurance Fund, is to enable health insurance services for persons insured by the Health Insurance Fund. In addition, it is the task of the institution to also assist with preparing standards and guidelines for treatment, motivate health care providers to improve the quality of health services, organise the performance of international agreements concerning health insurance and the Health Insurance Fund; participate in health care planning; provide opinions about legislations and international agreements related to the Health Insurance Fund and health insurance, and give advice on matters related to health insurance. In addition to this, the Health Insurance Fund collects documentation about treatment invoices of treatment cases from health care providers to have an overview of health insurance (see Annex 3.a). (5)

In this bachelor's thesis there are used the following variables from the Health Insurance Fund's data:

- *ttocode_HK* – commercial registry code of the health care provider (e.g. 90003434);
- *ID_HK* – a unique code assigned to each person (not personal identification number) (e.g. 10734533);

- *sex_HK* – the person's gender (male/female);
- *age_HK* – the age of the person at the moment of provision of health care service (e.g. 34, 81);
- *county_HK* – the county of residence of the person at the moment of provision of health care service (e.g. "Järvamaa", "Viljandimaa", "foreign country");
- *dgn_HK* – main and secondary diagnosis on the treatment invoice according to ICD-10 medical classification (e.g. "S00.01");
- *causecode_HK* – external cause diagnosis on the treatment invoice according to ICD-10 medical classification (e.g. "W00.01");
- *begin_HK* – start date of the treatment invoice (dd/mm/yyyy);
- *end_HK* – end date of the treatment invoice (dd/mm/yyyy);
- *sum_HK* – sum of treatment invoice in euros (e.g. 215 or 2125);
- *emergency_HK* – assistance from emergency medical care department (yes/no);
- *type_HK* – type of health care service (outpatient/inpatient);
- *inevitable_HK* – emergency care (yes/no);
- *insurance_HK* – existence of health insurance for the patient (yes/no).

In the Health Insurance Fund, the patient's age is generated based on the personal identification number. However, age is left uncalculated for all patients without health insurance. The reason is that the Health Insurance Fund automatically generates ages for patients from the personal identification number and as most uninsured persons are foreigners, the age is left uncalculated due to the peculiarity of their personal identification number. To produce high-quality statistics, it is important to know as much information about patients as possible.

The Health Insurance Fund's database had a total of 294,744 treatment invoices among injury related data for the year 2016.

2.2 Health Information System i.e. e-Health

The Health information system (TIS) i.e. e-Health managed and developed by the Health and Welfare Information Systems Centre (TEHIK) was created in 2008. It is a health sector cooperation model that incorporates various services, an important part of which is a database that is part of the state information system. Health care providers joined with the system send there the summaries of treatment documentation (or epicrises) and other medical documents in order to exchange information. The Health Information System processes healthcare-related data, inter alia, to keep registries reflecting medical condition and to produce health statistics. The controller of the Health Information System is the Ministry of Social Affairs and the processor is the Health and Welfare Information Systems Centre (see Annex 3.b). (5).

In this bachelor's thesis there is used the following data from the Health Information System:

- *ttocode_TIS* – commercial registry code of the health care provider (e.g. 90003434);
- *docnr_TIS* – unique code distinguishing epicrises (e.g. 603114342);
- *ID_TIS* – a unique code assigned to each person (not the personal identification number) (e.g. 476756325);
- *sex_TIS* – the person's gender (male/female);
- *age_TIS* – the age of the person at the moment of provision of health care service (e.g. 34, 81);
- *county_TIS* – the county of residence of the person at the moment of provision of health care service (e.g. "Järvamaa", "Viljandimaa", "foreign country");

- *dgn_TIS* – main and secondary diagnosis on the epicrisis according to ICD-10 medical classification (e.g. "S00.01");
- *causecode1_TIS, causecode2_TIS, ...* – external causes on the epicrisis according to ICD-10 medical classification (e.g. "W00.01");
- *begin_TIS* – start date of the treatment case (dd/mm/yyyy);
- *end_TIS* – end date of the treatment case (dd/mm/yyyy);
- *type_TIS* – type of health care service (outpatient/inpatient)*;*
- *inevitable_TIS* – emergency care (yes/no);
- *insurance_TIS* – existence of health insurance for the patient (yes/no).

In the Health Information System, the age of the patient is calculated from the personal identification number or taken from data manually entered by the doctors. Therefore, all patients have an age value. This is why the data of Health Information System was decided to be used to improve the Health Insurance Fund's data.

There were186,283 injury-related treatment case documents sent to Health Information System in 2016.

2.3 ICD-10 codes

ICD-10 medical classification can be defined as a system of categorisations where diseases categorised by established criteria. The main goal of ICD-10 is to enable the systematic registration, analysis, interpretation and comparison of data about mortality and morbidity collected at different times. This classification is used to change names of diagnoses and other health problems into alphanumeric code. This enables convenient storage, search and analysis of data even at the international level. (3)

In this bachelor's thesis, only the treatment cases that are related to injuries are observed. For this, only the data of the injury treatment cases where external cause diagnosis code start with letter "V", "W", "X" or "Y" or the main diagnosis code start with letter "S" or "T" of the ICD codes were sorted.

According to ICD-10 classification, the main diagnosis codes "S00–T98" cover injuries, poisoning and certain other consequences of external causes. These are in turn divided into non-intersecting groups for imputation:

- head and body area injuries;
- hands area Injuries;
- foot area injuries;
- other unspecified area and other types of injuries or complications (see Annex 1). (3)

External cause codes "V01–Y98" cover external causes of morbidity and mortality. External cause codes starting with the letter "V" cover the treatment case data of patients specifically injured in transport accidents. External cause codes starting with the letter "W" differentiate injury treatment cases caused by physical factors (e.g. falls, electricity, etc.). "X00–Y34" defines the data for injuries caused by different natural events and other factors (e.g. fire, burns, poisoning, etc.). Under "Y35–Y98", however, are the data for injuries caused by human and other factors. The external cause diagnosis codes are grouped based on the descriptions above (see Annex 2). (3)

3 Linking data

The first step of the bachelor's thesis was to link the documentation of treatment cases sent to the data submission systems of the Health Insurance Fund and Health Information System. The main problem turned out to be that the patients' ID codes differed between the databases. This was caused by a different patient ID code generation algorithm used in both databases. As the National Institute for Health Development is not the producer of official national statistics, they do not have the right to get the data of patients with the same ID codes because according to the principles of Estonian Data Protection Inspectorate, this is considered as a data leak. Therefore, it was not possible to link the summaries of treatment documentation of both databases based on patients' ID codes.

To nevertheless carry out the linking process, the documents of treatment cases from the different data-sets were tried to be linked based on the following variables:

- commercial registry code of the health care provider (*ttocode*);
- person's gender (*sex*);
- person's age (*age*);
- the county of residence of the person at the moment of provision of health care service (*county*);
- main and secondary diagnoses of the treatment case (*dgn*);
- external causes of the treatment case (*causecode*);
- start date of the treatment case (*begin*);
- end date of the treatment case (*end*);
- type of health care service (*type*);
- emergency care (yes/no) (*inevitable*);
- existence of health insurance for the patient (*insurance*).

As a result of the initial data analysis, it became evident from the observed variables that there are many instances of duplication in both the Health Insurance Fund's as well as the Health Information System's injury data. The analysis used the function *Merge* of the software STATA that did not allow linking repeating data. To facilitate the task, analytics at a meeting of NIHD's Health Statistics Department decided that in testing out all combinations, only such treatment cases were to be linked that are unique in terms of the observed combination.

Before the observed combinations were linked, the Health Insurance Fund's data was examined to find which uninsured patients had an age value existent on some other treatment invoice based on the same ID code. To get more accurate data-linking results, treatment invoices with incomplete information were temporarily imputed with the patient's age value from treatment invoices with existent data.

3.1 Entry errors and taking them into account

One hypothesis the data-linking ment to test was that doctors make entry errors i.e. enter data of the same treatment case differently into each system (values are different or a value has been entered into one system but not to the other). These differences and peculiarities were needed to be taken into account to find from the Health Information System's database ages for as many Health Insurance Fund's patients without insurance as possible.

The errors that were able to be discover in the linking processes (occurred in the data of both systems) were the following:

- the county of residence of the person at the moment of provision of health care service is different or missing;
- the main diagnose code on the treatment case is entered differently or left unspecified;
- the external cause code on the treatment case is entered differently or left unspecified;
- difference in the start date of the treatment case;
- difference in the end date of the treatment case;
- change of patient's age if the treatment case's dates differ;
- difference in the type of health care service;
- emergency care is entered differently or left unspecified;
- existence of health insurance for the patient is different.

The linking-process showed that the more the errors made by doctors are taken into account, the more situations occur in joined data where one ID code in one dataset is matched with two different ID codes of the other dataset. Therefore, what was needed was to select the most reasonable error occurrence combinations to be considered in the linking process.

A team of senior analysts and analysts from the National Institute for Health Development's Health Statistics Department was formed to choose the most reasonable combinations in the linking process.

3.2 Linking the observed combinations of the databases

In total, 187 different combinations were selected (see Annex 4). For the observed combinations, certain variables were allowed to be different in the linking process because it was assumed that doctors could have made errors in entering the data.

It was discovered during the linking process that doctors could have submitted documentation for the same treatment case at different times, i.e. later into one system than into the other. In cooperation with the team members of the NIHD's Health Statistics Department, it was decided that the maximum allowed period between two of the exact same epicrises is 30 days.

During the linking process, it was discovered that if the age variable was allowed to be different for all treatment cases of the Health Insurance Fund, then there could arise errors in the linking process, where the treatment cases of a 17-year-old patient of one database and an 80-year-old patient of another matched. As a result, it was unanimously decided that the patient's age variable could differ only for the treatment cases of persons without the age variable in the Health Insurance Fund's database (see Annex 4 "(age)", "no age").

In addition, it was taken into account in the linking process that the patient's age on the epicrisis is determined on the start date of the treatment case. If, however, the epicrisis was submitted later into one system than the other, the observed person could have got year older (see Annex 4 cases "156–171" and "184–187" "age ±1").

As a result of the linking, 157,620 observations linked from both databases. 28,663 observations from the Health Information System were unlinked and 137,124 observations from the Health Insurance Fund's database did not find a match.

3.3 Examples of linked rows

The data used in the following examples to describe the linking process are fictitious and do not reflect reality since sensitive personal data have been used in this bachelor's thesis. In the following examples, the data of one object are divided between two rows.

Example 1

Let's assume that the observation in the Health Insurance Fund's database of injury related data is the following

ttocode_HK	ID_HK	type_HK	age_HK	sex_HK	county_HK	dgn_HK
90001479	11436453	Stationary	77	Man	Järvamaa	S72.08

causecode_HK	begin_HK	end_HK	inevitable_HK	insurance_HK
W10.01	11.11.2016	12.11.2016	Yes	Yes

and correspondingly the Health Information System dataset has the object

ttocode_TIS	ID_TIS	type_TIS	age_TIS	sex_TIS	county_TIS	dgn_TIS
90001479	1098765	Stationary	77	Man	Järvamaa	S72.08

causecode1_TIS	begin_TIS	end_TIS	inevitable_TIS	insurance_TIS
W10.01	11.11.2016	12.11.2016	Yes	Yes

Based on tried and tested combinations, these observations would link in case 1.0 (see Annex 4).

Example 2

If in the Health Insurance Fund's database we would have the observation

ttocode_HK	ID_HK	type_HK	age_HK	sex_HK	county_HK	dgn_HK
90001498	10083912	Stationary	0	Woman	Tartumaa	T84.0

causecode_HK	begin_HK	end_HK	inevitable_HK	insurance_HK
Y83.1	15.02.2016	15.02.2016	Yes	Yes

then in the Health Information System this would be linked to the observation

ttocode_TIS	ID_TIS	type_TIS	age_TIS	sex_TIS	county_TIS	dgn_TIS
90001498	454357379	Stationary	0	Woman	Tartumaa	T84.0

causecode1_TIS	begin_TIS	end_TIS	inevitable_TIS	insurance_TIS
Y83.19	15.02.2016	15.02.2016	.	Yes

in case 5.2 (see Annex 4).

Example 3

If in the Health Insurance Fund's database we would have the following observation of an uninsured patient

ttocode_HK	ID_HK	type_HK	age_HK	sex_HK	county_HK	dgn_HK
90004587	11352739	Stationary	.	Woman	Viljandimaa	S06.6

causecode_HK	begin_HK	end_HK	inevitable_HK	insurance_HK
W19.48	11.09.2016	17.09.2016	No	No

then in the Health Information System this would be linked to the observation

ttocode_TIS	ID_TIS	type_TIS	age_TIS	sex_TIS	county_TIS	dgn_TIS
90001498	454357379	Stationary	34	Woman	Viljandimaa	S06.6

causecode2_TIS	begin_TIS	end_TIS	inevitable_TIS	insurance_TIS
W19.49	11.09.2016	17.09.2016	.	No

in case 6.2 (see Annex 4).

Example 4

The observation in the Health Insurance Fund’s database

ttocode_HK	ID_HK	type_HK	age_HK	sex_HK	county_HK	dgn_HK
90001478	172990769	Stationary	15	Man	Tartumaa	S53.40

causecode_HK	begin_HK	end_HK	inevitable_HK	insurance_HK
W51.01	17.03.2016	24.03.2016	Yes	Yes

would be linked to the observation in the Health Information system

ttocode_TIS	ID_TIS	type_TIS	age_TIS	sex_TIS	county_TIS	dgn_TIS
90001478	172990769	Stationary	16	Man	Tartumaa	S53.49

causecode2_TIS	begin_TIS	end_TIS	inevitable_TIS	insurance_TIS
W51.01	24.03.2016	24.03.2016	Yes	Yes

in case 156.1 (see Annex 4).

4 Imputation

Age values from the Health Information System were imputed to the treatment invoices of patients without insurance from the Health Insurance Fund. As the result of the data linking, ages were found to 5633 patients, and 3515 epicrisis remained without this variable's value.

Three different methods were used for imputation: general random Hot-Dock imputation, the k-nearest neighbour imputation and random Hot-Deck imputation within class combined with k-nearest neighbour imputation. Problems occurred with the latter method because in grouping the objects donor groups formed where no observations had an age value. The "VIM" package of the software program R gave such observations the age of 1 which in the opinion of the analysts of the National Institute for Health Development was not the right approach. Therefore, the k-nearest neighbour method was decided to be used all across the dataset for observations that ended up in empty donor groups. (see Code 3)

As an expert opinion, the team on analysts and senior analysts from the National Institute for Health Development decided to involve into the process the following original variables:

- *type* – type of health care service (1 - "outpatient", 2- "inpatient");
- *sex*– patients gender (1 - "Male", 2 -"Female");
- *age* – patient's age (integer value).

In addition, the following variables were decided to be created for the imputation (see Code 1):

- *familydoctor* – was this a family doctor's offices' visit or visit to a larger medical institution (0 - "larger medical institution", 1 - "family doctor");
- *sum_group* – patient's treatment invoice sum range ("...–100", "101–200", "201–..");
- *dgn_group* – group of treatment case's main diagnose code according to ICD-10 (see Annex 1);
- *causecode_group* – group of treatment case's external cause code according to ICD-10 (see Annex 2);
- *days_group* – duration of patient's treatment case in days („<=0", „0-5", „6-10", „>10");
- *begin_month* – month of start date of treatment invoice (1 - "January", ..., 12 - "December").

The following variables were left out of imputation:

- *inevitable* – whether it was emergency care or not (0 - "No", 1 - "Yes");
- *emergency* – whether it was care by emergency medical care department or not (0 - "No", 1 - "Yes");
- *county* – patient's residence at the moment of provision of health care service;
- *end* – end date of the treatment invoice.

Variables *inevitable* and *emergency* were left out of imputation because the expert opinion of the senior analysts was that the health care providers do not enter the values of these variables properly. The variable *county* was not included in the imputation process because the expert opinion was that the residence of a person does not influence the type of external cause of injury, patient's gender nor age. In addition, the values of this variable were rarely entered into the Health Insurance Fund's database. The variable *end* was not used in imputation because the variables *begin_month* and *days_group* were already included.

In the data of patients without medical insurance, there was a total of 9 missing values for the external cause code variable, 37 missing values for the gender variable and 3515 missing values for the age variable. In the process of imputation, it was decided to impute the values of external cause code group first, then gender values and then age values. In the imputation of each subsequent variable, the imputation results of the previous variable were also used.

4.1 Preliminary analysis of data

It was decided that to impute a variable, only the auxiliary variables were to be used that, based on existent data, would have a statistical relation to the variable being imputed. In this bachelor's thesis χ^2-tests and T-tests have been used to study statistical relations and they have been carried out with the software program R. The level of significance is $\alpha = 0.05$.

The relations of the variable *causecode_group* with the auxiliary variables were studied only with the help of χ^2-tests since this was a nominal variable (see Code 2.1). As a result, the following variables were obtained to impute the observed variable:

- *type*;
- *days_group*;
- *sum_group*;
- *begin_month*;
- *dgn_group*;
- *familydoctor*.

The relations of the variable *sex* with the auxiliary variables were studied only with the help of χ^2-tests since this was a binary variable (see Code 2.2). As a result, the following variables were obtained to impute the observed variable:

- *sum_group*;
- *dgn_group*.

The age relations of continuous variables with nominal auxiliary variables that have two values were studied with T-tests and variables with the more possible values with χ^2-tests. To carry out the latter test, the variable *agegroup* was created (see Code 2.3). As a result, the following variables were obtained to impute the age variable:

- *type*;
- *days_group*;
- *sum_group*;
- *begin_month*;
- *sex*;
- *dgn_group*;
- *causecode_group*.

4.2 Imputation results

The imputation was carried out with the software program R, using the package "VIM" (see Code 3).

4.2.1 Imputation results of the external cause of an injury group code

The variable *causecode_group* had a total of 9 missing values. Descriptions of imputed groups are available from Annex 2.

Table 1. Percentage of ICD-10 code groups of external causes of injuries and number of imputed values (see Code 4)

	Gr 1	Gr 2	Gr 3	Gr 4	SUM
General random Hot-Deck method	5.14% (0)	72.60% (9)	7.37% (0)	14.90% (0)	100.01%
K-Nearest neighbour method	5.14% (0)	72.58% (8)	7.38% (1)	14.92% (0)	100.00%
Random Hot-Deck method within classes and k-nearest neighbour method	5.14% (0)	72.55% (5)	7.39% (2)	14.90% (2)	100.00%
Existing data	5.14%	72.57%	7.37%	14.91%	99.99%

Table 1 shows that the imputation results are not very dissimilar in terms of percentages when different methods are used. It can be seen that the percentage of external causes from physical factors (*Gr 2*) has increased with random Hot-Deck method as well as the k-nearest neighbour method – 0.03% and 0.01%, respectively, where 9 and 8 values were imputed, respectively. The percentage of external causes of injuries from transport accidents (*Gr 1*) has not changed in any of the methods because no values were imputed. The percentage of external causes of injuries from human and other factors (*Gr 4*) has only increased with random Hot-Deck imputation within class (the k-nearest neighbour method was not used in the combination when imputing this variable, see Code 4) and by 0.01% – 2 values were imputed. With other methods, the percentage of this value has decreased 0.01% because no values were added. The percentage of external causes of injuries from natural events and other factors (*Gr 3*) has increased with the k-nearest neighbour method by 0.01% (1 value was imputed) and with Hot-Deck imputation within class by 0.02% (2 values were added).

4.2.2 The imputation results of the gender variable

The variable *sex* had a total of 37 missing values.

Table 2. Percentages of genders (see Code 4)

	Presentage	
	Women	Men
General random Hot-Deck method	16.48% (5)	83.52% (32)
K-Nearest neighbour method	16.43% (0)	83.57% (37)
Random Hot-Deck method within classes and k-nearest neighbour method	16.46% (3)	83.54% (34)
Existing data	16.50%	83.50%

Table 2 shows that there are no large differences in the percentages of the gender variable when data is imputed with different methods. In the existent data, 16.5% of treatment cases involved female patients and 83.5% of treatment cases involved male patients. The percentage of male patients in the data increased with all methods. With the k-nearest neighbour method, the variable *sex* was imputed with the value "Man" in all 37 observations and consequently, the percentage of male patients increased by 0.07% in the data. With the general random Hot Deck method, 5 female and 32 male gender values were imputed – consequently, the percentage of treatment invoices of men increased and the percentage of treatment invoices of women decreased by 0.02% in the data. With the random Hot-Deck imputation within class, the variable *sex* was imputed the value "Woman" 3 times and value "Man" 34 times. The percentage of the latter increased by 0.04% in the data (the k-nearest neighbour method was not used in the imputation of this variable, see Code 4).

4.2.3 The imputation results of the patient's age variable

In the observed data, patient's age was represented on 5633 treatment invoices and was missing from 3515 treatment invoices.

Table 3. Age characteristics across all data (see Code 4)

	Mean	Std	Min	Max	Med
General random Hot-Deck method	36.969	11.856	0	94	35
$\boldsymbol{K}$-Nearest neighbour method	36.715	11.862	0	94	35
Random Hot-Deck method within classes and $\boldsymbol{k}$-nearest neighbour method	36.519	11.837	0	94	35
Existing data	36.890	11.858	0	94	35

There were no differences in minimal and maximum age when different methods were used – the minimal age remained 0 and maximum 94, which is understandable because imputation is based on existent data (see Chapter 1.2). The median age remained 35 with all methods. Based on existent data, the average age was initially 36.890 years and the dispersion of the variable was 11.858. Table 3 shows that there are no large differences between the imputed data. The dispersion of the value being imputed has increased only with the k-nearest neighbour method, but only by 0.004. The average age has increased with the general random Hot-Deck method to 36.969 years and decreased with all other methods – to 36.715 years with the k-nearest neighbour method and to 36.519 years with the random Hot-Deck imputation within class combined with the k-nearest neighbour method.

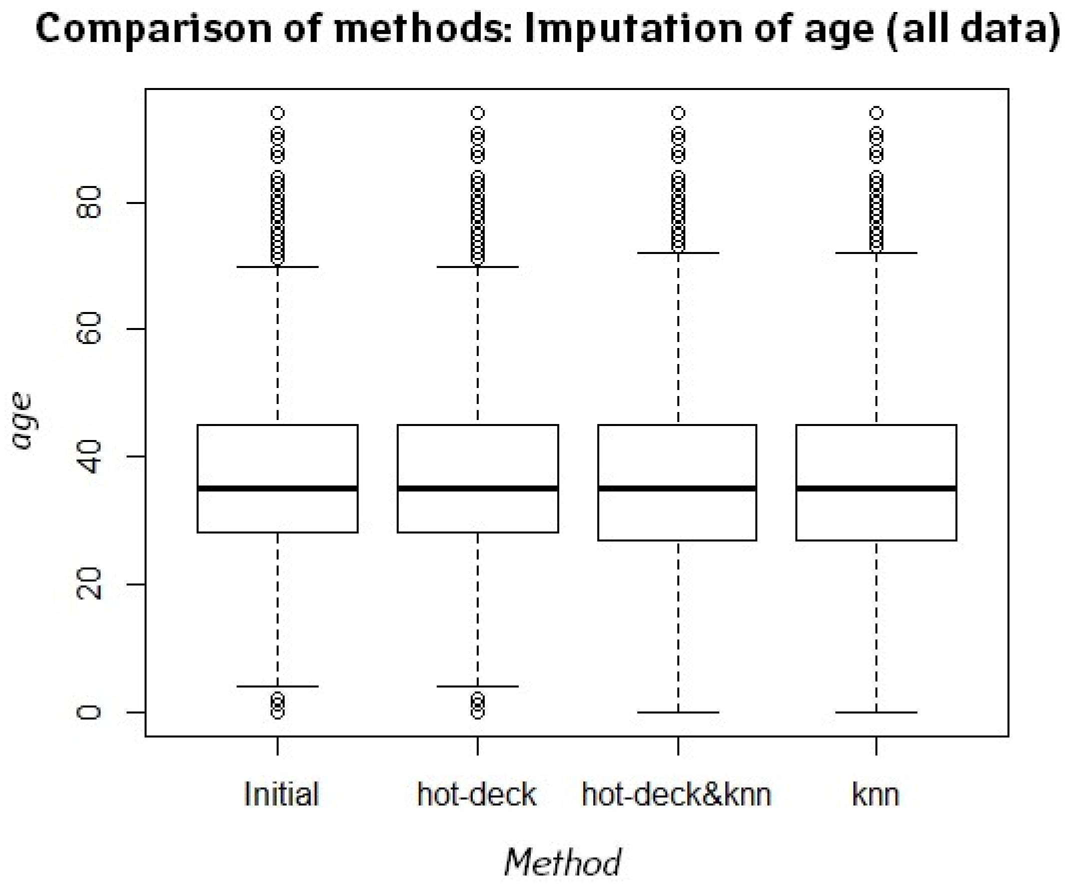


Figure 1. Age box plot in initial and final data-sets after imputation (see Code 4)

The box plot shows quartiles and median of numerical characteristics with horizontal lines the end points of which are connected with vertical lines. Table 1 shows that in the final datasets the observable numerical characteristics of the age variable are the same in all results and do not differ from initially existent data. Maximum and minimum values of the observable variable are shown by the end points of the whiskers that in this case are also quite similar between different datasets. It can be seen that the aforementioned characteristics are a little more similar with initial values in the case of general random Hot-Deck method (*hot-deck*), small differences can be seen with the combination of group-based Hot-Deck method and the k-nearest neighbour method (*hot-deck&knn*) and with the k-nearest neighbour method *(knn).* Dots demarcate the observations that are further than one and a half quartiles away from the median. It can be seen that these have also distributed similarly with different methods.

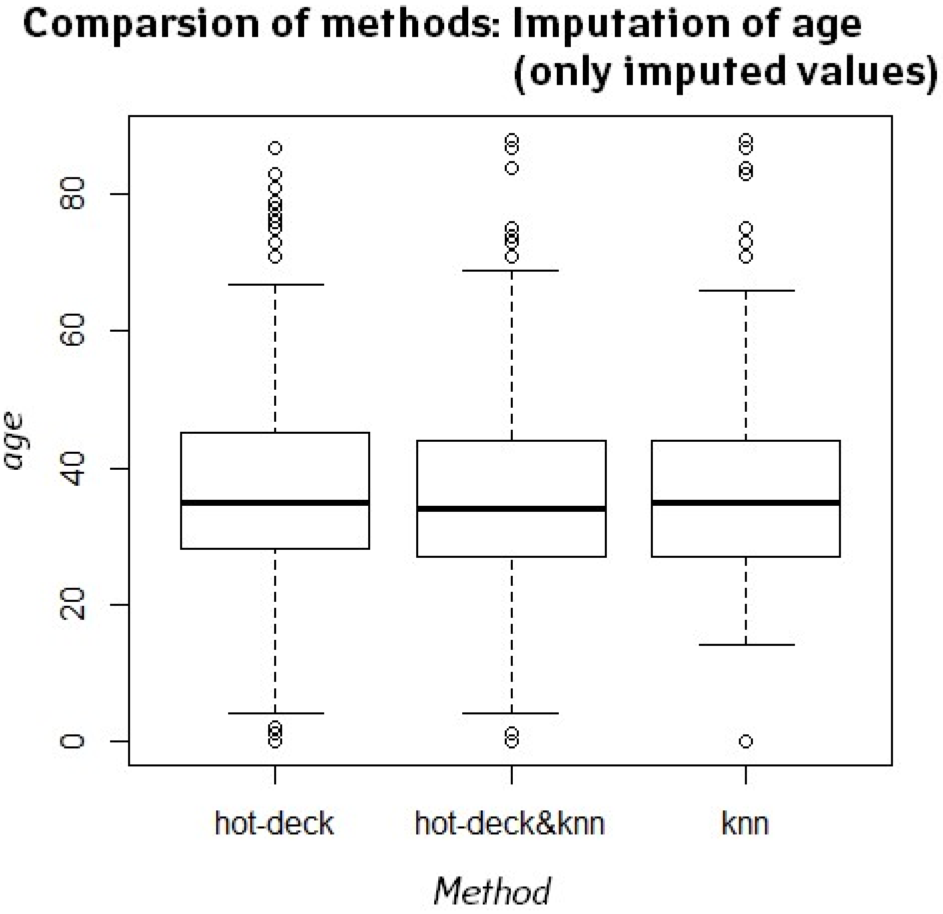


Figure 2. Comparison of the methods for the imputation of age in the final dataset (see Code 4)

Figure 2 only takes imputed values into acount. This figure shows that the k-nearest neighbour method *(knn)* has the least variability in the imputation of patients' ages on treatment invoices. The imputed data shows more variability of the values of the age variable when the general random Hot-Deck method (*hot-deck*) and the combination of the random Hot-Deck imputation within class and the k-nearest neighbour method (*hot-deck&knn*) is used. According to the prediction by the analysts of the National Institute for Health Development, the patients without health insurance should be in the age group to which the k-nearest neighbour method imputed ages the most.

5 Quality of imputation methods

In this bachelor's thesis, the existence of the age value for the patient without health insurance in the Health Insurance Fund's database depends on the quality of linking the data of the Health Information System and Health Insurance Fund. Treatment cases could have been left unlinked in this process if the doctor made considerable errors in entering the data or if some treatment cases were too similar to be linked (duplicates were generated). However, these situations could have arisen completely randomly. Therefore, the absence of age can be considered as missing completely at random (MCAR). In this case, the studied variable does not depend on any auxiliary variables and the distribution of existent values is the same as for missing values (2).

To compare the imputation methods, those treatment invoices of uninsured patients were selected from linked data which had values for all variables (code group of external cause of injury, gender, age). In this chapter, two tests are conducted. In the first, at least 70% and in the second, at least 50% of the age value is retained at completely random (see Code 5.1 and 6.1) and then the imputations based on the methods used in this bachelor's thesis are carried out (see Code 5.2 and 6.2). The imputation results obtained are compared to initial values and based on this analysis, the best imputation method is selected.

The values of imputation results have been examined alone as well, i.e. what is the difference with the patient's real age. For this, only those rows where the age value was imputed have been separated and the observations are linked with the initial data. In the linked data, the age difference variable was created that shows the difference of the imputation result from the real value (see Code 5.3 and 6.3).

In order to ensure the quality of the imputation method, 100 simulations were carried out. In the first test, the methods were used 100 times on the exact same data to impute ages for patients of treatment case. In the second test, 70% and 50% of data were respectively retained before each step of the simulation with imputation methods used at each step on different initial data (see Code 5.4 and 6.4). The results obtained were analysed with box plots (see Code 5.5 and 6.5).

5.1 Quality of imputation upon 70% of information

Table 4 shows the characteristics of results obtained upon imputation with 30% of data missing.

Table 4. Age characteristics across all data (see Code 5.3)

	Mean	Std	Min	Max	Med
General random Hot-Deck method	37.200	12.009	0	94	36
$\boldsymbol{K}$-Nearest neighbour method	36.942	12.020	0	94	35
Random Hot-Deck method within classes and $\boldsymbol{k}$-nearest neighbour method	37.126	12.002	0	94	36
Actual	36.897	11.858	0	94	35

Table 4 shows that with the general random Hot-Deck method and the combination of the donor group based Hot-Deck method and the k-nearest neighbour method the median age is different from the actual – instead of 35, it is 36. The minimums and maximums of age remained unchanged because values are imputed from existing data (see Chapter 1.2) and if maximum and minimal values were not deleted, they will also remain the same in the imputed data. The table also shows that the dispersion of the observable age variable has increased with all imputation methods. This situation is probably due to the donor based imputation methods having an imputation bias (see Chapter 1.3) that in turn increases the standard deviation. The real average age was 36.897 years and there were no big changes in the value of the observable characteristic upon imputation. The most accurate average was given by the data imputed with the k-nearest neighbour method – 36.942. With the general random Hot-Deck method and the combination of the donor group based Hot-Deck method and the k-nearest neighbour method the average age of the data was over 37 years – 37.200 and 37.126 years, respectively.

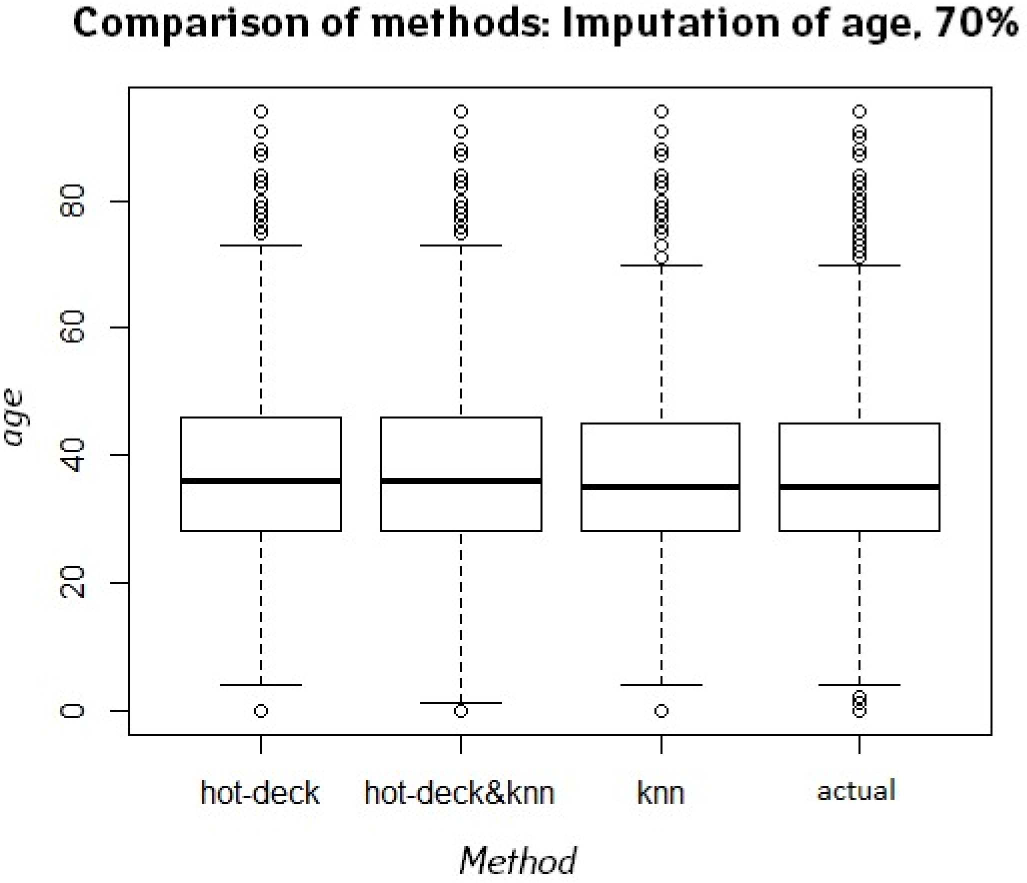


Figure 3. Box plot of age imputation results across all data (see Code 5.3)

The box plot of Figure 3 shows that there are no significant differences between the results of imputation methods. Although the age variable's quartiles and median are the same with different methods, there are slight differences primarily in the minimum and maximum values of the sample for which the best results are produced with the k-nearest neighbour method *(knn)*. With general Hot-Deck method (*hot-deck*), there is a slight increase of the sample's maximum value. With the combination of group-based Hot-Deck method and the k-nearest neighbour method (*hot-deck&knn*), the both the maximum and minimum absolute value of the sample's age variable has increased. Observations demarcated with dots that are further than one and a half quartiles away from the median have also distributed similarly with different methods.

Table 5 shows the characteristics of age differences obtained upon imputation with 30% of data missing. The results are only observed across imputed data.

Table 5. The characteristics of imputation results and real ages based on imputed data (see Code 5.3)

	Mean	Std	Min	Max	Med
General random Hot-Deck method	-0.709	16.566	-53	57	-1
K-Nearest neighbour method	0.215	16.666	-69	64	0
Random Hot-Deck method within classes and k-nearest neighbour method	-0.11	16.426	-69	64	1

The examination of Table 5 shows that the smallest bias of average age estimate was with the combination of random Hot-Deck imputation within class and the k-nearest neighbour method: -0.11. The largest bias of -0.709 was with general random Hot-Deck Method and the average result of 0.215 was achieved with the k-nearest neighbour method. However, the highest variability of age difference was given by the k-nearest neighbour method with 16.666. The results of general random Hot-Deck method with 16.566 and the combination of group-based random Hot-Deck imputation and the k-nearest neighbour method with 16.426 were slightly better. It should be mentioned here that with the general random Hot-Deck method, the minimum and maximum value are the smallest in terms of absolute values: -53 and 57, respectively. With the other two methods, the minimum age difference was -69 and maximum 64. The best median age value was given by the k-nearest neighbour method with 0, the general random Hot-Deck method obtained 1 as the value, and the third examined method obtained -1 as the value of the characteristic.

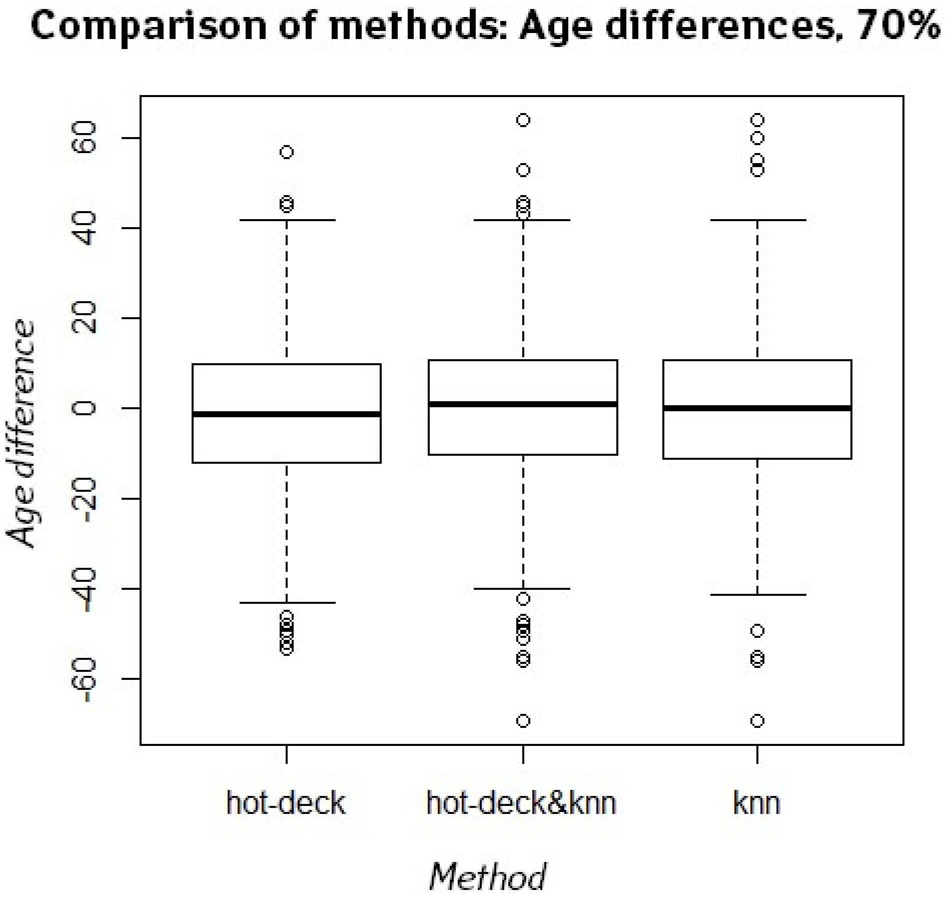


Figure 4. The box plot of imputation results and real ages based on imputed data (see Code 5.3)

It can be seen from the box plot of Figure 4 that the characteristics of the differences between real and imputed ages are quite similar between different methods. The observations demarcated as dots show that the outliers are not so different with the general Hot-Deck method (*hot-deck*) as they are with other methods (as can be seen from Table 5 as well).

5.2 Simulation with at least 70% of information

In the first simulation test, data was imputed with different methods on the same data, i.e. 70% of data was retained once and after this, data was imputed 99 times (see Code 5.4). Taking into account the results of previous imputations was regarded as one step of the simulation and therefore, 100 simulations were carried out in total.

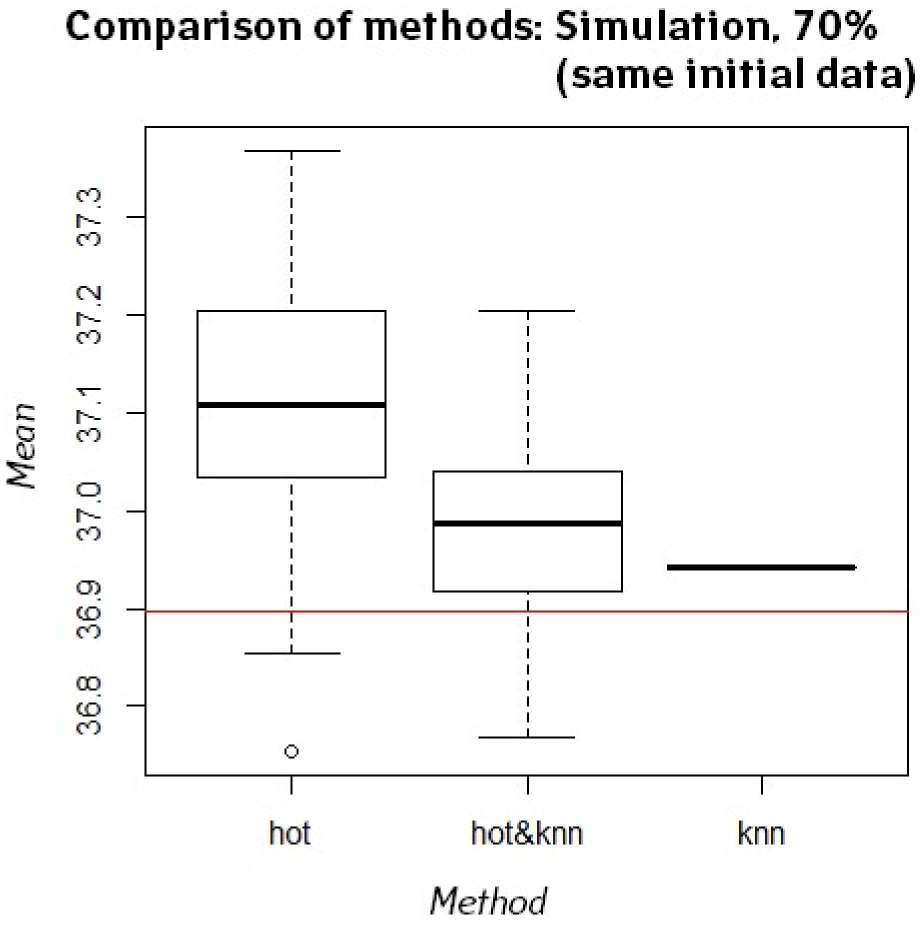


Figure 5. Comparison of methods in simulation (70%) – same initial data (see Code 5.4)

The examination of Figure 5 shows that the k-nearest neighbour method *(knn)* gives the same result on every step of the simulation. This may be due to always imputing to missing values the values of the same neighbours (see Chapter 1.6). Regardless of this, the median of the average of the results obtained with the observed imputation method is the closest to the actual average (the red line on the figure). The next best result was given by the group-based random Hot-Deck imputation combined with the k-nearest neighbour method *(hot&knn)* because the quartiles and median of the average imputation results are closer than with the general random Hot-Deck method *(hot)*. Unfortunately, the best model could not be decided based on this figure alone.

In the second simulation, data was imputed on different initial data at each of the 99 steps, i.e. on each imputation attempt, approximately 30% of the data was deleted and after this, the data was imputed with different methods (see Code 5.5). In this simulation, taking into account the results of the first imputation was also regarded as one step of the simulation – therefore, 100 simulations were carried out in total. This test was carried out to examine closer the k-nearest neighbour method.

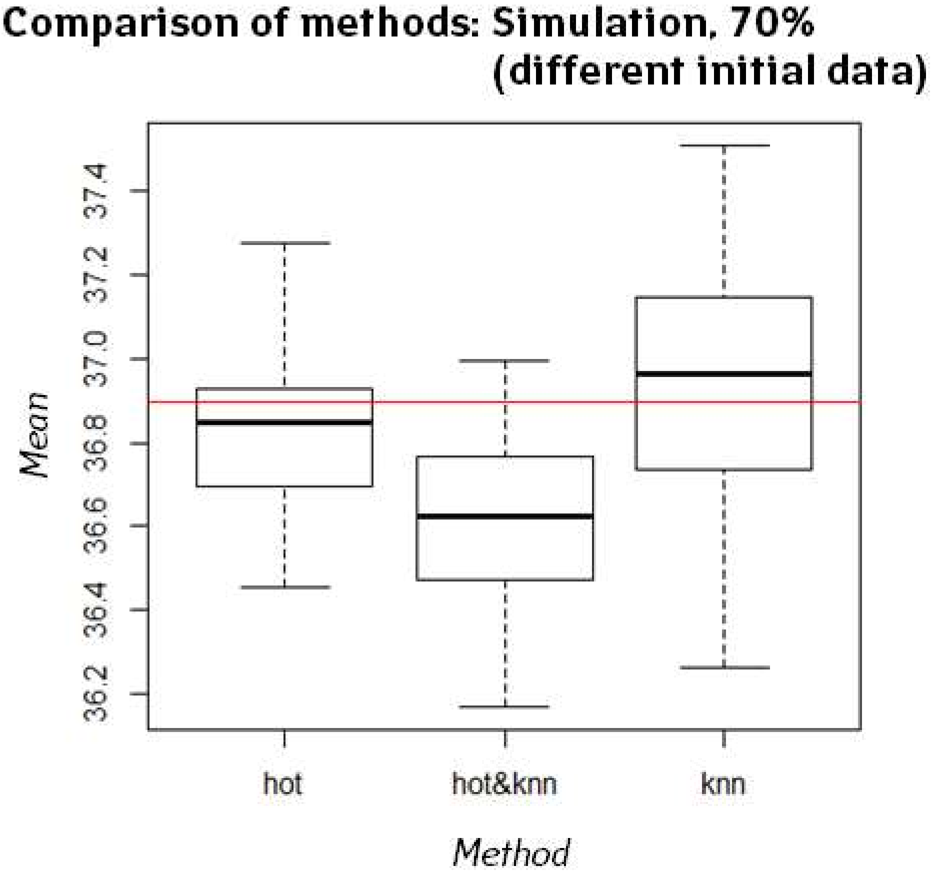


Figure 6. Comparison of methods in simulation (70%) – different initial data (see Code 5.5)

It can be seen from Figure 6 that the k-nearest neighbour method *(knn)* and the general random Hot-Deck method *(hot)* give the best results because the medians of the results obtained with the observed imputation methods are the closest to the real average (the red line on the figure). With the general Hot-Deck method, the upper quartile of the observed statistics, and with the k-nearest neighbour method, the lower quartile are closer to the real value. It should be pointed out that the averages of the results imputed with the k-nearest neighbour method have the most variability which made the investigators suspicious of the stability of the method. Therefore, the conclusion was made from this test that the best results were given by the general random Hot-Deck implementation. The most imprecise results were given by the group-based random Hot-Deck implementation combined with the k-nearest neighbour method *(hot&knn)*.

5.3 Quality of imputation with at least 50% of information

Table 6 shows the characteristics of results obtained upon imputation with 50% of data missing.

Table 6. Age characteristics across all data (see Code 6.3)

	Mean	Std	Min	Max	Med
General random Hot-Deck method	36.909	11.949	0	94	35
$\boldsymbol{K}$-Nearest neighbour method	37.158	11.761	0	94	36
Random Hot-Deck method within classes and $\boldsymbol{k}$-nearest neighbour method	36.488	12.075	0	94	35
Actual	36.897	11.858	0	94	35

It can be seen from table 6 that the age median is different from the real when the k-nearest neighbour method is used – 36 instead of 35. Minimum and maximum age values have remained the same. The actual dispersion of the age variable was 11.858. The table shows that with the k-nearest neighbour method, the dispersion of the observed characteristic has even decreased by 0.097. It has increased with other methods – by 0.091 with the general random Hot-Deck method and by 0.217 with the combination of the donor group based Hot-Deck method and the k-nearest neighbour method. The real average age was 36.897 years and there were no big changes in the value of the observable characteristic upon imputation. The most accurate average was given by the data imputed with the general random Hot-Deck method −36.909. With the k-nearest neighbour method, the average age was 37.158 and with the combination of the donor group based Hot-Deck and k-nearest neighbour method, the value of the observed characteristic was 36.488.

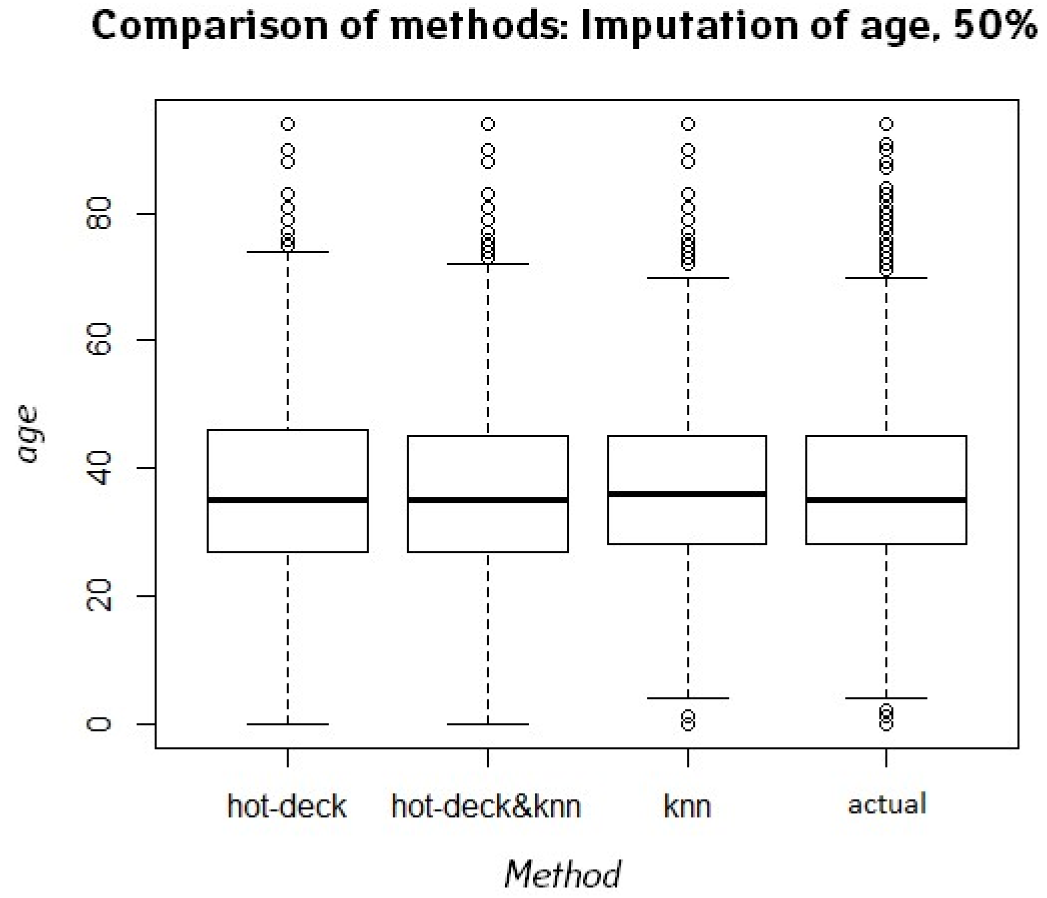


Figure 7. Box plot of age imputation results across all data (see Code 6.3)

The box plot of Figure 7 shows that there are no large differences between the characteristics of the imputation results. With all methods, the quartiles and median of the age variable are the same as with the actual data. The k-nearest neighbour method *(knn)* has given a slightly better result for the minimum and maximum values of the sample's age variable. With general Hot-Deck method (*hot-deck*), there is an increase of the sample's maximum value and a decrease of the minimum value. Results are better compared to the previous test with the group-based Hot-Deck method and the k-nearest neighbour method (*hot-deck&knn*) – although the minimum value of the sample's observable variable has decreased, then the maximum value does not differ from reality that much anymore. The reason behind this may be that if 50% of data was missing, then more empty groups were formed and therefore the k-nearest neighbour method was still used on the majority of data. This, however, improved the results obtained. Observations demarcated with dots have distributed similarly upon the use of different methods.

Table 7 shows the characteristics of age differences obtained upon imputation with 50% of data missing. The results are only observed across imputed data.

Table 7. The characteristics of imputation results and real ages based on imputed data (see Code 6.3)

	Mean	Std	Min	Max	Med
General random Hot-Deck method	-0.212	16.839	-63	65	0
$\boldsymbol{K}$-Nearest neighbour method	-0.619	16.128	-69	62	0
Random Hot-Deck method within classes and $\boldsymbol{k}$-nearest neighbour method	-0.774	16.606	-72	60	-1

The examination of Table 7 shows that the smallest bias of average age estimate in this test was with the general random Hot-Deck imputation: -0.212. The largest bias of -0.744 was with the group-based random Hot-Deck method combined with the k-nearest neighbour method and the average result of -0.619 was achieved with the k-nearest neighbour method. The smallest variability of age difference was given by the k-nearest neighbour method with 16.128. The results of the general random Hot-Deck method with 16.839 and the combination of group-based random Hot-Deck imputation and the k-nearest neighbour method with 16.606 were worse. With the general random Hot-Deck imputation, the minimum and maximum age difference is -63 and 65. With the k-nearest neighbour method, the values of the observable characteristics were -69 and 62 respectively and with the third observed method, -72 and 60. The best median was given by the general random Hot-Deck and k-nearest neighbour method with 0. The value of the characteristic was yet again -1 with the group-based random Hot-Deck method combined with the k-nearest neighbour method.

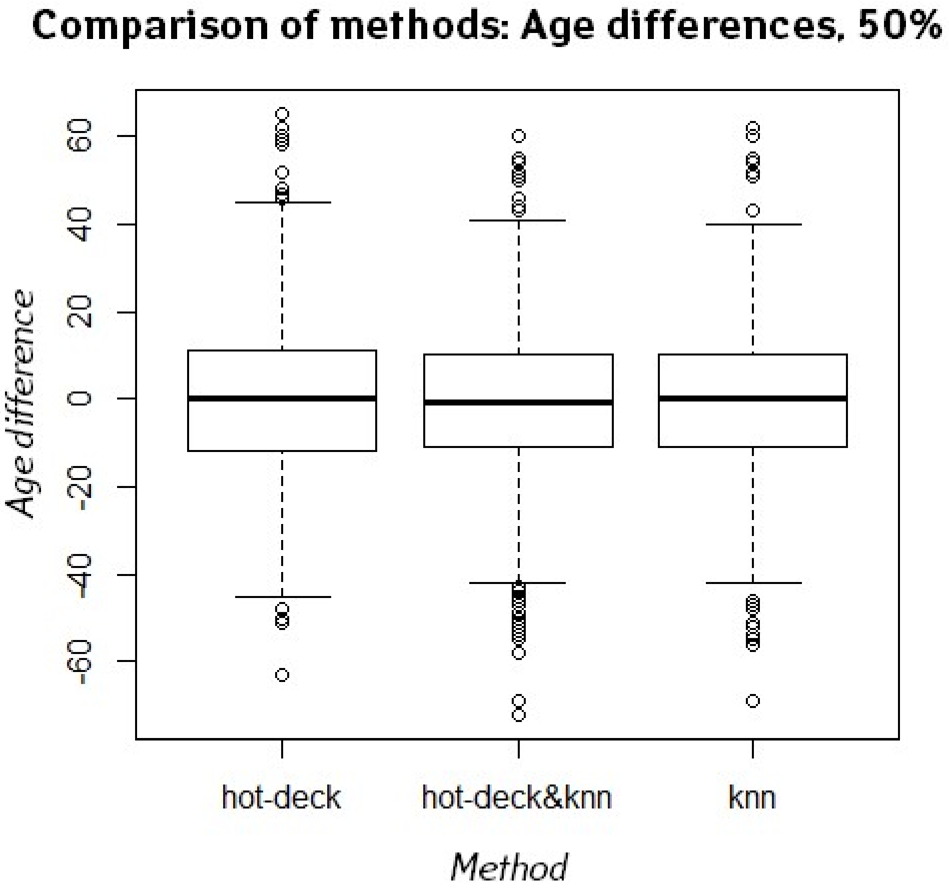


Figure 8. The box plot of imputation results and real ages based on imputed data (see Code 6.3)

The examination of the box plot of Figure 8 shows that the characteristics of the differences between real and imputed ages are again quite similar between different methods. The group-based random Hot-Deck method combined with the k-nearest neighbour method (*hot-deck&knn*) and the k-nearest neighbour method *(knn)* also give similar minimum and maximum values for the sample. The reason behind this might also be that many empty groups were formed with the use of the first method and therefore the k-nearest neighbour method was still used on the majority of data. With the general Hot-Deck method (*hot-deck*), the maximum value of the sample's observed variable is somewhat larger and the minimum value slightly smaller than with other methods.

5.4 Simulation with at least 50% of information

In the first test, data was attempted to be imputed with different methods on the same data, i.e. 50% of data was retained once and after this, data was imputed 100 times (see Code 6.4). Taking into account the results of previous imputations was regarded as one step of the simulation and therefore, 100 simulations were carried out in total.

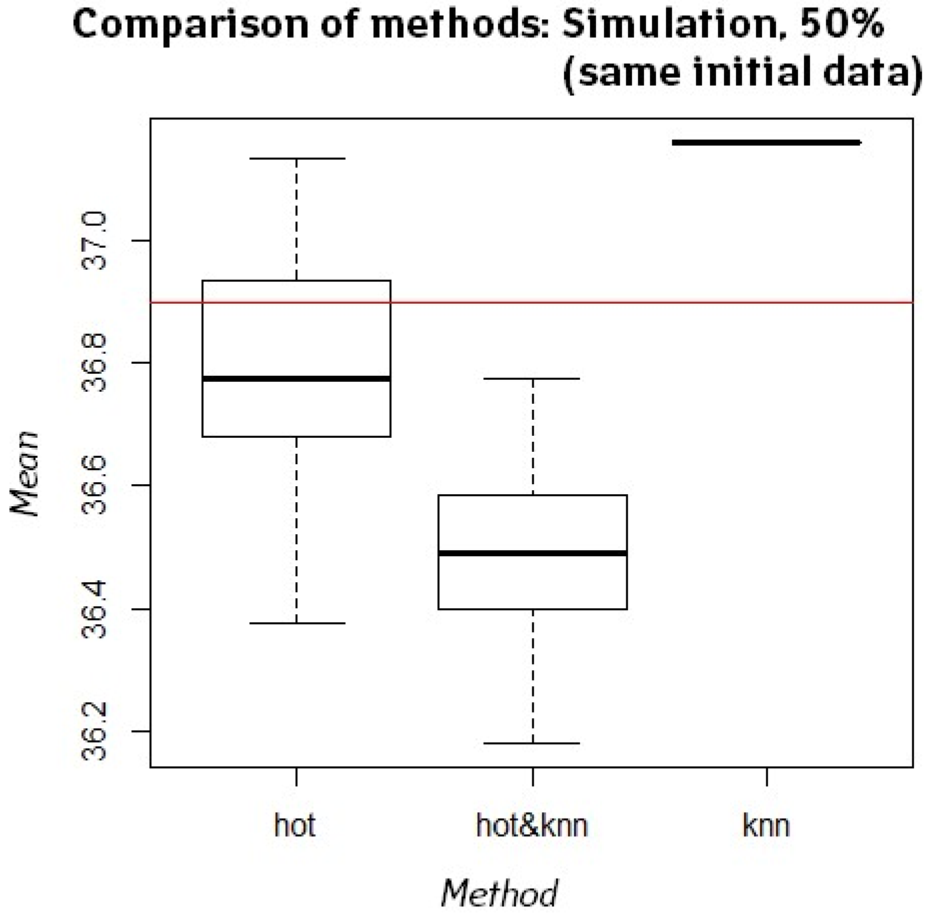


Figure 9. Comparison of methods in simulation (50%) – same initial data (see Code 6.4)

The examination of Figure 9 shows that the k-nearest neighbour method *(knn)* again gives the same result on every step of the simulation since on each step, the missing values are imputed with the same ages (see Chapter 1.6). Based on this box plot, it can be said that the most accurate imputation results are given by the general random Hot-Deck imputation *(hot)* with which the median and quartile of averages of age values are the closest to the real average age (the red line on the figure). In this case, the group-based random Hot-Deck implementation combined with the k-nearest neighbour method *(hot&knn)* gives the most imprecise results. Unfortunately, this figure does not enable to draw conclusions on the quality of the models because yet again there was lack of overview of the reliability of the k-nearest neighbour method.

In the second test, data was attempted to be imputed on different initial data, i.e. on each imputation attempt, approximately 50% of the data was deleted and after this, the data was imputed with different methods (see Code 6.5). In this simulation, taking into account the results of the first imputation was also regarded as one step of the simulation – therefore, 100 simulations were carried out in total. This process was carried out again to examine closer the quality of the k-nearest neighbour method.

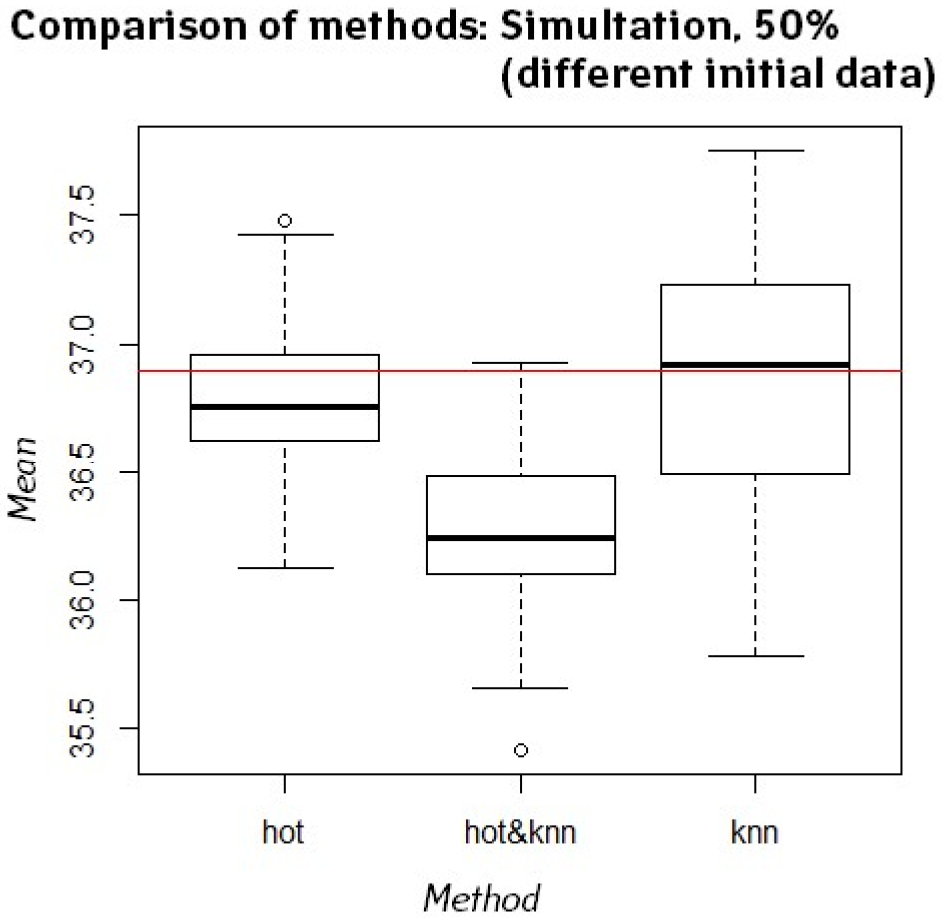


Figure 10. Comparison of methods in simulation (50%) – different initial data (see Code 6.5)

It can be seen from Figure 10 that the k-nearest neighbour method *(knn)* has given the most accurate median value for the averages of ages because this figure is the closest to the real average (the red line on the figure). The next best result for the median value was given by the general Hot-Deck method and the worst results by the combination of group-based random Hot-Deck imputation and the k-nearest neighbour method *(hot&knn)*. Also, it should be mentioned that with the general Hot-Deck method, the quartiles are closer to the real average values of age than with the k-nearest neighbour method. Based on this, it was decided that the k-nearest neighbour method gives the most unstable results and it would be more reasonable to use the general random Hot-Deck imputation here.

References

1. Andridge RR, Little RJA. A Review of Hot Deck Imputation for Survey Non-response. Int Stat Rev 2010 April;78(1):40–64. doi: 10.1111/j.1751-5823.2010.00103.x.
2. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? Int J Epidemiol 2014 August;43(4): 1336–1339. doi: 10.1093/ije/dyu080.
3. Bogovsi P. RHK-10: Rahvusvaheline haiguste ja nendega seotud terviseprobleemide statistiline klassifikatsioon. Tallinn: Tallinn Book Printers; 1996.
4. Estonian Language Foundation: Eesti õigekeelsussõnaraamat ÕS 2013; 2013. http://www.eki.ee/dict/qs/. Accessed 20 Feb 2018.
5. Kirpu V, Eigo N. Handling missing data and errors in Estonian eHealth information system. Tallinn: National Institute for Health Development; 2018. https://www.tai.ee/images/kirpu_eigo_university_of_tartu_2018_handling_missing_data_and_errors_in_estonian_ehealth_information_system.pdf. Accessed 10 April 2018.
6. Kowarik A, Templ M. Imputation with the R Package VIM. Journal of Statistical Software 2016 October;74(7). doi: 10.18637/jss.v074.i07.
7. Lepik N, Traat I. Tõenäosuslik valikuuring I. Lecture notes. Tartu: University of Tartu, Institute of Mathematical Statistics; 2016. https://courses.ms.ut.ee/MTMS.01.003/2016_fall/uploads/Main/loengud2016.pdf. Accessed 08 Feb 2018.
8. Prostakova J. Mittevastamine ja selle kompenseerimine [Bachelor's thesis]. Tartu: University of Tartu, Institute of Mathematical Statistics; 2007.
9. Toompere K. Imputeerimis- ja kaalumismeetodite mõju hinnangute nihkele [Master's thesis]. Tartu: University of Tartu, Institute of Mathematical Statistics; 2009.

Annexes

Annex 1. Explanations of main diagnose codes of injuries according to ICD-10

Nr	Name of the group	Range of codes	Meaning
1	Head and body area injuries	*S00–S09*	Injuries to the head
		S10–S19	Injuries to the neck
		S20–S29	Injuries to the thorax
		S30–S39	Injuries to the abdomen, lower back, lumbar spine and pelvis
2	Hands area Injuries	*S40–S49*	Injuries to the shoulder and upper arm
		S50–S59	Injuries to the elbow and forearm
		S60–S69	Injuries to the wrist and hand
3	Foot area injuries	*S70–S79*	Injuries to the hip and thigh
		S80–S89	Injuries to the knee and lower leg
		S90–S99	Injuries to the ankle and foot
4	Other unspecified area and other types of injuries or complications	*T00–T07*	Injuries involving multiple body regions
		T08–T14	Injuries to unspecified part of trunk, limb or body region
		T15–T19	Effects of foreign body entering through natural orifice
		T20–T32	Burns and corrosions
		T20–T25	...Burns and corrosions of external body surface, specified by site
		T26–T28	...Burns and corrosions confined to eye and internal organs
		T29–T32	...Burns and corrosions of multiple and unspecified body regions
		T33–T35	Frostbite
		T36–T50	Poisoning by drugs, medicaments and biological substances
		T51–T65	Toxic effects of substances chiefly nonmedicinal as to sourc
		T66–T78	Other and unspecified effects of external causes
		T79	Certain early complications of trauma
		T80–T88	Complications of surgical and medical care, not elsewhere classified
		T90–T98	Sequelae of injuries, of poisoning and of other consequences of external causes

Annex 2. Definitions of external cause codes of injuries according to ICD-10

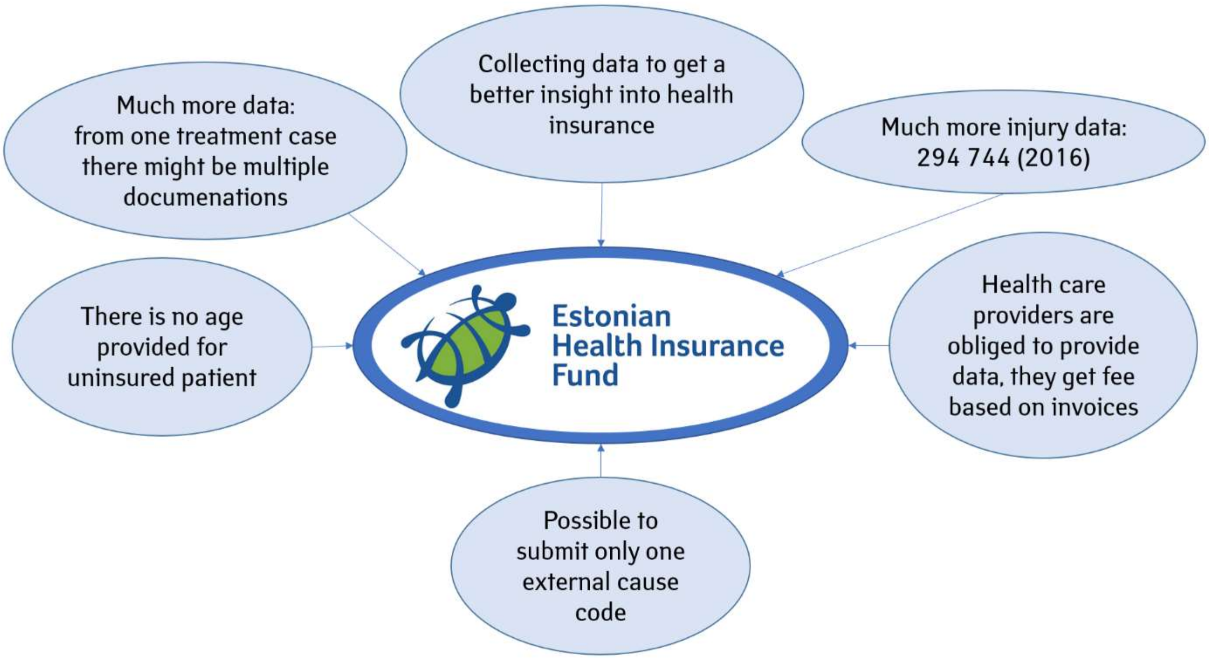
Nr	Name of the group	Range of codes	Meaning
1	External causes of injuries caused by transport accidents	*V01–V09*	Pedestrian injured in transport accident
		V10–V19	Pedal cyclist injured in transport accident
		V20–V29	Motorcycle rider injured in transport accident
		V30–V39	Occupant of three-wheeled motor vehicle injured in transport accident
		V40–V49	Car occupant injured in transport accident
		V50–V59	Occupant of pick-up truck or van injured in transport accident
		V60–V69	Occupant of heavy transport vehicle injured in transport accident
		V70–V79	Bus occupant injured in transport accident
		V80–V89	Other land transport accidents
		V90–V94	Water transport accidents
		V95–V97	Air and space transport accidents
		V98–V99	Other and unspecified transport accidents
2	External causes of injuries caused by physical factors (e.g. falls, electricity, etc.)	*W00–W19*	Falls
		W20–W49	Exposure to inanimate mechanical forces
		W50–W64	Exposure to animate mechanical forces
		W65–W74	Accidental drowning and submersion
		W75–W84	Other accidental threats to breathing
		W85–W99	Exposure to electric current, radiation and extreme ambient air temperature and pressure
3	External causes of injuries caused by different natural events and other factors (e.g. fire, burns, poisoning, etc.)	*X00–X09*	Exposure to smoke, fire and flames
		X10–X19	Contact with heat and hot substances
		X20–X29	Contact with venomous animals and plants
		X30–X39	Exposure to forces of nature
		X40–X49	Accidental poisoning by and exposure to noxious substances
		X50–X57	Overexertion, travel and privation
		X58–X59	Accidental exposure to other and unspecified factors
4	External causes of injuries caused by human and other external causes	*X60–X84*	Intentional self-harm
		X85–Y09	Assault
		Y10–Y34	Event of undetermined intent
		Y35–Y36	Legal intervention and operations of war
		Y40–Y84	Complications of medical and surgical care
		Y40–Y59	...Drugs, medicaments and biological substances causing adverse effects in therapeutic use
		Y60–Y69	...Misadventures to patients during surgical and medical care
		Y70–Y82	...Medical devices associated with adverse incidents in diagnostic and therapeutic use

Previous **table** continued.

Nr	Name of the group	Range of codes	Meaning
		Y83–Y84	…Surgical and other medical procedures as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure
		Y85–Y89	Sequelae of external causes of morbidity and mortality
		Y90–Y98	Supplementary factors related to causes of morbidity and mortality classified elsewhere

Annex 3. Peculiarities of data from the Health Information System and the Health Insurance Fund

Annex 3.a. Peculiarities of data from the Health Insurance Fund



Estonian Health Insurance Fund

Much more data: from one treatment case there might be multiple documentations

Collecting data to get a better insight into health insurance

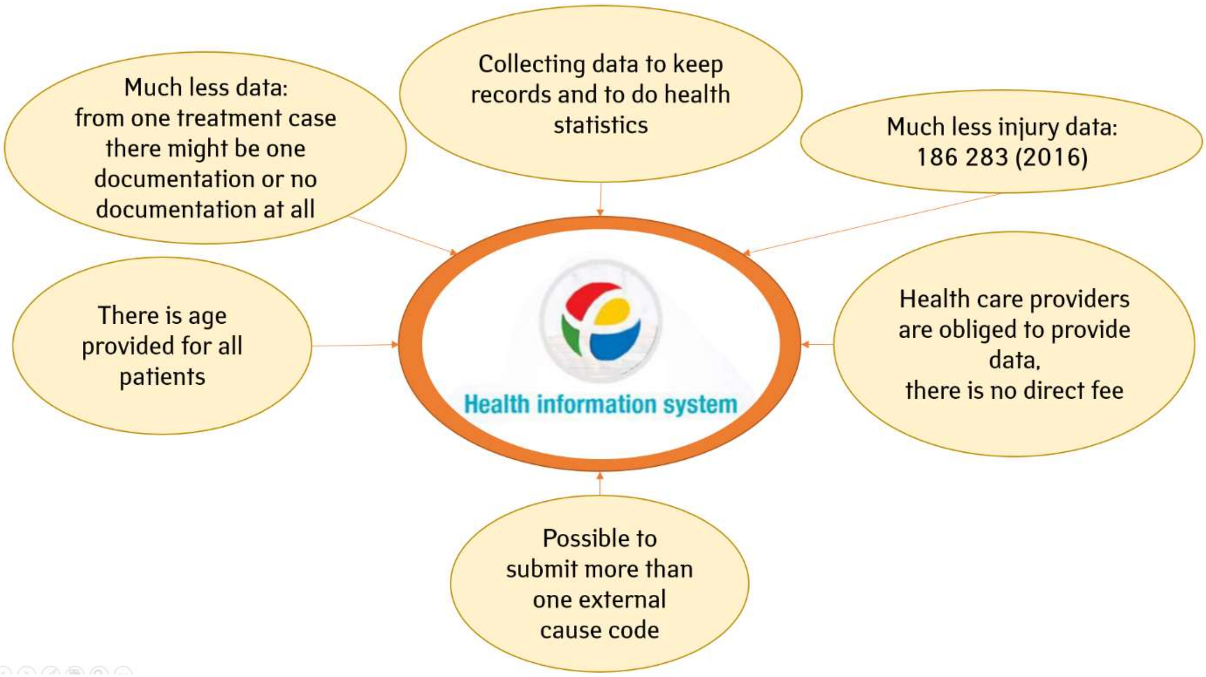
Much more injury data: 294 744 (2016)

There is no age provided for uninsured patient

Health care providers are obliged to provide data, they get fee based on invoices

Possible to submit only one external cause code

Annex 3.b. Peculiarities of data from the Health Information System (e-Health)



Health information system

Much less data: from one treatment case there might be one documentation or no documentation at all

Collecting data to keep records and to do health statistics

Much less injury data: 186 283 (2016)

There is age provided for all patients

Health care providers are obliged to provide data, there is no direct fee

Possible to submit more than one external cause code

Annex 4. Combinations of linking

Case	No	Variables, that are allowed to be different	Connected
1	0	**Everything matches** (with age)	11 361
	1	Different (or missing) 6th symbol of main diagnosis code	2
	2	Different (or missing) 6th symbol of external cause code	23
	3	Different (or missing) 6th symbol of both codes	0
2	0	***age* missing**	382
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	2
	3	Different (or missing) 6th symbol of both codes	0
3	0	***insurance*** (with age)	148
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
4	0	***insurance*; *age* missing**	9
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
5	0	***inevitable*** (with age)	119 245
	1	Different (or missing) 6th symbol of main diagnosis code	5
	2	Different (or missing) 6th symbol of external cause code	452
	3	Different (or missing) 6th symbol of both codes	0
6	0	***inevitable*; *age* missing**	3 459
	1	Different (or missing) 6th symbol of main diagnosis code	1
	2	Different (or missing) 6th symbol of external cause code	35
	3	Different (or missing) 6th symbol of both codes	0
7	0	***begin*** (with age)	10
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
8	0	***begin*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
9	0	***end*** (with age)	61
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
10	0	***end*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
11	0	***county*** (with age)	131
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	1
	3	Different (or missing) 6th symbol of both codes	0
12	0	***county*; *age* missing**	37
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
13	0	***type*** (with age)	14
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
14	0	***type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
15	0	***insurance, inevitable*** (with age)	372
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
16	0	***insurance, inevitable*; *age* missing**	34
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	2
	3	Different (or missing) 6th symbol of both codes	0
17	0	***insurance, begin*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
18	0	***insurance, begin*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
19	0	***insurance, end*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
20	0	***insurance, end*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
21	0	***insurance, county*** (with age)	2
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
22	0	***insurance, county*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
23	0	***insurance, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
24	0	***insurance, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
25	0	***inevitable, begin*** (with age)	4 148
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
26	0	***inevitable, begin*; *age* missing**	113
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
27	0	***inevitable, end*** (with age)	3 784
	1	Different (or missing) 6th symbol of main diagnosis code	1
	2	Different (or missing) 6th symbol of external cause code	3
	3	Different (or missing) 6th symbol of both codes	0
28	0	***inevitable, end*; *age* missing**	101
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	1
	3	Different (or missing) 6th symbol of both codes	0
29	0	***inevitable, county*** (with age)	933
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	5
	3	Different (or missing) 6th symbol of both codes	0
30	0	***inevitable, county*; *age* missing**	167
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	4
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
31	0	***inevitable, type*** (with age)	9
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
32	0	***inevitable, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
33	0	***begin, end* (max difference 30 days)** (with age)	3
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	1
34	0	***begin, end*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
35	0	***begin, county*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
36	0	***begin, county*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
37	0	***begin, type*** (with age)	1
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
38	0	***begin, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
39	0	***end, county*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
40	0	***end, county*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
41	0	***end, type*** (with age)	4
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
42	0	***end, type***; ***age*** **missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
43	0	***county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
44	0	***county, type***; ***age*** **missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
45	0	***insurance, inevitable, begin*** (with age)	11
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
46	0	***insurance, inevitable, begin***; ***age*** **missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
47	0	***insurance, inevitable, end*** (with age)	18
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
48	0	***insurance, inevitable, end***; ***age*** **missing**	2
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
49	0	***insurance, inevitable, county*** (with age)	26
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
50	0	***insurance, inevitable, county***; ***age*** **missing**	2
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
51	0	***insurance, inevitable, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
52	0	***insurance, inevitable, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
53	0	***insurance, begin, end*** (with age)	1
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
54	0	***insurance, begin, end*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
55	0	***insurance, begin, county*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
56	0	***insurance, begin, county*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
57	0	***insurance, begin, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
58	0	***insurance, begin, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
59	0	***insurance, end, county*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
60	0	***insurance, end, county*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
61	0	***insurance, end, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
62	0	***insurance, end, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
63	0	***insurance, county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
64	0	***insurance, county, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
65	0	***inevitable, begin, end*** (with age)	375
	1	Different (or missing) 6th symbol of main diagnosis code	6
	2	Different (or missing) 6th symbol of external cause code	18
	3	Different (or missing) 6th symbol of both codes	0
66	0	***inevitable, begin, end*; *age* missing**	32
	1	Different (or missing) 6th symbol of main diagnosis code	1
	2	Different (or missing) 6th symbol of external cause code	4
	3	Different (or missing) 6th symbol of both codes	1
67	0	***inevitable, begin, county*** (with age)	33
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
68	0	***inevitable, begin, county*; *age* missing**	8
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
69	0	***inevitable, begin, type*** (with age)	1
	1	Different (or missing) 6th symbol of main diagnosis code	1
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
70	0	***inevitable, begin, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
71	0	***inevitable, end, county*** (with age)	19
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
72	0	***inevitable, end, county*; *age* missing**	6
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
73	0	***inevitable, end, type*** (with age)	8
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
74	0	***inevitable, end, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
75	0	***inevitable, county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
76	0	***inevitable, county, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
77	0	***begin, end, county*** (with age)	1
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
78	0	***begin, end, county*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
79	0	***begin, end, type*** (with age)	11
	1	Different (or missing) 6th symbol of main diagnosis code	2
	2	Different (or missing) 6th symbol of external cause code	3
	3	Different (or missing) 6th symbol of both codes	0
80	0	***begin, end, type*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
81	0	***begin, county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
82	0	***begin, county, type***; ***age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
83	0	***end, county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
84	0	***end, county, type***; ***age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
85	0	***insurance, inevitable, begin, end*** (with age)	13
	1	Different (or missing) 6th symbol of main diagnosis code	1
	2	Different (or missing) 6th symbol of external cause code	2
	3	Different (or missing) 6th symbol of both codes	0
86	0	***insurance, inevitable, begin, county*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
87	0	***insurance, inevitable, begin, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
88	0	***insurance, inevitable, end, county*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
89	0	***insurance, inevitable, end, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
90	0	***begin, end, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
91	0	***insurance, begin, end, county*** (with age)	13
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	2
	3	Different (or missing) 6th symbol of both codes	0
92	0	***insurance, begin, end, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
93	0	***insurance, begin, county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
94	0	***insurance, end, county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
95	0	***inevitable, begin, end, county*** (with age)	15
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	3
	3	Different (or missing) 6th symbol of both codes	0
96	0	***inevitable, begin, end, type*** (with age)	23
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	1
	3	Different (or missing) 6th symbol of both codes	1
97	0	***inevitable, begin, county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
98	0	***inevitable, end, county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
99	0	***begin, end, county, type*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
100	0	***causecode*** (with age)	255
	1	Different (or missing) 6th symbol of main diagnosis code	0
101	0	***causecode; age*** **missing**	17
	1	Different (or missing) 6th symbol of main diagnosis code	0
102	0	***dgn*** (with age)	1
	2	Different (or missing) 6th symbol of external cause code	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
103	0	***dgn*; *age* missing**	4
	2	Different (or missing) 6th symbol of external cause code	0
104	0	***dgn, inevitable*** (with age)	19
	2	Different (or missing) 6th symbol of external cause code	0
105	0	***dgn, inevitable*; *age* missing**	0
	2	Different (or missing) 6th symbol of external cause code	0
106	0	***causecode, inevitable*** (with age)	9 648
	1	Different (or missing) 6th symbol of main diagnosis code	1
107	0	***causecode, inevitable*; *age* missing**	251
	1	Different (or missing) 6th symbol of main diagnosis code	0
108	0	***different (or missing) last 3 symbols of dgn*** (with age)	20
	2	Different (or missing) 6th symbol of external cause code	0
109	0	***different (or missing) last 3 symbols of dgn*; *age* missing**	0
	2	Different (or missing) 6th symbol of external cause code	0
110	0	***different (or missing) last 3 symbols of causecode*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
111	0	***different (or missing) last 3 symbols of causecode*; *age* missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
112	0	***dgn* and *different (or missing) last 3 symbols of causecode*** (with age)	2
113	0	***dgn* and *different (or missing) last 3 symbols of causecode*; *age* missing**	0
114	0	***inevitable, different (or missing) last 3 symbols of dgn*** (with age)	357
	2	Different (or missing) 6th symbol of external cause code	50
115	0	***inevitable, different (or missing) last 3 symbols of dgn*; *age* missing**	25
	2	Different (or missing) 6th symbol of external cause code	2
116	0	***inevitable, different (or missing) last 3 symbols of causecode*** (with age)	2
	1	Different (or missing) 6th symbol of main diagnosis code	0
117	0	***inevitable, different (or missing) last 3 symbols of causecode*; *age* missing**	3
	1	Different (or missing) 6th symbol of main diagnosis code	0
118	0	***inevitable, dgn* and *different (or missing) last 3 symbols of causecode*** (with age)	127
119	0	***inevitable, dgn* and *different (or missing) last 3 symbols of causecode*; *age* missing**	0
120	0	***begin, different (or missing) last 3 symbols of dgn*** (with age)	0
	2	Different (or missing) 6th symbol of external cause code	0
121	0	***begin, different (or missing) last 3 symbols of dgn*; *age* missing**	0
	2	Different (or missing) 6th symbol of external cause code	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
122	0	***end, different (or missing) last 3 symbols of dgn*** (with age)	0
	2	Different (or missing) 6th symbol of external cause code	0
123	0	***end, different (or missing) last 3 symbols of dgn; age*** **missing**	0
	2	Different (or missing) 6th symbol of external cause code	0
124	0	***begin, end, different (or missing) last 3 symbols of dgn*** (with age)	0
	2	Different (or missing) 6th symbol of external cause code	1
125	0	***begin, end, different (or missing) last 3 symbols of dgn; age*** **missing**	0
	2	Different (or missing) 6th symbol of external cause code	0
126	0	***begin, different (or missing) last 3 symbols of causecode*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
127	0	***begin, different (or missing) last 3 symbols of causecode; age*** **missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
128	0	***end, different (or missing) last 3 symbols of causecode*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
129	0	***end, different (or missing) last 3 symbols of causecode; age*** **missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
130	0	***begin, end, different (or missing) last 3 symbols of causecode*** (with age)	1
	1	Different (or missing) 6th symbol of main diagnosis code	0
131	0	***begin, end, different (or missing) last 3 symbols of causecode; age*** **missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
132	0	***begin, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	1
133	0	***begin, dgn*** **and** ***different (or missing) last 3 symbols of causecode; age*** **missing**	0
134	0	***end, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	0
135	0	***end, dgn*** **and** ***different (or missing) last 3 symbols of causecode; age*** **missing**	0
136	0	***begin, end, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	1
137	0	***begin, end, dgn*** **and** ***different (or missing) last 3 symbols of causecode; age*** **missing**	0
138	0	***inevitable, begin, different (or missing) last 3 symbols of dgn*** (with age)	4
	2	Different (or missing) 6th symbol of external cause code	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
139	0	***inevitable, begin, different (or missing) last 3 symbols of dgn; age*** **missing**	0
	2	Different (or missing) 6th symbol of external cause code	0
140	0	***inevitable, end, different (or missing) last 3 symbols of dgn*** (with age)	28
	2	Different (or missing) 6th symbol of external cause code	0
141	0	***inevitable, end, different (or missing) last 3 symbols of dgn; age*** **missing**	1
	2	Different (or missing) 6th symbol of external cause code	0
142	0	***inevitable, begin, end, different (or missing) last 3 symbols of dgn*** (with age)	33
	2	Different (or missing) 6th symbol of external cause code	8
143	0	***inevitable, begin, end, different (or missing) last 3 symbols of dgn; age*** **missing**	2
	2	Different (or missing) 6th symbol of external cause code	0
144	0	***inevitable, begin, different (or missing) last 3 symbols of causecode*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
145	0	***inevitable, begin, different (or missing) last 3 symbols of causecode; age*** **missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
146	0	***inevitable, end, different (or missing) last 3 symbols of causecode*** (with age)	30
	1	Different (or missing) 6th symbol of main diagnosis code	0
147	0	***inevitable, end, different (or missing) last 3 symbols of causecode; age*** **missing**	4
	1	Different (or missing) 6th symbol of main diagnosis code	0
148	0	***inevitable, begin, end, different (or missing) last 3 symbols of causecode*** (with age)	50
	1	Different (or missing) 6th symbol of main diagnosis code	1
149	0	***inevitable, begin, end, different (or missing) last 3 symbols of causecode; age*** **missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
150	0	***inevitable, begin, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	1
151	0	***inevitable, begin, dgn*** **and** ***different (or missing) last 3 symbols of causecode; age*** **missing**	0
152	0	***inevitable, end, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	4
153	0	***inevitable, end, dgn*** **and** ***different (or missing) last 3 symbols of causecode; age*** **missing**	
154	0	***inevitable, begin, end, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	35
155	0	***inevitable, begin, end, dgn*** **and** ***different (or missing) last 3 symbols of causecode; age*** **missing**	1

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
156	0	***age*** ±1, ***begin*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
157	0	***age*** ±1, ***begin, end*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
158	0	***age*** ±1, ***begin, different (or missing) last 3 symbols of dgn*** (with age)	25
	2	Different (or missing) 6th symbol of external cause code	0
159	0	***age*** ±1, ***begin, end, different (or missing) last 3 symbols of dgn*** (with age)	0
	2	Different (or missing) 6th symbol of external cause code	0
160	0	***age*** ±1, ***begin, different (or missing) last 3 symbols of causecode*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
161	0	***age*** ±1, ***begin, end, different (or missing) last 3 symbols of causecode*** (with age)	1
	1	Different (or missing) 6th symbol of main diagnosis code	0
162	0	***age*** ±1, ***begin, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	0
163	0	***age*** ±1, ***begin, end, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	0
164	0	***age*** ±1, ***inevitable, begin*** (with age)	484
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	2
	3	Different (or missing) 6th symbol of both codes	0
165	0	***age*** ±1, ***inevitable, begin, end*** (with age)	7
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
166	0	***age*** ±1, ***inevitable, begin, different (or missing) last 3 symbols of dgn*** (with age)	4
	2	Different (or missing) 6th symbol of external cause code	0
167	0	***age*** ±1, ***inevitable, begin, end, different (or missing) last 3 symbols of dgn*** (with age)	237
	2	Different (or missing) 6th symbol of external cause code	9
168	0	***age*** ±1, ***inevitable, begin, different (or missing) last 3 symbols of causecode*** (with age)	3
	1	Different (or missing) 6th symbol of main diagnosis code	0
169	0	***age*** ±1, ***inevitable, begin, end, different (or missing) last 3 symbols of causecode*** (with age)	63
	1	Different (or missing) 6th symbol of main diagnosis code	18
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
170	0	***age*** ±1, ***inevitable, begin, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	0
171	0	***age*** ±1, ***inevitable, begin, end, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	9
172	0	***county, different (or missing) last 3 symbols of dgn*** (with age)	1
	2	Different (or missing) 6th symbol of external cause code	0
173	0	***county, different (or missing) last 3 symbols of dgn; age*** **missing**	0
	2	Different (or missing) 6th symbol of external cause code	0
174	0	***county, different (or missing) last 3 symbols of causecode*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
175	0	***county, different (or missing) last 3 symbols of causecode; age*** **missing**	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
176	0	***county, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	0
177	0	***county, dgn*** **and** ***different (or missing) last 3 symbols of causecode; age*** **missing**	0
178	0	***inevitable, county, different (or missing) last 3 symbols of dgn*** (with age)	4
	2	Different (or missing) 6th symbol of external cause code	1
179	0	***inevitable, county, different (or missing) last 3 symbols of dgn; age*** **missing**	0
	2	Different (or missing) 6th symbol of external cause code	1
180	0	***inevitable, county, different (or missing) last 3 symbols of causecode*** (with age)	14
	1	Different (or missing) 6th symbol of main diagnosis code	0
181	0	***inevitable, county, different (or missing) last 3 symbols of causecode; age*** **missing**	5
	1	Different (or missing) 6th symbol of main diagnosis code	0
182	0	***inevitable, county, dgn*** **and** ***different (or missing) last 3 symbols of causecode*** (with age)	1
183	0	***inevitable, county, dgn*** **and** ***different (or missing) last 3 symbols of causecode; age*** **missing**	1
184	0	***age*** ±1, ***county, begin*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
185	0	***age*** ±1, ***county, begin, end*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Previous **table** continued.

Case	No	Variables, that are allowed to be different	Connected
186	0	***age*** ±1, ***inevitable, county, begin*** (with age)	8
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0
187	0	***age*** ±1, ***inevitable, county, begin, end*** (with age)	0
	1	Different (or missing) 6th symbol of main diagnosis code	0
	2	Different (or missing) 6th symbol of external cause code	0
	3	Different (or missing) 6th symbol of both codes	0

Annex 5. Creating new variables for imputation (software: R)

```
library(readstata13)
dat <- read.dta13("C://data.dta")

# Create begin_month variable:
dat$begin_month <- format(dat$begin, "%m")
dat$begin_month=factor(as.numeric(dat$begin_month))

# Create sum_group variable:
dat$sum_group <- cut(dat$sum, breaks = c(-Inf, 100, 200, Inf),
    labels=c("...-100", "101-200", "201-..."))
table(dat$sum_group)

# Create days_group variable:
dat$days_group <- cut(as.numeric(dat$end-dat$begin), breaks = c(-Inf, 0, 5, 10, Inf),
    labels=c("<=0", "0-5", "6-10", ">10"))
table(dat$days_group)

# Create dgn_group variable:
dat$dgn_group <- 4
dat$dgn_group[substr(dat$dgn,0,1) == "S" & as.numeric(substr(dat$dgn,2,3)) < 40] <- 1
dat$dgn_group[substr(dat$dgn,0,1) == "S" & as.numeric(substr(dat$dgn,2,3)) >= 40 &
    as.numeric(substr(dat$dgn,2,3)) < 70] <- 2
dat$dgn_group[substr(dat$dgn,0,1) == "S" & as.numeric(substr(dat$dgn,2,3)) >= 70] <- 3
dat$dgn_group=factor(as.numeric(dat$dgn_group))
table(dat$dgn_group)

# Create causecode_group variable:
dat$causecode_group <- NA
dat$causecode_group[substr(dat$causecode,0,1) == "V"] <- 1
dat$causecode_group[substr(dat$causecode,0,1) == "W"] <- 2
dat$causecode_group[substr(dat$causecode,0,1) == "Y"] <- 4
dat$causecode_group[substr(dat$causecode,0,1) == "X" &
    as.numeric(substr(dat$causecode,2,3)) < 60] <- 3
dat$causecode_group[substr(dat$causecode,0,1) == "X" &
    as.numeric(substr(dat$causecode,2,3)) >= 60] <- 4
dat$causecode_group=factor(as.numeric(dat$causecode_group))
table(dat$causecode_group)

# Create familydoctor variable using ttocode:
dat$familydoctor <- 0
dat$familydoctor[substr(dat$ttocode,0,1) == "1"] <- 1
table(dat$familydoctor)

# Factorize some variables:
dat$type=factor(dat$type); table(dat$type)
dat$sex=factor(dat$sex); table(dat$sex)
```

Annex 6. Preliminary analysis of variables that will be imputed

Annex 6.a. Variable's *causecode_group* relationship with other variables

```
sum(is.na(dat$causecode_group)) # 9 missing values

#-----------------------

# causecode_group and type
tbl <- table(dat$causecode_group, dat$type); (tbl <- tbl[complete.cases(tbl),])
#      1    2
# 1  381   89
# 2 6207  425
# 3  594   80
# 4 1228  135
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 119.94, df = 3, p-value < 2.2e-16

# causecode_group and days_group
tbl <- table(dat$causecode_group, dat$days_group); (tbl <- tbl[complete.cases(tbl),])
#    <=0  0-5 6-10  >10
# 1  371   57   25   17
# 2 6029  418  104   81
# 3  576   58   15   25
# 4 1178  136   25   24
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 119.45, df = 9, p-value < 2.2e-16

# causecode_group and sum_group
tbl <- table(dat$causecode_group, dat$sum_group); (tbl <- tbl[complete.cases(tbl),])
#  ...-100 101-200 201-...
#1     286      51     133
#2    5520     503     609
#3     508      64     102
#4     777     272     314
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 570.48, df = 6, p-value < 2.2e-16

# causecode_group and begin_month
tbl <- table(dat$causecode_group, dat$begin_month); (tbl <- tbl[complete.cases(tbl),])
#     1   2   3   4   5   6   7   8   9  10  11  12
# 1  18  18  22  39  58  63  80  64  41  25  21  21
# 2 520 520 505 457 659 665 751 682 487 478 432 476
# 3  68  46  41  56  60  58  95  63  48  52  47  40
# 4 123 100 118 111 107 123 140 124 116 108  89 104
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 98.359, df = 33, p-value = 2.033e-08

# causecode_group and familydoctor
tbl <- table(dat$causecode_group, dat$familydoctor) ; (tbl <- tbl[complete.cases(tbl),])
#      0    1
# 1  341  129
# 2 4315 2317
# 3  440  234
# 4  862  501
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 13.556, df = 3, p-value = 0.003576

# causecode_group and dgn_group
tbl <- table(dat$causecode_group, dat$dgn_group); (tbl <- tbl[complete.cases(tbl),])
#      1    2    3    4
# 1  163  155  118   34
# 2 1609 2232 1836  955
# 3   73   97  112  392
# 4  747  126   45  445
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 1791.5, df = 9, p-value < 2.2e-16
```

Annex 6.b. Variable's *sex* relationship with other variables

```
sum(is.na(dat$sex)) # 37 missing values

#-----------------------

# sex and type
tbl <- table(dat$sex, dat$type); (tbl <- tbl[complete.cases(tbl),])
#      1    2
# 1 6995  613
# 2 1393  110
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 0.83883, df = 1, p-value = 0.3597

# sex and days_group
tbl <- table(dat$sex, dat$days_group); (tbl <- tbl[complete.cases(tbl),])
#    <=0  0-5 6-10  >10
# 1 6788  545  146  129
# 2 1343  121   22   17
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 5.2603, df = 3, p-value = 0.1537

# sex and sum_group
tbl <- table(dat$sex, dat$sum_group); (tbl <- tbl[complete.cases(tbl),])
#   ...-100 101-200 201-...
# 1    5867     763     978
# 2    1212     119     172
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 9.764, df = 2, p-value = 0.007582

# sex and begin_month
tbl <- table(dat$sex, dat$begin_month); (tbl <- tbl[complete.cases(tbl),])
#     1   2   3   4   5   6   7   8   9  10  11  12
# 1 599 572 573 559 756 743 861 775 591 565 487 527
# 2 130 110 110 103 125 163 203 152  98  97  98 114
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 16.562, df = 11, p-value = 0.1215

# sex and familydoctor
tbl <- table(dat$sex, dat$familydoctor); (tbl <- tbl[complete.cases(tbl),])
#      0    1
# 1 4944 2664
# 2  984  519
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 0.10934, df = 1, p-value = 0.7409

# sex and dgn_group
tbl <- table(dat$sex, dat$dgn_group); (tbl <- tbl[complete.cases(tbl),])
#      1    2    3    4
# 1 2208 2127 1674 1599
# 2  369  476  429  229
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 59.087, df = 3, p-value = 9.21e-13

# sex and causecode_group
tbl <- table(dat$sex, dat$causecode_group); (tbl <- tbl[complete.cases(tbl),])
#      1    2    3    4
# 1  406 5504  552 1139
# 2   61 1101  118  222
```

```
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 4.7798, df = 3, p-value = 0.1887
```

Annex 6.c. Variable’s *age* relationship with other variables

```
sum(is.na(dat$age)) # 3515 missing values

#-----------------------

# T-tests:

# age and type
tbl <- cbind(dat$age, dat$type); tbl <- tbl[complete.cases(tbl),]
colnames(tbl) <- c("age", "type"); tbl <- as.data.frame(tbl)
t.test(tbl$age[tbl$type==1], tbl$age[tbl$type==2])
# Welch Two Sample t-test
# data:  tbl$age[tbl$type == 1] and
#   tbl$age[tbl$type == 2]
# t = -8.5011, df = 828.1, p-value < 2.2e-16
# alternative hypothesis:
#   true difference in means is not equal to 0
# 95 percent confidence interval:
#   -5.617629 -3.510119
# sample estimates:
#   mean of x mean of y
#    36.33657  40.90044

# age and familydoctor
tbl <- cbind(dat$age, dat$familydoctor); tbl <- tbl[complete.cases(tbl),]
colnames(tbl) <- c("age", "familydoctor"); tbl <- as.data.frame(tbl)
t.test(tbl$age[tbl$familydoctor==0], tbl$age[tbl$familydoctor==1])
# Welch Two Sample t-test
# data:  tbl$age[tbl$familydoctor == 0] and
#   tbl$age[tbl$familydoctor == 1]
# t = 0.17912, df = 4923.1, p-value = 0.8579
# alternative hypothesis:
#   true difference in means is not equal to 0
# 95 percent confidence interval:
#    -0.5696324  0.6841880
# sample estimates:
#   mean of x mean of y
#    36.91256  36.85528

# age and sex
tbl <- cbind(dat$age, dat$sex); tbl <- tbl[complete.cases(tbl),]
colnames(tbl) <- c("age", "sex"); tbl <- as.data.frame(tbl)
t.test(tbl$age[tbl$sex==1], tbl$age[tbl$sex==2])
# Welch Two Sample t-test
# data:  tbl$age[tbl$sex == 1] and
    tbl$age[tbl$sex == 2]
# t = -2.9694, df = 1162.6, p-value = 0.003045
# alternative hypothesis:
#   true difference in means is not equal to 0
# 95 percent confidence interval:
#    -2.4412552 -0.4986938
# sample estimates:
#   mean of x mean of y
#    36.64933  38.11931

#-----------------------


# Chi-squared tests:

# Let create a new variable age_group for these tests:
summary(dat$age)
dat$age_group <- cut(dat$age, breaks = c(-Inf, 20, 40, 60, Inf),
    labels=c("0-20", "21-40", "41-60", "61-..."))
summary(dat$age_group)

# age_group and days_group
tbl <- table(dat$age_group, dat$days_group); (tbl <- tbl[complete.cases(tbl),])
#          <=0  0-5 6-10  >10
#   0-20   228   17    5    3
#  21-40  2978  278   50   38
#  41-60  1561  190   86   71
# 61-...    93   16    9   10
chisq.test(tbl, simulate.p.value = TRUE)
# Pearson's Chi-squared test with simulated p-value
    (based on 2000 replicates)
# data:  tbl
# X-squared = 126.25, df = NA, p-value = 0.0004998

# age_group and sum_group
tbl <- table(dat$age_group, dat$sum_group); (tbl <- tbl[complete.cases(tbl),])
#       ...-100 101-200 201-...
#   0-20    199      23      31
#  21-40   2592     279     473
#  41-60   1278     207     423
# 61-...     67      17      44
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 108.38, df = 6, p-value < 2.2e-16

# age_group and begin_month
tbl <- table(dat$age_group, dat$begin_month); (tbl <- tbl[complete.cases(tbl),])
#           1    2    3    4    5    6    7    8    9   10   11   12
#   0-20   20   15   27   16   25   20   48   21   13   24   13   11
#  21-40  275  247  273  234  332  338  380  351  240  249  216  209
#  41-60  165 135 128 134 164 174 206 190 157 162 138 155
# 61-...   14  19  10   6  15   6  14   8   7  17   7   5
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 67.922, df = 33, p-value = 0.0003267

# age_group and dgn_group
tbl <- table(dat$age_group, dat$dgn_group); (tbl <- tbl[complete.cases(tbl),])
#           1    2    3    4
#   0-20   63   90   64   36
#  21-40  985  904  742  713
#  41-60  557  450  435  466
# 61-...   43   39   21   25
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 33.93, df = 9, p-value = 9.19e-05

# age_group and causecode_group
tbl <- table(dat$age_group, dat$causecode_group); (tbl <- tbl[complete.cases(tbl),])
#            1    2    3    4
#  0-20     26  176   24   27
#  21-40   176 2361  211  592
#  41-60   102 1366  165  275
# 61-...     4   99   11   14
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 54.69, df = 9, p-value = 1.395e-08

dat$age_group <- NULL
```

Annex 7. Imputation (Software: R)

```
library(VIM)
library(dplyr)
#-----------------------
#-----------------------
#-----------------------
# GENERAL RANDOM HOT-DECK METHOD:
#-----------------------
# causecode_group imputation:
sum(is.na(dat$causecode_group)) # 9 missing values
hot_deck <- hotdeck(dat, variable=c("causecode_group"))
sum(is.na(hot_deck$causecode_group)) # 0 missing values
#-----------------------
# sex imputation:
sum(is.na(hot_deck$sex))  # 37 missing values
hot_deck <- hotdeck(hot_deck, variable=c("sex"))
sum(is.na(hot_deck$sex)) # 0 missing values
#-----------------------
# age imputation:
sum(is.na(hot_deck$age))  # 3515 missing values
hot_deck <- hotdeck(hot_deck, variable=c("age"))
sum(is.na(hot_deck$age)) # 0 missing values
#-----------------------
#-----------------------
# NEAREST NEIGHBOUR METHOD:
#-----------------------
# causecode_group imputation:
sum(is.na(dat$causecode_group)) # 9 missing values
knn<- kNN(dat, variable=c("causecode_group"), dist_var = c("sum_group", "begin_month",
                       "familydoctor", "days_group", "dgn_group", "type"), k=1)
sum(is.na(knn$causecode_group)) # 0 missing values
#-----------------------
# sex imputation:
sum(is.na(knn$sex))  # 37 missing values
knn <- kNN(knn, variable=c("sex"), dist_var = c("sum_group", "dgn_group"), k=1)
sum(is.na(knn$sex)) # 0 missing values
#-----------------------
# age imputation:
sum(is.na(knn$age)) # 3515 missing values
knn <- kNN(knn, variable=c("age"), dist_var = c("sum_group", "begin_month", "sex","days_group",
                        "dgn_group", "type", "causecode_group"), k=1)
sum(is.na(knn$age)) # 0 missing values
#-----------------------
#-----------------------
# RANDOM HOT-DECK METHOD WITHIN CLASSES AND NEAREST NEIGHBOUR METHOD:
#-----------------------
# causecode_group imputation:
sum(is.na(dat$causecode_group)) # 9 missing values
```

```
# Lets examine if there are groups, where there are only missing values for causecode_group:
data <- dat; data$missing <- 0; data$missing[is.na(data$causecode_group)] <- 1
data <- transform(data, group = as.numeric(factor(paste(sum_group, days_group, dgn_group,
        begin_month, type, familydoctor, sep=""))))
max(data$group) # 692 groups formed
# Lets determine group value:
data <- group_by(data, group)
data <- mutate(data, miss = sum(missing), all = length(missing))
data$empty <- 0
data$empty[data$miss == data$all] <- 1
sum(data$empty) # no empty groups

# So we only use random hot-deck method within class:
hot_knn <- hotdeck(data, variable=c("causecode_group"), domain_var = c("sum_group",
                         "begin_month", "familydoctor","days_group", "dgn_group", "type"))
sum(is.na(hot_knn$causecode_group)) # 0 missing values
hot_knn$empty <- NULL; hot_knn$all <- NULL; hot_knn$miss <- NULL; hot_knn$group <- NULL
hot_knn$missing <- NULL
#-----------------------
# sex imputation:
sum(is.na(hot_knn$sex))  # 37 missing values

# Lets examine if there are groups, where there are only missing values for sex:
data <- hot_knn; data$missing <- 0; data$missing[is.na(data$sex)] <- 1
data <- transform(data, group = as.numeric(factor(paste(sum_group, dgn_group, sep=""))))
max(data$group) # 12 groups formed
# Lets determine group value:
data <- group_by(data, group)
data <- mutate(data, miss = sum(missing), all = length(missing))
data$empty <- 0
data$empty[data$miss == data$all] <- 1
sum(data$empty) # no empty groups

# So we only use random hot-deck method within class:
hot_knn <- hotdeck(data, variable=c("sex"), domain_var = c("sum_group", "dgn_group"))
sum(is.na(hot_knn$sex))  # 0 missing values
hot_knn$empty <- NULL; hot_knn$all <- NULL; hot_knn$miss <- NULL; hot_knn$group <- NULL
hot_knn$missing <- NULL
#-----------------------
# age imputation:
sum(is.na(hot_knn$age)) # 3515 missing values

# Lets examine if there are groups, where there are only missing values for age:
data <- hot_knn; data$missing <- 0; data$missing[is.na(data$age)] <- 1
data <- transform(data, group = as.numeric(factor(paste(sum_group, days_group, dgn_group,
        begin_month, type, sex, causecode_group, sep=""))))
max(data$group) # 1226 groups formed
# Lets determine group value:
data <- group_by(data, group)
data <- mutate(data, miss = sum(missing), all = length(missing))
data$empty <- 0; data$empty[data$miss == data$all] <- 1
```

```
sum(data$empty) # 236 objects belong to the empty groups

hot_knn <- hotdeck(data, variable=c("age"), domain_var = c("sum_group", "begin_month",
                                  "sex","days_group", "dgn_group", "type", "causecode_group"))
sum(is.na(hot_knn$age)) # 0 missing values
# Lets see what value was imputed in empty groups
hot_knn$age[data$empty==1] # 1 is always assigned
hot_knn$age_imputed <- hot_knn$age_imp
hot_knn$age_imp <- NULL

# Lets impute in empty groups using nearest neighbour method:
hot_knn$age[data$empty==1] <- NA; sum(is.na(hot_knn$age))
hot_knn<- kNN(hot_knn, variable=c("age"), dist_var = c("sum_group", "begin_month",
                           "sex","days_group", "dgn_group", "type", "causecode_group"), k=1)
sum(is.na(knn$age)) # 0 missing values
# Lets see what value was imputed in empty groups
hot_knn$age[data$empty==1] # different ages were assigned
hot_knn$age_imp[hot_knn$age_imputed==TRUE]<-TRUE
hot_knn$empty <- NULL; hot_knn$all <- NULL; hot_knn$miss <- NULL; hot_knn$group <- NULL
hot_knn$missing <- NULL; hot_knn$age_imputed <- NULL
```

Annex 8. Analysis of the results (Software: R)

```
# Lets create variable to separate different methods:
hot_knn$method <- "hot-deck&knn"; hot_deck$method <- "hot-deck"
knn$method <- "knn"; dat$method <- "initial"
dat$age_imp <- NA; dat$sex_imp <- NA; dat$causecode_group_imp <- NA
#-----------------------
# causecode_group
#-----------------------
# Lets Connect the datasets for the analysis:
compare <- rbind(hot_deck, knn, hot_knn, subset(dat, !is.na(causecode_group)))
tbl <- table(compare$method, compare$causecode_group)
names(dimnames(tbl)) <- c(„method", "causecode_group")
summy <- apply(tbl,1,sum)
addmargins(round(sweep(tbl,1,summy,"/"),4)*100)
compare <- rbind(subset(hot_deck, causecode_group_imp==TRUE),
    subset(knn, causecode_group_imp==TRUE), subset(hot_knn, causecode_group_imp==TRUE))
tbl <- table(compare$method, compare$causecode_group)
names(dimnames(tbl)) <- c("method", "causecode_group")
tbl
#-----------------------
# sex
#-----------------------
# Lets connect the datasets for the analysis:
compare <- rbind(hot_deck, knn, hot_knn, subset(dat, !is.na(sex)))
# Lets check the weights:
round(table(compare$sex[compare$method=="hot-deck"])/
 (0.01*length(compare$sex[compare$method=="hot-deck"])), 2)
round(table(compare$sex[compare$method=="knn"])/
 (0.01*length(compare$sex[compare$method=="knn"])), 2)
round(table(compare$sex[compare$method=="hot-deck&knn"])/
  (0.01*length(compare$sex[compare$method=="hot-deck&knn"])), 2)
round(table(compare$sex[compare$method=="initial"])/
 (0.01*length(compare$sex[compare$method=="initial"])), 2)
compare <- rbind(subset(hot_deck, sex_imp==TRUE), subset(knn, sex_imp==TRUE),
    subset(hot_knn, sex_imp==TRUE))
tbl <- table(compare$method, compare$sex)
names(dimnames(tbl)) <- c("method", "sex")
tbl
#-----------------------
# age
#-----------------------
# Lets connect the datasets for the analysis:
compare <- rbind(hot_deck, knn, hot_knn, subset(dat, !is.na(age)))
round(mean(compare$age[compare$method=="hot-deck"]), 3)
round(sd(compare$age[compare$method=="hot-deck"]), 3)
min(compare$age[compare$method=="hot-deck"])
max(compare$age[compare$method=="hot-deck"])
median(compare$age[compare$method=="hot-deck"])
```

```
round(mean(compare$age[compare$method=="knn"]), 3)
round(sd(compare$age[compare$method=="knn"]), 3)
min(compare$age[compare$method=="knn"])
max(compare$age[compare$method=="knn"])
median(compare$age[compare$method=="knn"])

round(mean(compare$age[compare$method=="hot-deck&knn"]), 3)
round(sd(compare$age[compare$method=="hot-deck&knn"]), 3)
min(compare$age[compare$method=="hot-deck&knn"])
max(compare$age[compare$method=="hot-deck&knn"])
median(compare$age[compare$method=="hot-deck&knn"])

round(mean(compare$age[compare$method=="initial"]), 3)
round(sd(compare$age[compare$method=="initial"]), 3)
min(compare$age[compare$method=="initial"])
max(compare$age[compare$method=="initial"])
median(compare$age[compare$method=="initial"])

boxplot(age~method,data=compare)

compare <- rbind(subset(hot_deck, age_imp==TRUE), subset(knn, age_imp==TRUE),
    subset(hot_knn, age_imp==TRUE))
boxplot(age~method,data=compare)
```

Annex 9. Imputation when 70% of the data exists (Software: R)

Annex 9.a. Creating missing data

```
# Lets use only complete existing data:
data_try <- subset(dat, !is.na(causecode_group))
data_try <- subset(data_try, !is.na(sex)); data_try <- subset(data_try, !is.na(age))
data_try$age_imp <- NULL; data_try$sex_imp <- NULL; data_try$causecode_group_imp <- NULL
data_try$method <- NULL

data_try$ID <- seq.int(nrow(data_try))
# Lets see how much data should be left after deleting 30% of the age values:
delete_thirty <- floor(nrow(data_try)*0.3)
delete_rows_thirty <- sample(data_try$ID, delete_thirty); data_thirty <- data_try
data_thirty$age[is.element(data_thirty$ID, delete_rows_thirty)] <- NA
sum(!is.na(data_thirty$age)) # remains 3941 values
```

Annex 9.b. New data imputation

```
# age imputation:
sum(is.na(data_thirty$age)) # missing 1688 values
#-----------------------
#-----------------------
# GENERAL RANDOM HOT-DECK METHOD:
hot_thirty <- hotdeck(data_thirty, variable=c("age"))
sum(is.na(hot_thirty$age)) # 0 missing values
#-----------------------
#-----------------------
# NEAREST NEIGHBOUR METHOD:
knn_thirty <- kNN(data_thirty, variable=c("age"), dist_var = c("sum_group", "begin_month",
        "sex","days_group", "dgn_group", "type", "causecode_group"), k=1)
sum(is.na(knn_thirty$age)) # 0 missing values
#-----------------------
#-----------------------
# RANDOM HOT-DECK METHOD WITHIN CLASSES AND NEAREST NEIGHBOUR METHOD
# Lets examine if there are groups, where there are only missing values for age:
data <- data_thirty; data$missing <- 0; data$missing[is.na(data$age)] <- 1
data <- transform(data, group = as.numeric(factor(paste(sum_group, days_group, dgn_group,
        begin_month, type, sex, causecode_group, sep=""))))
max(data$group) # 1041 groups formed
# Lets determine group value:
data <- group_by(data, group); data <- mutate(data, miss = sum(missing), all = length(missing))
data$empty <- 0; data$empty[data$miss == data$all] <- 1

hot_knn_thirty <- hotdeck(data, variable=c("age"), domain_var = c("sum_group", "begin_month",
        "sex", "days_group", "dgn_group", "type", "causecode_group"))
sum(is.na(hot_knn_thirty$age)) # 0 missing values
# Lets see what value was imputed in empty groups
hot_knn_thirty$age[data$empty==1] # 1 is always assigned
```

```
hot_knn_thirty$age_imputed <- hot_knn_thirty$age_imp; hot_knn_thirty$age_imp <- NULL

# Lets impute in empty groups using nearest neighbour method:
hot_knn_thirty$age[data$empty==1] <- NA; sum(is.na(hot_knn_thirty$age))
hot_knn_thirty<- kNN(hot_knn_thirty, variable=c("age"), dist_var = c("sum_group",
        "begin_month", "sex", "days_group", "dgn_group", "type", "causecode_group"), k=1)
sum(is.na(hot_knn_thirty$age)) # 0 missing values
# Lets see what value was imputed in empty groups
hot_knn_thirty$age[data$empty==1] # different ages were assigned
hot_knn_thirty$age_imp[hot_knn_thirty$age_imputed==TRUE]<-TRUE
hot_knn_thirty$empty <- NULL; hot_knn_thirty$all <- NULL; hot_knn_thirty$miss <- NULL
hot_knn_thirty$group <- NULL; hot_knn_thirty$missing <- NULL
hot_knn_thirty$age_imputed <- NULL
```

Annex 9.c. Imputed data analysis

```
# Lets create variable to separate different methods:
hot_knn_thirty$method <- "hot-deck&knn"; hot_thirty$method <- "hot-deck"
knn_thirty$method <- "knn"; data_pr <- data_try; data_pr$method <- "actual"
data_pr$age_imp <- NA

# Lets connect the datasets for the analysis:
compare <- rbind(hot_thirty, knn_thirty, hot_knn_thirty, data_pr)
round(mean(compare$age[compare$method=="hot-deck"]), 3)
round(sd(compare$age[compare$method=="hot-deck"]), 3)
min(compare$age[compare$method=="hot-deck"])
max(compare$age[compare$method=="hot-deck"])
median(compare$age[compare$method=="hot-deck"])
round(mean(compare$age[compare$method=="knn"]), 3)
round(sd(compare$age[compare$method=="knn"]), 3)
min(compare$age[compare$method=="knn"])
max(compare$age[compare$method=="knn"])
median(compare$age[compare$method=="knn"])
round(mean(compare$age[compare$method=="hot-deck&knn"]), 3)
round(sd(compare$age[compare$method=="hot-deck&knn"]), 3)
min(compare$age[compare$method=="hot-deck&knn"])
max(compare$age[compare$method=="hot-deck&knn"])
median(compare$age[compare$method=="hot-deck&knn"])
round(mean(compare$age[compare$method=="actual"]), 3)
round(sd(compare$age[compare$method=="actual"]), 3)
min(compare$age[compare$method=="actual"])
max(compare$age[compare$method=="actual"])
median(compare$age[compare$method=="actual"])

boxplot(age~method,data=compare)

# Lets examine age differences:
data_trying <- subset(data_pr, is.element(data_thirty$ID, delete_rows_thirty))
hot_three_compare <- merge(subset(hot_thirty, age_imp==TRUE), data_trying,by="ID")
hot_three_compare$age_difference <- hot_three_compare$age.x-hot_three_compare$age.y
```

```
round(mean(hot_three_compare$age_difference), 3)
round(sd(hot_three_compare$age_difference), 3)
round(min(hot_three_compare$age_difference), 3)
round(max(hot_three_compare$age_difference), 3)
round(median(hot_three_compare$age_difference), 3)
knn_three_compare <- merge(subset(knn_thirty, age_imp==TRUE), data_trying,by="ID")
knn_three_compare$age_difference <- knn_three_compare$age.x-knn_three_compare$age.y
round(mean(knn_three_compare$age_difference), 3)
round(sd(knn_three_compare$age_difference), 3)
round(min(knn_three_compare$age_difference), 3)
round(max(knn_three_compare$age_difference), 3)
round(median(knn_three_compare$age_difference), 3)
hot_knn_three_compare <- merge(subset(hot_knn_thirty, age_imp==TRUE), data_trying, by="ID")
hot_knn_three_compare$age_difference <-
         hot_knn_three_compare$age.x-hot_knn_three_compare$age.y
round(mean(hot_knn_three_compare$age_difference), 3)
round(sd(hot_knn_three_compare$age_difference), 3)
round(min(hot_knn_three_compare$age_difference), 3)
round(max(hot_knn_three_compare$age_difference), 3)
round(median(hot_knn_three_compare$age_difference), 3)

compare <- rbind(hot_three_compare, knn_three_compare, hot_knn_three_compare)
boxplot(age_difference~method.x,data=compare)
```

Annex 9.d. Simulation of imputation and the analysis (imputation using same initial data in all steps)

```
hot_vec_means <- c(mean(hot_thirty$age)); knn_vec_means <- c(mean(knn_thirty$age))
hot_knn_vec_means <- c(mean(hot_knn_thirty$age))

# Lets do the imputation 99 times more:
for (i in 1:99){
    # GENERAL RANDOM HOT-DECK METHOD:
    hot_thirty_sim <- hotdeck(data_thirty, variable=c("age"))
    hot_vec_means <- append(hot_vec_means, mean(hot_thirty_sim$age))
    #-----------------------
    # NEAREST NEIGHBOUR METHOD:
    knn_thirty_sim <- kNN(data_thirty, variable=c("age"), dist_var = c("sum_group",
            "begin_month", "sex", "days_group", "dgn_group", "type", "causecode_group"), k=1)
    knn_vec_means <- append(knn_vec_means, mean(knn_thirty_sim$age))
    #-----------------------
    # RANDOM HOT-DECK METHOD WITHIN CLASSES AND NEAREST NEIGHBOUR METHOD
    data <- data_thirty; data$missing <- 0; data$missing[is.na(data$age)] <- 1
    data <- transform(data, group = as.numeric(factor(paste(sum_group, days_group, dgn_group,
            begin_month, type, sex, causecode_group, sep=""))))
    # Lets determine grouping value:
    data <- group_by(data, group)
    data <- mutate(data, miss = sum(missing), all = length(missing))
    data$empty <- 0
    data$empty[data$miss == data$all] <- 1
    hot_knn_thirty_sim <- hotdeck(data, variable=c("age"), domain_var = c("sum_group",
            "begin_month", "sex", "days_group", "dgn_group", "type", "causecode_group"))
    hot_knn_thirty_sim$age_imputed <- hot_knn_thirty_sim$age_imp
    hot_knn_thirty_sim$age_imp <- NULL

    # Lets impute in empty groups using nearest neighbour method:
    hot_knn_thirty_sim$age[data$empty==1] <- NA
    hot_knn_thirty_sim<- kNN(hot_knn_thirty_sim,variable=c("age"),
        dist_var = c("sum_group","begin_month", "sex", "days_group", "dgn_group", "type",
            "causecode_group"), k=1)
    hot_knn_thirty_sim$age_imp[hot_knn_thirty_sim$age_imputed==TRUE]<-TRUE
    hot_knn_thirty_sim$empty <- NULL; hot_knn_thirty_sim$all <- NULL
    hot_knn_thirty_sim$miss <- NULL; hot_knn_thirty_sim$group <- NULL
    hot_knn_thirty_sim$missing <- NULL; hot_knn_thirty_sim$age_imputed <- NULL
    hot_knn_vec_means <- append(hot_knn_vec_means, mean(hot_knn_thirty_sim$age))
}

inbetween_30 <- data.frame(hot_vec_means, knn_vec_means, hot_knn_vec_means)
means_hot <- data.frame(inbetween_30$hot_vec_means)
colnames(means_hot) <- "mean"
means_hot$method <- "hot"
means_knn <- data.frame(inbetween_30$knn_vec_means)
colnames(means_knn) <- "mean"
means_knn$method <- "knn"
means_hot_knn <- data.frame(inbetween_30$hot_knn_vec_means)
colnames(means_hot_knn) <- "mean"
means_hot_knn$method <- "hot&knn"
compare <- rbind(means_hot, means_knn, means_hot_knn)
boxplot(mean~method,data=compare)
abline(h = mean(data_try$age), col = "red")
```

Annex 9.e. Simulation of imputation and the analysis (imputation using different initial data in every step)

```
hot_vec_different_means <- c(mean(hot_thirty$age))
knn_vec_different_means <- c(mean(knn_thirty$age))
hot_knn_vec_different_means <- c(mean(hot_knn_thirty$age))

# Lets do the imputation 99 times more:
for (i in 1:99){
    # Lets delete the data and get different initial data every time:
    delete_thirty <- floor(nrow(data_try)*0.3)
    delete_rows_thirty <- sample(data_try$ID, delete_thirty)
    data_thirty <- data_try
    data_thirty$age[is.element(data_thirty$ID, delete_rows_thirty)] <- NA
    sum(!is.na(data_thirty$age))
    #-----------------------
    # GENERAL RANDOM HOT-DECK METHOD:
    hot_thirty_sim <- hotdeck(data_thirty, variable=c("age"))
    hot_vec_different_means <- append(hot_vec_different_means, mean(hot_thirty_sim$age))
    #-----------------------
    # NEAREST NEIGHBOUR METHOD:
    knn_thirty_sim <- kNN(data_thirty, variable=c("age"), dist_var = c("sum_group",
            "begin_month", "sex","days_group", "dgn_group", "type", "causecode_group"), k=1)
    knn_vec_different_means <- append(knn_vec_different_means, mean(knn_thirty_sim$age))
    #-----------------------
    # RANDOM HOT-DECK METHOD WITHIN CLASSES AND NEAREST NEIGHBOUR METHOD:
    data <- data_thirty; data$missing <- 0; data$missing[is.na(data$age)] <- 1
    data <- transform(data, group = as.numeric(factor(paste(sum_group, days_group, dgn_group,
            begin_month, type, sex, causecode_group, sep=""))))
    # Lets determine grouping value:
    data <- group_by(data, group)
    data <- mutate(data, miss = sum(missing), all = length(missing))
    data$empty <- 0; data$empty[data$miss == data$all] <- 1

    hot_knn_thirty_sim <- hotdeck(data, variable=c("age"), domain_var = c("sum_group",
            "begin_month", "sex", "days_group", "dgn_group", "type", "causecode_group"))
    hot_knn_thirty_sim$age_imputed <- hot_knn_thirty_sim$age_imp
    hot_knn_thirty_sim$age_imp <- NULL

    # Lets impute in empty groups using nearest neighbour method:
    hot_knn_thirty_sim$age[data$empty==1] <- NA
    hot_knn_thirty_sim<- kNN(hot_knn_thirty_sim, variable=c("age"), dist_var = c("sum_group",
            "begin_month", "sex", "days_group", "dgn_group", "type","causecode_group"), k=1)
    hot_knn_thirty_sim$age_imp[hot_knn_thirty_sim$age_imputed==TRUE]<-TRUE
    hot_knn_thirty_sim$empty <- NULL; hot_knn_thirty_sim$all <- NULL
    hot_knn_thirty_sim$miss <- NULL; hot_knn_thirty_sim$group <- NULL
    hot_knn_thirty_sim$missing <- NULL; hot_knn_thirty_sim$age_imputed <- NULL
    hot_knn_vec_different_means <- append(hot_knn_vec_different_means,
        mean(hot_knn_thirty_sim$age))
}

inbetween_30_different <- data.frame(hot_vec_different_means, knn_vec_different_means,
    hot_knn_vec_different_means)

means_hot <- data.frame(inbetween_30_different$hot_vec_different_means)
colnames(means_hot) <- "mean“
means_hot$method <- "hot"
means_knn <- data.frame(inbetween_30_different$knn_vec_different_means)
colnames(means_knn) <- "mean"
means_knn$method <- "knn"
means_hot_knn <- data.frame(inbetween_30_different$hot_knn_vec_different_means)
colnames(means_hot_knn) <- "mean"
means_hot_knn$method <- "hot&knn"
compare <- rbind(means_hot, means_knn,means_hot_knn)
boxplot(mean~method,data=compare); abline(h = mean(data_try$age), col = "red")
```

Annex 10. Imputation when 50% of the data exists (Software: R)

Annex 10.a. Creating missing data

```
# Lets use only complete existing data:
data_try <- subset(dat, !is.na(causecode_group))
data_try <- subset(data_try, !is.na(sex)); data_try <- subset(data_try, !is.na(age))
data_try$age_imp <- NULL; data_try$sex_imp <- NULL; data_try$causecode_group_imp <- NULL
data_try$method <- NULL

data_try$ID <- seq.int(nrow(data_try))
# Lets see how much data should be left after deleting 50% of the age values:
delete_fifty <- floor(nrow(data_try)*0.5)
delete_rows_fifty <- sample(data_try$ID, delete_fifty); data_fifty <- data_try
data_fifty$age[is.element(data_fifty$ID, delete_rows_fifty)]<-NA
sum(!is.na(data_fifty$age)) # remains 2815 values
```

Annex 10.b. New data imputation

```
# age imputation:
sum(is.na(data_fifty$age)) # missing 2814 values
#-----------------------
#-----------------------
# GENERAL RANDOM HOT-DECK METHOD:
hot_fifty <- hotdeck(data_fifty, variable=c("age"))
sum(is.na(hot_fifty$age)) # 0 missing values
#-----------------------
#-----------------------
# NEAREST NEIGHBOUR METHOD:
knn_fifty <- kNN(data_fifty, variable=c("age"), dist_var = c("sum_group", "begin_month",
        "sex","days_group", "dgn_group", "type", "causecode_group"), k=1)
sum(is.na(knn_fifty$age)) # 0 missing values
#-----------------------
#-----------------------
# RANDOM HOT-DECK METHOD WITHIN CLASSES AND NEAREST NEIGHBOUR METHOD
# Lets examine if there are groups, where there are only missing values:
data <- data_fifty; data$missing <- 0; data$missing[is.na(data$age)] <- 1
data <- transform(data, group = as.numeric(factor(paste(sum_group, days_group, dgn_group,
        begin_month, type, sex, causecode_group, sep=""))))
max(data$group) # 1041 groups formed
# Lets determine grouping value:
data <- group_by(data, group)
data <- mutate(data, miss = sum(missing), all = length(missing))
data$empty <- 0´; data$empty[data$miss == data$all] <- 1

hot_knn_fifty <- hotdeck(data, variable=c("age"), domain_var = c("sum_group", "begin_month",
    "sex", "days_group", "dgn_group", "type", "causecode_group"))
sum(is.na(hot_knn$age)) # 0 missing values
# Lets see what value was imputed in empty groups:
```

```
hot_knn_fifty$age[data$empty==1] # 1 is always assigned

hot_knn_fifty$age_imputed <- hot_knn_fifty$age_imp; hot_knn_fifty$age_imp <- NULL

# Lets impute in empty groups using nearest neighbour method:
hot_knn_fifty$age[data$empty==1] <- NA
sum(is.na(hot_knn_fifty$age))
hot_knn_fifty<- kNN(hot_knn_fifty, variable=c("age"), dist_var = c("sum_group",
        "begin_month","sex", "days_group", "dgn_group", "type", "causecode_group"), k=1)
sum(is.na(hot_knn_fifty$age)) # 0 missing values
# Lets see what value was imputed in empty groups:
hot_knn_fifty$age[data$empty==1] # different ages were assigned
hot_knn_fifty$age_imp[hot_knn_fifty$age_imputed==TRUE]<-TRUE
hot_knn_fifty$empty <- NULL; hot_knn_fifty$all <- NULL; hot_knn_fifty$miss <- NULL
hot_knn_fifty$group <- NULL; hot_knn_fifty$missing <- NULL
hot_knn_fifty$age_imputed <- NULL
```

Annex 10.c. Imputed data analysis

```
# Lets create variable to separate different methods:
hot_knn_fifty$method <- "hot-deck&knn"; hot_fifty$method <- "hot-deck"
knn_fifty$method <- "knn"; data_pr <- data_try; data_pr$method <- "actual"
data_pr$age_imp <- NA

# Lets connect the datasets for the analysis:
compare <- rbind(hot_fifty, knn_fifty, hot_knn_fifty, data_pr)
round(mean(compare$age[compare$method=="hot-deck"]), 3)
round(sd(compare$age[compare$method=="hot-deck"]), 3)
min(compare$age[compare$method=="hot-deck"])
max(compare$age[compare$method=="hot-deck"])
median(compare$age[compare$method=="hot-deck"])
round(mean(compare$age[compare$method=="knn"]), 3)
round(sd(compare$age[compare$method=="knn"]), 3)
min(compare$age[compare$method=="knn"])
max(compare$age[compare$method=="knn"])
median(compare$age[compare$method=="knn"])
round(mean(compare$age[compare$method=="hot-deck&knn"]), 3)
round(sd(compare$age[compare$method=="hot-deck&knn"]), 3)
min(compare$age[compare$method=="hot-deck&knn"])
max(compare$age[compare$method=="hot-deck&knn"])
median(compare$age[compare$method=="hot-deck&knn"])
round(mean(compare$age[compare$method=="actual"]), 3)
round(sd(compare$age[compare$method=="actual"]), 3)
min(compare$age[compare$method=="actual"])
max(compare$age[compare$method=="actual"])
median(compare$age[compare$method=="actual"])

boxplot(age~method,data=compare)

# Lets examine age differences:
data_trying <- subset(data_pr, is.element(data_fifty$ID, delete_rows_fifty))
```

```
hot_five_compare <- merge(subset(hot_fifty, age_imp==TRUE), data_trying,by="ID")
hot_five_compare$age_difference <- hot_five_compare$age.x-hot_five_compare$age.y
round(mean(hot_five_compare$age_difference), 3)
round(sd(hot_five_compare$age_difference), 3)
round(min(hot_five_compare$age_difference), 3)
round(max(hot_five_compare$age_difference), 3)
round(median(hot_five_compare$age_difference), 3)
knn_five_compare <- merge(subset(knn_fifty, age_imp==TRUE), data_trying,by="ID")
knn_five_compare$age_difference <- knn_five_compare$age.x-knn_five_compare$age.y
round(mean(knn_five_compare$age_difference), 3)
round(sd(knn_five_compare$age_difference), 3)
round(min(knn_five_compare$age_difference), 3)
round(max(knn_five_compare$age_difference), 3)
round(median(knn_five_compare$age_difference), 3)
hot_knn_five_compare <- merge(subset(hot_knn_fifty, age_imp==TRUE), data_trying,by="ID")
hot_knn_five_compare$age_difference <- hot_knn_five_compare$age.x-hot_knn_five_compare$age.y
round(mean(hot_knn_five_compare$age_difference), 3)
round(sd(hot_knn_five_compare$age_difference), 3)
round(min(hot_knn_five_compare$age_difference), 3)
round(max(hot_knn_five_compare$age_difference), 3)
round(median(hot_knn_five_compare$age_difference), 3)

compare <- rbind(hot_five_compare, knn_five_compare, hot_knn_five_compare)
boxplot(age_difference~method.x,data=compare)
```

Annex 10.d. Simulation of imputation and the analysis (imputation using same initial data in all steps)

```
hot_vec_means <- c(mean(hot_fifty$age))
knn_vec_means <- c(mean(knn_fifty$age))
hot_knn_vec_means <- c(mean(hot_knn_fifty$age))

# Lets do the imputation 99 times more:
for (i in 1:99){
    # GENERAL RANDOM HOT-DECK METHOD:
    hot_fifty_sim <- hotdeck(data_fifty, variable=c("age"))
    hot_vec_means <- append(hot_vec_means, mean(hot_fifty_sim$age))
    #-----------------------
    # NEAREST NEIGHBOUR METHOD:
    knn_fifty_sim <- kNN(data_fifty, variable=c("age"), dist_var = c("sum_group",
            "begin_month", "sex","days_group", "dgn_group", "type", "causecode_group"), k=1)
    knn_vec_means <- append(knn_vec_means, mean(knn_fifty_sim$age))
    #-----------------------
    # RANDOM HOT-DECK METHOD WITHIN CLASSES AND NEAREST NEIGHBOUR METHOD
    data <- data_fifty; data$missing <- 0; data$missing[is.na(data$age)] <- 1
    data <- transform(data, group = as.numeric(factor(paste(sum_group, days_group, dgn_group,
            begin_month, type, sex, causecode_group, sep=""))))
    # Lets determine grouping value:
    data <- group_by(data, group)
    data <- mutate(data, miss = sum(missing), all = length(missing))
    data$empty <- 0; data$empty[data$miss == data$all] <- 1

    hot_knn_fifty_sim <- hotdeck(data, variable=c("age"), domain_var = c("sum_group",
            "begin_month", "sex", "days_group", "dgn_group", "type", "causecode_group"))
    hot_knn_fifty_sim$age_imputed <- hot_knn_fifty_sim$age_imp
    hot_knn_fifty_sim$age_imp <- NULL

    # Lets impute in empty groups using nearest neighbour method:
    hot_knn_fifty_sim$age[data$empty==1] <- NA
    hot_knn_fifty_sim<- kNN(hot_knn_fifty_sim, variable=c("age"), dist_var = c("sum_group",
            "begin_month", "sex", "days_group", "dgn_group", "type", "causecode_group"), k=1)
    hot_knn_fifty_sim$age_imp[hot_knn_fifty_sim$age_imputed==TRUE]<-TRUE
    hot_knn_fifty_sim$empty <- NULL; hot_knn_fifty_sim$all <- NULL
    hot_knn_fifty_sim$miss <- NULL; hot_knn_fifty_sim$group <- NULL
    hot_knn_fifty_sim$missing <- NULL; hot_knn_fifty_sim$age_imputed <- NULL
    hot_knn_vec_means <- append(hot_knn_vec_means, mean(hot_knn_fifty_sim$age))
}

inbetween_50 <- data.frame(hot_vec_means, knn_vec_means, hot_knn_vec_means)
means_hot <- data.frame(inbetween_50$hot_vec_means)
colnames(means_hot) <- "mean"
means_hot$method <- "hot"
means_knn <- data.frame(inbetween_50$knn_vec_means)
colnames(means_knn) <- "mean"
means_knn$method <- "knn"
means_hot_knn <- data.frame(inbetween_50$hot_knn_vec_means)
colnames(means_hot_knn) <- "mean"
means_hot_knn$method <- "hot&knn"
compare <- rbind(means_hot, means_knn, means_hot_knn)
boxplot(mean~method,data=compare)
abline(h = mean(data_try$age), col = "red")
```

Annex 10.e. Simulation of imputation and the analysis (imputation using different initial data in every step)

```
hot_vec_different_means <- c(mean(hot_fifty$age))
knn_vec_different_means <- c(mean(knn_fifty$age))
hot_knn_vec_different_means <- c(mean(hot_knn_fifty$age))

# Lets do the imputation 99 times more:
for (i in 1:99){
    # Lets delete the data and get different initial data every time:
    delete_fifty <- floor(nrow(data_try)*0.5)
    delete_rows_fifty <- sample(data_try$ID, delete_fifty)
    data_fifty <- data_try
    data_fifty$age[is.element(data_fifty$ID, delete_rows_fifty)]<-NA
    #-----------------------
    # GENERAL RANDOM HOT-DECK METHOD:
    hot_fifty_sim <- hotdeck(data_fifty, variable=c("age"))
    hot_vec_different_means <- append(hot_vec_different_means, mean(hot_fifty_sim$age))
    #-----------------------
    # NEAREST NEIGHBOUR METHOD:
    knn_fifty_sim <- kNN(data_fifty, variable=c("age"), dist_var = c("sum_group",
            "begin_month", "sex","days_group", "dgn_group", "type", "causecode_group"), k=1)
    knn_vec_different_means <- append(knn_vec_different_means, mean(knn_fifty_sim$age))
    #----------------------
    # RANDOM HOT-DECK METHOD WITHIN CLASSES AND NEAREST NEIGHBOUR METHOD
    data <- data_fifty; data$missing <- 0; data$missing[is.na(data$age)] <- 1
    data <- transform(data, group = as.numeric(factor(paste(sum_group, days_group, dgn_group,
            begin_month, type, sex, causecode_group, sep=""))))
    # Lets determine grouping value:
    data <- group_by(data, group)
    data <- mutate(data, miss = sum(missing), all = length(missing))
    data$empty <- 0
    data$empty[data$miss == data$all] <- 1

    hot_knn_fifty_sim <- hotdeck(data, variable=c("age"), domain_var = c("sum_group",
            "begin_month", "sex", "days_group", "dgn_group", "type", "causecode_group"))
    hot_knn_fifty_sim$age_imputed <- hot_knn_fifty_sim$age_imp
    hot_knn_fifty_sim$age_imp <- NULL

    # Lets impute in empty groups using nearest neighbour method:
    hot_knn_fifty_sim$age[data$empty==1] <- NA
    hot_knn_fifty_sim<- kNN(hot_knn_fifty_sim, variable=c("age"), dist_var = c("sum_group",
            "begin_month","sex", "days_group", "dgn_group", "type", "causecode_group"), k=1)
    hot_knn_fifty_sim$age_imp[hot_knn_fifty_sim$age_imputed==TRUE]<-TRUE
    hot_knn_fifty_sim$empty <- NULL; hot_knn_fifty_sim$all <- NULL
    hot_knn_fifty_sim$miss <- NULL; hot_knn_fifty_sim$group <- NULL
    hot_knn_fifty_sim$missing <- NULL; hot_knn_fifty_sim$age_imputed <- NULL
    hot_knn_vec_different_means <- append(hot_knn_vec_different_means,
        mean(hot_knn_fifty_sim$age))
}

inbetween_50_different <- data.frame(hot_vec_different_means, knn_vec_different_means,
         hot_knn_vec_different_means)
means_hot <- data.frame(inbetween_50_different$hot_vec_different_means)
colnames(means_hot) <- "mean"
means_hot$method <- "hot"
means_knn <- data.frame(inbetween_50_different$knn_vec_different_means)
colnames(means_knn) <- "mean"
means_knn$method <- "knn"
means_hot_knn <- data.frame(inbetween_50_different$hot_knn_vec_different_means)
colnames(means_hot_knn) <- "mean"
means_hot_knn$method <- "hot&knn"
compare <- rbind(means_hot, means_knn, means_hot_knn)
boxplot(mean~method,data=compare)
abline(h = mean(data_try$age), col = "red")
```

Health and health care statistics:

- **Health statistics and health research database**
  http://www.tai.ee/tstua
- **Website of Health Statistics Department of National Institute for Health Development**
  http://www.tai.ee/en/r-and-d/health-statistics/activities
- **Dataquery to National Institute for Health Development**
  tai@tai.ee
- **Database of Statistics Estonia**
  http://www.stat.ee/en
- **Statistics of European Union**
  http://ec.europa.eu/eurostat
- **European health for all database (HFA-DB)**
  http://data.euro.who.int/hfadb/
- **OECD's statistical databases (OECD.Stat)**
  http://stats.oecd.org/index.aspx?DataSetCode=HEALTH_STAT