UNIVERSITY OF TARTU

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Institute of Computer Science

Computer Science Curriculum

Ants-Oskar Mäesalu

Disease Comorbidity Analysis

Bachelor’s thesis (9 ECTS)

Supervisor: Jaak Vilo, PhD

Tartu 2015

Disease Comorbidity Analysis

Abstract:

Personalised medicine is a new approach to health care, in which the focus is on the individuality of patients, and disease prediction and prevention are emphasised, as opposed to only reacting to the consequences of medical disorders. As much data about the patients and diseases as possible, as well as other medical information, is taken into account while attempting to find if and how they are linked to each other. The main objective of personalised medicine is to offer more effective treatment to every patient in a shorter period of time at a lower cost in the future.

The aim of this thesis is to study and analyse disease comorbidity in the Estonian population. 2x2 contingency tables are constructed about every pair of co-occurring ICD-10 diagnose codes in epicrises gathered by the Estonian E-Health Foundation in the years 2012-2013. The potential correlation between diseases is measured with Fisher's exact test and diagnose pairs with a stronger association are filtered. The results are visualised using heat maps. Disease comorbidity analysis is a prerequisite for future research about disease episode mining.

Keywords:

bioinformatics, personalised medicine, epidemiology, disease comorbidity, ICD-10, 2x2 contingency tables, Fisher's exact test

Haiguste komorbiidsusanalüüs

Lühikokkuvõte:

Personaalmeditsiin on uus lähenemine tervisekaitsele, milles tuuakse esile patsientide individuaalsus ja asetatakse rõhku haiguste ennetamisele nendest tekkinud tagajärgedele reageerimise asemel. Sealjuures võetakse arvesse võimalikult palju nii patsientide kui haiguste kohta teadaolevast ja ka muust meditsiinilisest teabest ning üritatakse nende vahel seoseid leida. Personaalmeditsiini üldeesmärgid on pakkuda tulevikus kõigile senisest lühema aja jooksul efektiivsemat ravi madalate kuludega.

Käesoleva töö eesmärgiks on uurida haiguste komorbiidsust Eesti populatsioonis. Töös koostatakse Eesti E-tervise Sihtasutuse 2012.-2013. aasta epikriiside andmete põhjal kõigi sama patsiendi puhul koosesinevate RHK-10 registri haiguste paaride kohta 2x2 sõltuvustabelid. Haigustevahelist võimalikku seost hinnatakse Fisheri täpse testiga, filtreeritakse välja tugevamini assotsieeritud paarid ja visualiseeritakse tulemusi nn kuumuskaartide abil. Haiguste koosesinemise uurimine on eelduseks tulevases teadustöös haigusepisoodide kaevandamisele.

Võtmesõnad:

bioinformaatika, personaalmeditsiin, epidemioloogia, haiguste komorbiidsus, RHK-10, 2x2 sõltuvustabelid, Fisheri täpne test

Acknowledgements

I would like to express my deepest gratitude to my supervisor, professor Jaak Vilo for the instruction, support and encouragement he provided me throughout the course of this thesis. His guidance helped me start from nowhere in bioinformatics, excite my interest in this field, and influenced me to carry on even when I had almost given up. I acknowledge his tremendous contribution to this thesis and the research to follow.

I am thankful to my sister who has always supported me in every way possible, and who helped me proof-read this thesis multiple times in the process of writing it. English is not my first language and as such, the assistance of a talented multilingual translator is much appreciated.

I would also like to recognise STACC and its employees for employing and tutoring me in order for me to pursue this thesis. Without their assistance, this research would have been impossible.

Table of Contents

List of Abbreviations

API	Application Programming Interface
CNS	Central Nervous System
GB	Gigabyte
ICD-10	International Classification of Diseases, revision 10
IDE	Integrated/Interactive Development Environment
MB	Megabyte
STACC	Software Technology and Applications Competence Center
SQL	Structured Query Language
SSH	Secure Shell

1. Introduction

1.1. Introduction

Personalised medicine is mainly described as a new approach to healthcare taking into account all of the individual differences of patients in all stages of the medical process, consisting of disease prevention, its diagnosis, and the treatment of health issues [1]. Its aim is to tailor everything used in healthcare to the exact needs of specific patients, thus improving the prevention and treatment of as many diseases as possible.

Even though throughout history, medicine has often included the patients' age and gender as risk factors for certain diseases, the way their medical history is considered is mostly generic and abstract, and does not involve exact statistical measurements. Intricate risk analysis taking into account all possible aspects of patients has been difficult as there has not been a lot of analysable data collected until recently and it has not had a high enough quality or good enough characterisation for studying [1]. Complex e-health systems with distinctly specified data collected in real time are in the process of solving the issue with data gathering, and data science is being employed in the analysis process as the amount of data collected is enormous.

In Estonia, there is a widespread political and social consensus on the need for a more individual approach in health care systems referring not only to genetic analysis and data usage but to the usage of all kinds of complex medical data gathered about the patients, including their genetic information, their behaviour and the effect of the environment. The development of personalised medicine is increasingly prioritised all over the world, however, in Estonia, there is already an initial nationwide e-health system in use, providing the data needed for extensive medical analysis and research [2].

One of personalised medicine's aims is to be able to predict diseases instead of reacting to their consequences [2]. Research into the medical history of patients could serve that exact purpose as previous health issues could be the causes or at least risk factors for the appearance of other disease cases. This thesis aims to contribute to that field of bioinformatics.

1.2. Motivation

Healthcare data gathering has proved the need to use big data analytics in order to sufficiently analyse medical data and to pave the road towards personalised medicine. There are tools available for use to analyse genetic data and to find the correct ways to treat diseases connected to genetic data [3]. Similarly, data mining techniques can be used to analyse not only one patient’s medical history at a time but aggregately all of the patients’ data in an attempt to find new information about connections and comorbidities between medical issues.

Even though there have been some articles on specific disease comorbidity in Estonia, for example between migraine and epilepsy and other neurological and psychiatric problems, such as depression, panic attacks and essential tremor [4], there has not yet been an extensive research of disease comorbidity conducted on the Estonian population. As e-health systems have already been applied in Estonian medicine from 2008 [5], and medical personnel are obligated to forward medical documents into the Estonian National Health Information System [6], it is about time the data is used in extensive comorbidity analysis.

Even in global bioinformatics research, comorbidity analysis on a whole population considering all of the possible disease cases is rare. Comorbidity is mostly researched taking a specific index disease into consideration [7] but approaching the problem this way evidentially takes too much time – the widely used International Classification of Diseases (ICD-10) allows more than 14,400 different diagnose codes to be specified [8]. Thus, there has been some research into the comorbidity of all of these diseases, for example in Denmark [9] [10]. This thesis mostly considers the research in those articles.

1.3. Contributions

The main aim of this thesis is to apply the comorbidity analysis methods employed on the Denmark population [9] [10] on the Estonian population to find any similarities and differences between those two datasets and to present a basis for further comorbidity research in Estonia. The thesis will provide exact measurements and numerical results for the connections between disease cases in the Estonian population and could thus serve as a way to analyse the overall medical state of the population as well as the main causes of diseases. The thesis contributes to the increasingly active field of personalised medicine in Estonia and is necessary for the development of systems capable of temporal disease prediction for specific patients.

1.4. Outline

The thesis and the current state of personalised medicine are described in the first, introductory chapter. The motivation for the thesis and the author’s contributions to the subject are expressed in detail. A short review is also given about the outline of the thesis.

Chapter 2 discusses comorbidity analysis. In it, comorbidity, multimorbidity, disease trajectories and contingency tables are defined and the scoring methods used in this thesis to rate data in contingency tables are also specified. The methods used for data visualisation and clustering are also detailed in this chapter. Furthermore, the Simpson’s paradox is explained.

In Chapter 3, the data used in this thesis is described in detail. In this chapter, the initial dataset is described and the way it was extracted, filtered and read is detailed. The errors that appeared in the process, are also discussed. In order to better describe data, some analysis on it is also described.

The implementation of the analysis described in Chapter 3 executed on the data described in Chapter 3 is specified in detail in Chapter 4. Flow diagrams and other details about the implementation of the analysis are presented. An automatic analyser built for quicker testing and studying is also detailed.

Chapter 5 outlines the results of the analysis performed in the previous chapter. The chapter provides an analysis and interpretation of the outcome of this research. The reasoning behind the results is in no way conclusive as it could be further studied and even more definitive rules could be found, however, it attempts to cover and construe all of the main points.

Conclusions and the summary of the thesis are presented in Chapter 6, the last chapter of the thesis. The main directions for further research in the field are also described in this chapter.

2. Comorbidity Analysis

Chapter 2 focuses on setting the background on comorbidity analysis. In this chapter, the terms of comorbidity, multimorbidity, disease trajectories, contingency tables are defined and the main methods for scoring and measuring the association of diseases are also detailed. The visualisation and clustering of data resulting from comorbidity analysis is also discussed in this chapter.

2.1. Comorbidity

In different kinds of clinical research, the term of comorbidity has had very different ambiguous definitions. The nature and importance of the conditions in question, chronological factors such as the time span and sequence of the conditions, and the burden of disease are usually the parameters used to differentiate between definitions of comorbidity. Other non-health-related characteristics of the patient, such as socioeconomic, cultural, environmental and behavioural data, have proved to be significant, as well. The overall term of comorbidity refers to the presence of multiple specific conditions in an individual. However, in epidemiology, a field that involves identifying the causal relationships between medical disorders, and public health research, comorbidity is usually defined as the coexistence of distinct diseases. In such analysis, an index disease is emphasised [7].

Sometimes, temporal parameters are also taken into account while analysing comorbidity, such as the sequence in which diseases or their comorbidities appear or their simultaneous co-appearance in the same exact time period to some extent [7]. It is possible that a patient suffers from a disease their whole life and it could affect the appearance of other diseases, however, it is diagnosed only once. For example, myopia, the most commonly diagnosed eye condition, has been proved to cause serious visual impairment, such as optic nerve crescent, white-without-pressure, lattice degeneration, microcystoid degeneration and pigmentary degeneration [11]. On the other hand, some disease cases only appear discretely and do not always affect the patient their whole life. Smaller accidents fall into this category.

Comorbidity analysis of diseases is used to identify the relationship between these diseases. There could, for example, be no cause between the diseases and their co-occurrence in certain patients could only be a chance [7] but with a sufficiently large set of patients, the random cases could be eliminated through cut-off values as it was done in the research in Denmark [9]. The association between two specific diseases could be described as either a direct causation, associated risk factors between the diseases, heterogeneity or independence [7].

2.2. Multimorbidity

The presence of multiple distinct disease conditions within one individual is referred to as multimorbidity. Multimorbidity analysis does not refer to a specific index condition, or usually take into account the order of the conditions in question [7], however, some research also considers the disease trajectories. Multimorbidity is a commonplace medical problem, yet there still appears to be an insufficiency of exhaustive analysis with conclusive results. [12]

As multimorbidity refers to the co-occurrence of multiple disease cases in a patient and comorbidity to the co-occurrence of only two diseases at once, multimorbidity could be described through multiple comorbid relationships between all of the possible pairs of the diseases in question. Doing so, the multimorbidity of diseases in patients can be analysed more thoroughly as there are more comprehensive measurements between the pairs of specific diagnoses. As the comorbidity between all of those pairs is taken into account concurrently during the same analysis, the multimorbidity does not affect the results too much and the description of the relationships between all of the diseases in question can be versatile and complex.

2.3. Inverse comorbidity

Sometimes, a patient who has suffered from one illness, could have a lower chance of catching some other disease than a patient who has not suffered from the first disease. This phenomenon is called inverse comorbidity. For example, certain central nervous system (CNS) disorders, such as Down’s syndrome, Alzheimer’s disease, Parkinson’s disease, multiple sclerosis and Huntington’s disease, could cause inverse cancer comorbidity in certain types of cancers [13]. This thesis does not focus on the explicit analysis of inverse comorbidity but creates the necessary utilities to perform such analysis in further work, described in Section 6.2.

2.4. Disease Trajectories

In order to analyse the comorbidity of diseases in patients, their disease trajectories are to be identified. A disease trajectory is identified as a temporal disease progression and describes the sequence diseases have followed for a patient [10]. These disease trajectories can be used to simplify the counting of comorbid disease cases.

If data used for research covers a long enough time span, these disease trajectories could be used to identify more discrete time-critical relationships between diseases and thus, to predict and prevent future diseases. Clustering significant disease trajectories could, in turn, refer to even more associations between disease groups [10]. In Denmark, researchers had 14.9 years of registry data on 6.2 million patients and could provide substantial results in disease trajectory analysis [10] but in the Estonian dataset, there was only 2 years of medical data currently available for research, described in Chapter 3, and thus, exhaustive temporal disease progression analysis was left for future work, described in Section 6.2.

2.5. Contingency Tables

A table in which a sample of some population is parted or categorised by two or more discrete qualitative variables, is widely referred to as a contingency table. Contingency tables denote frequencies at which certain parts of the population fall into these categories. They are used for frequency analysis in population groups to find associations between the groups and between the qualitative variables used [14].

A 2x2 contingency table or, in other words, a two-dimensional contingency table is the simplest form of a contingency table [14], and is defined as a table which compares two distinct groups with each other; each of them is further parted into groups either with a certain attribute or character or without it [15].

Table 1: Description of an abstract 2x2 contingency table [15]

Group	With attribute	Without attribute	Total
1	a	A – a	A
2	b	B – b	B
Total	r = a + b	N – r	N = A + B

In Table 1, group 1 consists of *A* people out of which a people have the specified attribute. Similarly, group 2 consists of *B* people, *b* of which have the same attribute. In total, there are *N = A + B* people in the groups, out of which *r = a + b* have the specified attribute.

Table 2: An alternate description of an abstract 2x2 contingency table

Group	With attribute	Without attribute	Total
1	a	b	a + b
2	c	d	c + d
Total	a + c	b + d	N = a + b + c + d

Often, as is the case in Table 2, the variables denoting the values in the fields of the contingency table are changed into *a*, *b*, *c* and *d* in order to simplify the notation of formulae applied on contingency tables. Such formulae are further described in Section 2.6.

A 2x2 contingency table is the most frequently used contingency table. It usually appears when two different independent variables are compared to each other on the basis of the same population [14]. The patient corpus used in this thesis can be divided and counted into four different fields in the two-dimensional contingency table with respect to whether or not they have suffered from some disease X and whether or not they have suffered from some disease Y. The same approach was used in research conducted in Denmark [9].

Table 3: Description of a 2x2 contingency table in the context of comorbidity between diagnoses A and B

	With Y	**Without Y**	**Total**
With X	a	b	a + b
Without X	c	d	c + d
Total	a + c	b + d	N = a + b + c + d

Here, in Table 3, the description of the table is exactly the same as for the abstract 2x2 contingency table with the exception that in place of the usual groups, people are divided into them based on whether they have been diagnosed with X or not, and the attribute studied is having been diagnosed with Y.

This kind of contingency tables can be constructed about all possible pairs of diagnoses. In this thesis, if there are *N* diagnoses accounted for, we construct an *NxN* matrix with a 2x2 contingency table on some two diseases in all of the fields of the matrix. For those cases where those two diagnoses happen to be the same diagnosis X, 0 is written into the upper left field, the amount of patients that have suffered from diagnosis X in the data set in total is written into the upper right and lower left fields, and the amount of patients who have never suffered from diagnosis X into the lower right field of the two-dimensional contingency table. Similar analysis was done in Denmark [9]. The way diagnoses are chosen for analysis is further described in Chapter 3.

2.6. Scoring Methods

There are many possible ways to evaluate the data set constructed in Section 2.5. A straightforward way would be to just compare all of the frequencies of diagnose co-associations with each other but that method would not provide much additional information. A test or scoring method needs to be chosen to evaluate the independence of the diagnoses in a pair.

Common ways to measure the contents of 2x2 contingency tables include absolute risk reduction, relative risk, relative risk reduction and odds ratio. These are, however, usually used specifically in treatment risk analysis [16]. There are better scoring methods available for testing the independence of two variables in a contingency table – for example, the chi-squared test, further described in Section 2.6.1, and Fisher's exact test, detailed in Section 2.6.2.

2.6.1. Chi-Squared Test

A chi-squared test is often used to test the independence of two variables. The test relies on investigating the validity of the null hypothesis that $p_{ij} = p_{i.} \cdot p_{.j}$, where "$p_{ij}$ represents the probability of an observation belonging to the *i*-th category of the row variable and to the *j*-th category of the column variable" – thus "$p_{i.}$ represents the probability of any observation belonging to the *i*-th category of the row variable" and "$p_{.j}$ the probability of any observation belonging to the *j*-th category of the column variable" [14].

Pearson's chi-squared test is defined as the statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}$$

where *r* denotes the number of rows and *c* the number of columns in the contingency table, and E_{ij} is the estimate of the population observation frequency in the *i*-th category of the row variable and the *j*-th category of the column variable in the contingency table, given by

$$E_{ij} = N\widehat{p_{i.}}\widehat{p_{.j}} = N\frac{p_{i.}}{N}\frac{p_{.j}}{N} = \frac{p_{i.} \cdot p_{.j}}{N}$$

where $\widehat{p_{i.}}$ and $\widehat{p_{.j}}$ are estimates of the probabilities $p_{i.}$ and $p_{.j}$, and *N* is the size of the population [14].

If the variables being compared are independent, the difference between n_{ij} and E_{ij} is minimal and the statistic χ^2 will be smaller than that of associated variables. The statistic should be compared to the corresponding value from the chi-squared distribution of the same degree of freedom as the initial contingency table, calculated as $(c - 1) \cdot (r - 1)$ [14]. Usually, 0.01 or 0.05 are used as the significance level.

For 2x2 contingency tables, this statistic can be described as [14]

$$\chi^2 = \frac{N \cdot (ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

where the variables a, b, c, d and $N = a + b + c + d$ are the same as described in Section 2.5, Table 2.

2.6.2. Fisher's Exact Test

The chi-squared test can be replaced by Fisher's exact test. While using the chi-squared test, corrections sometimes have to be applied because it uses the chi-squared approximation [17]. Fisher's exact test uses the exact probability distribution of the observed frequencies instead and corrections need not be applied. Because of its computational complexity, Fisher's exact test has usually only been applied to contingency tables where the expected frequencies are small [14].

With the availability of bigger computational power becoming more prevalent, Fisher's exact test can be applied to data sets with larger expected frequencies, too. Whenever possible, an exact test without approximations or corrections should be used [17]. This thesis mostly relies on Fisher's exact test as used in research conducted in Denmark [9].

Fisher's exact test is usually only used for 2x2 contingency tables, however, it can also be used for larger contingency tables [17]. Under the null hypothesis $p_1 = p_2$ where p_1 and p_2 represent respective population proportions having a characteristic, Fisher's exact test is identified by the hypergeometric distribution

$$f(a|r) = \frac{\binom{A}{a}\binom{B}{b}}{\binom{N}{r}} = \frac{\binom{A}{a}\binom{B}{r-a}}{\binom{N}{r}}$$

where the variables a, b, A, B and r are the same as described in Section 2.5, Table 1, and r, A and B as marginal are fixed [15]. This distribution can also be identified as

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{N}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,N!}$$

in terms of the variables a, b, c, d and N described in Section 2.5, Table 2 [14]. One-sided alternatives to the two-sided Fisher's exact test above can also be used with alternative hypotheses $p_1 > p_2$ or $p_1 < p_2$ [15].

2.6.3. Filtering Interesting Pairs

When using the *NxN* matrix of contingency tables constructed at the end of Section 2.5 where *N* denotes the amount of diagnoses in the data set, Fisher's exact test could be conducted on all of the contingency tables in the matrix. A cut-off can be imposed to filter the list of p-values returned from Fisher's exact test in order to find interesting pairs [9].

In the aforementioned Danish analysis, the measure used for choosing interesting pairs was defined as

$$s_{XY} = \log_2\left(\frac{Observed\ Value + 1}{Expected\ Value + 1}\right)$$

where the observed value is the observed number of diagnosis co-associations, and the expected value was defined as

$$Expected\ Value = \frac{n_X \cdot n_Y}{n_{total}}$$

where X and Y denote the diagnoses in question [9]. The value defined in the previous formulae can be expressed as

$$s_{XY} = \log_2\left(\frac{a+1}{\frac{(a+b)\cdot(a+c)}{n} + 1}\right)$$

using the variables *a*, *b*, *c* and *n* as described in Section 2.5, Table 3. As in Denmark, 1 was added to the nominator and the denominator of the measure in order to lower the presence of lower associated pairs in the final data set. A cut-off value of 1.0 was used in this thesis to "restrict our focus to pairs with a higher than two-fold (approximately) over co-association" [9].

In Denmark, the Benjamini-Hochberg false discovery rate method was also used on the data set to correct for multiple testing [9], however, this thesis does not rely on multiple testing and the false discovery rate method was not applied. For each pair of a patient and diagnosis, only one instance is used, as described in Chapter 3.

2.7. Heat Maps

A heat map is a method of visualising data sets in a matrix on the screen in a compact way by conveying the information in the matrix as a clustered coloured rectangular tiling with hierarchical cluster trees or, in other words, dendrograms attached to its sides. Heat maps are used because they are an effective way to depict moderately large data sets on the screen for a quick overview of data in matrices. They are, perhaps, one of the most used visualising tools used in biological sciences [18].

A heat map is practically worthless without first clustering the data set to group similar values together. When clustered, associations between different values can be taken note of and used in research. Clustering is further described in Section 2.8.

As Fisher's exact test described in Section 2.6. returns values between 0 and 1, those measurements can be used for the heat map's colour scale. 0 might be depicted as a blue rectangle in the heat map, and 1 as red. That way, after hierarchically clustering the data set and then depicting it as a heat map, red areas can be looked for on the plot and conclusions deducted from their appearance.

Heat maps can be plotted with the help of the Python library Matplotlib in the SciPy Stack [19]. The implementation is further discussed in Section 4.4.4.

2.8. Hierarchical Clustering

Clustering is a method of categorising data into similar groups, otherwise known as clusters. Its aim is to divide measurements in such a way that similar values are positioned closer to each other, in the same cluster, and dissimilar values farther away from each other, in separate clusters [20]. Clustering is used whenever there appears a need for comparative analysis on big data sets.

Hierarchical clustering is one of two most used methods of clustering – the other being partitioning –, in which each cluster is divided into subclusters, altogether forming a tree-shaped structure known as a dendrogram. One of the most usual algorithms used for hierarchical clustering, agglomerative clustering, is implemented by iteratively joining the closest items or clusters into larger clusters until every item belongs to one final supercluster consisting of all of the initial items [20].

Hierarchical clustering can be divided into several submethods based on different definitions for the distance between clusters, often referred to as the linkage function. Some of the different linkage functions include single linkage, which focuses on the shortest distance between cluster members, complete linkage, which uses the largest distance between items in clusters, average distance, which uses the arithmetic averages of distances, and centroid linkage, which uses the distance between cluster centroids [20].

Hierarchical clustering and the dendrograms resulted from performing hierarchical clustering on a data set can be used to plot comprehensible heat maps in order to better visualise the data set under investigation, as described in Section 2.7. They could also be used in other fields, for example, in temporal analysis to find similarities between different diagnoses by clustering them by periodicity. This falls out of the scope of this thesis but is briefly outlined in Section 6.2.

In this thesis, the complete linkage function is used for hierarchical clustering performed on the NxN 2x2 contingency table score matrix before visualising the data on a heat map. Hierarchical clustering can be performed in Python with the help of the SciPy Stack [19]. The implementation is further discussed in Section 4.4.4.

2.9. The Simpson's Paradox

The Simpson's paradox, also referred to as reversal paradox, describes the reasoning behind a phenomenon in contingency tables in which association trends observed in some independent population groups might reverse or disappear when those groups are combined [21]. In medicine, the Simpson's Paradox has mainly been observed in treatment analysis [22] and clinical trials [23]. The paradox has also appeared in epidemiology [24].

As comorbidity analysis relies heavily on information in 2x2 contingency tables, the Simpson's Paradox might appear in the results of the research conducted in this thesis when the population under investigation is split into multiple different groups before comorbidity analysis. Such groups may be constructed by splitting the patients either by their gender, by their specific age groups, or both. The implementation of data slicing in this thesis is further discussed in Section 4.5.2. Other parameters may also be incorporated into data splitting operations in the future, as described in Section 6.2.

3. Data

Chapter 3 describes the data used in this thesis in detail. In this chapter, the initial data sets are defined, their extraction, slicing and usage is explained, and the errors occurring in the database used are also described. Methods used to further illustrate the data used in this research are also detailed in this chapter.

3.1. Description

The data used in this thesis was gathered from the Estonian E-Health Foundation’s system. Permission to use it for research purposes was granted to the Software Technology and Applications Competence Center (STACC). The complete dataset consists of ambulatory and stationary epicrises in the years 2012 and 2013. All of the data is anonymised and can hence be used for all kinds of medical analysis. The size of the dataset is detailed in Table 4.

Table 4: Description of the complete dataset

	2012	**2013**	**Total**
Stationary epicrises	214,754	211,355	426,109
Ambulatory epicrises	1,156,732	2,090,043	3,246,775
Space requirements	49 GB	61 GB	110 GB

The dataset used during analysis in this particular thesis consists of data about over 900,000 distinct patients in the Estonian population during the years 2012 and 2013. In total, there are about 15,000 different diagnose codes in that data set. Since the original data contained missing or erroneous values in important fields, all of the data could not be taken into account. The errors in question are described in detail in Section 3.2.

The set of patients is described by an unique identification number, their birth year and gender. All other data about the patients must be gathered from medical documents in the database. However, as the documents are used in real life and doctors need to manage their time effectively, all of the data is not categorised into different types of fields. Most of the data is hidden in raw text and data mining has to be used in order to gain access to the knowledge hidden there.

The epicrises consist of raw text, structured fields filled in the epicrisis documents and data mined specifically from the text. There is all kinds of data saved, for example, medical visits, different epicrises, patients' complaints, blood pressure measurements, drugs prescribed, death cases, etc. The diagnose codes in all of the epicrisis documents are marked as ICD-10 codes.

The International Classification of Diseases (ICD) allows for diagnoses to be categorised into different groups by two or three parts of each code, altogether consisting of more than 14,400 different possible health issues. Using ICD-10 codes is mandatory for health-related documents [8]. In the ICD-10 code "H40.1", the H stands for "diseases of the eye and adnexa", H40 stands for "glaucoma" and H40.1 stands for "primary open-angle glaucoma" [25]. A reference to the ICD-10 can be found on the World Health Organization's web page [25].

In STACC, the raw data and data mined from it are kept separately. The "work" database consists of either data collected from structured fields in the e-health system or already mined data. However, the initial information is kept in the "original" and "original_2013" databases in order for it to not get mixed up. These contain epicrises from the years 2012 and 2013 correspondingly.

In this thesis, all of the patients' gender and birth years are used. From the epicrises' data, the main diagnosis codes and epicrises types are used. In addition, in order to get the correct dates of medical disorders, a date field is used from the "original" and "original_2013" databases. The reasoning behind this is discussed in Section 3.2.

Not all of the raw data has, yet, been mined, analysed or verified. Only about a third of the original data about the years 2012 and 2013 has been successfully analysed and transferred into the "work" database. Nonetheless, it is still enough for a comprehensive comorbidity analysis.

3.2. Errors

Even though some of the fields, such as the diagnosis codes, are filled in by doctors in structured fields, some of the data has still been gathered by using data mining methods on the raw text data in the medical epicrises – for example, the actual date of illness. There are bound to be errors in the dataset. Causes of errors in the dataset can vary. Mostly, they can be caused by mistyping or spelling errors in the epicrisis documents' text, or by misanalysis of borderline cases during the extraction process while data mining.

In order for the analysis in this thesis to be as exact and useful for future research as possible, filtering is needed to clean the data. Different kinds of errors must be categorised and taken out of the initial dataset, as false data could seriously affect the outcome of the research. If one field in a database row is erroneous, no field of that row should be trusted until the cause of the error is found and fixed.

Some of the patients' identification codes are non-numeric and contain falsely mined data. In this thesis, these data rows from the patients' data table are not taken into account. All of the patients' gender fields are consistent but some of the patients do not have birth years marked. In addition, some of the birth years are illogical – some of the patients seem to be either born hundreds of years before their epicrises or later than their corresponding epicrises. Thus, only the patients with existing birth years ranging from 1912 to 2013 are considered in this thesis.

In the diagnoses database table, the diagnosis codes could, sometimes, be faulty. They might not always correspond with the ICD-10 code syntax or not be filled at all. Sometimes, the date of the diagnosis is either erroneous or falls out of the specified 2012 – 2013 year range. The explicit date fields in the "work" database document the date of the epicrisis – however, the date of the actual illness is mined in the "original" and "original_2013" databases. Only rows where none of the previously mentioned errors existed and that could be linked with specific patients and initial epicrises' data were used in this research.

3.3. Extraction

In order to perform specific scientific analysis, input data should have an appropriate format. In this thesis, it is easiest if we denote information about diagnoses for all of the patients grouped by the patient identificator, as opposed to epicrises saved as separate rows in the database. Information about the patients could then be stored about each patient separately. This kind of data notation would make it easy to slice data by patients' biological background information to perform analysis on different kinds of data subsets.

In order for the analysis to be as fast as possible and for it to not need a continuously running database connection, the large data sets were first extracted from the database, and then the data that is useful for this thesis, saved into three different text files – "diagnoses.txt", "epicrises.txt" and "patients.txt".

The diagnoses file contains a list of all of the diagnoses in the complete dataset along with their frequency. It is used to identify the most often diagnosed disorders in order to plot them on comorbidity heat maps. Each row consists of an ICD-10 code and its frequency, in that order.

The epicrises file contains the medical conditions each patient has been diagnosed with. Each row consists of a patient's identification code and their diagnose history. Each element of the diagnose history is made up of the time of the diagnose, its ICD-10 code, the epicrisis type and the diagnosis type. The epicrisis types are described as “1 – main diagnosis”, “2 – by-illness”, “3 – outer cause”, or “4 – complication”. The diagnosis types can either be “S” or “A”, indicating a stationary or an ambulatory epicrisis.

The patients’ file contains their biological background information. Each row in the file contains a patient’s anonymised identification code, their gender and their birth year, in that specific order. The patients’ file can be used to split the epicrises’ dataset into smaller subgroups by the patients’ biological background information.

SQL queries are used to read the appropriate data from the database. Queries can be constructed for all of the data files separately as per the data description above but as the data in different files has to fit together, joins have to be used. Where clauses are used to filter all possible erroneous fields. The SQL queries used in this thesis are implementation specific and can be seen from the code in Appendix B.

In case data files already exist, data does not usually need to be re-read from the database and the same already existing data files could be used for analysis instead. Granted, the need to retain the same structure the data was in before saving it into text files needs to be taken into account. The implementation of data extraction, storing and reading is further described in Section 4.2.

3.4. Slicing

As can be seen from Section 3.1, Table 4, only about 12% of all of the epicrises recorded in the national e-health system were stationary epicrises. Moreover, while most of the stationary epicrises are taken into account in the database, only about half of all of the ambulatory epicrises are analysed. If the complete dataset is analysed only as a whole, explicit results might not emerge from the data gathered from stationary epicrises. As in Denmark [10], data was split by either stationary (inpatient) or ambulatory (outpatient) epicrises.

In order to find whether comorbidity relationships differ between different age groups or genders, the complete population could also be split by the patients’ gender and birth year data. The same analysis can then be carried out on all of the different data slices. The implementation of an abstract data slicer is discussed in Section 4.5.2.

3.5. Disease Frequency

In order to better illustrate the dataset under investigation, disease frequencies could be plotted to emphasize the differences in the most frequent diseases different age groups suffer from. As the complete dataset consists of about 15,000 different diagnoses, as described in Section 3.1, diseases have to be grouped so the plots would be able to portray actual information. Each ICD-10 code can be reduced into a corresponding disease class chapter, as can be seen from the ICD-10 codes register [25].

When grouped by the amount of people of a certain age who have suffered from disorders in a certain ICD-10 chapter, a stacked plot can be made to illustrate disease class frequency in certain age groups. Each disease chapter could then be depicted with a different colour. Similar graphs were plotted in research conducted in Denmark [10]. This thesis will refer to such a plot as the absolute disease frequency plot. Its method of plotting is further discussed in Section 4.3.2.

As the female population usually exceeds the male population by numbers, and as there are more people in some age groups than in others, the absolute disease frequency plot might not depict information on which substantial conclusions could be made. To eliminate the effects of population differences in different population groups, the absolute disease frequency plot values could be divided with their corresponding values in the datasets population pyramid. Such a method to retrieve a relative disease frequency plot is also discussed in Section 4.3.2.

4. Implementation

Chapter 4 describes the implementation of the analysis specified in Chapter 2 performed on the data detailed in Chapter 3. In Chapter 4, the tools and technologies used to perform the analysis in this thesis are specified, and the reasoning behind the code provided in Appendix B is explained in detail.

4.1. Technologies Used

The main programming language for the implementation phase of the research in this thesis was Python 2.7. Python is an open source multi-platform fast easy to learn powerful general purpose programming language [19] [26], which was chosen for its portability, ease of use and the ability to quickly prototype code. Python was written in the PyCharm Community Edition 4.0.5, which is considered as one of the best IDE-s for Python for its intelligent code assistance, refactoring, debugging and testing features [27].

Scientific calculations and graph plotting were implemented with the help of SciPy, an open source Python library used for scientific calculations and programming, consisting of multiple different packages, namely NumPy, the package used for numerical calculations, pandas, the package providing easy to use data structures, and Matplotlib, a high-quality plotting software [19].

For confidentiality reasons, all of the work had to be stored on the official server the data described in Chapter 3 was stored in. Access to the server was gained through an SSH connection [28] in an encrypted partition of a VMware Player [29] virtual machine running the Debian operating system [30].

For database access, SQL was used. The database was running on MySQL, the most popular open source database used for its speed, reliability and easy handling [31], and access to the database through Python was acquired through the usage of the Python Database API MySQLdb [32]. Access into the database out of the Python application was obtained through MySQL Workbench 5.2.40, a visual tool providing easier and more intuitive ways of designing, modelling, generating and managing databases used by database architects, developers and administrators [33].

4.2. Data Extraction and Storing

4.2.1. Database Operations

As previously described in Section 3.3, there are 3 main data files used in this thesis – “diagnoses.txt”, “epicrises.txt” and “patients.txt”. These files describe directly what is stored in the actual SQL database, hence the data should be extracted and saved using similar methods. For this reason, the classes DataFilesCreator, QueryConstructor and DataFileReader were created. The latter will be discussed in Section 4.2.2, the two first, however, are used during the data extraction process from the database.

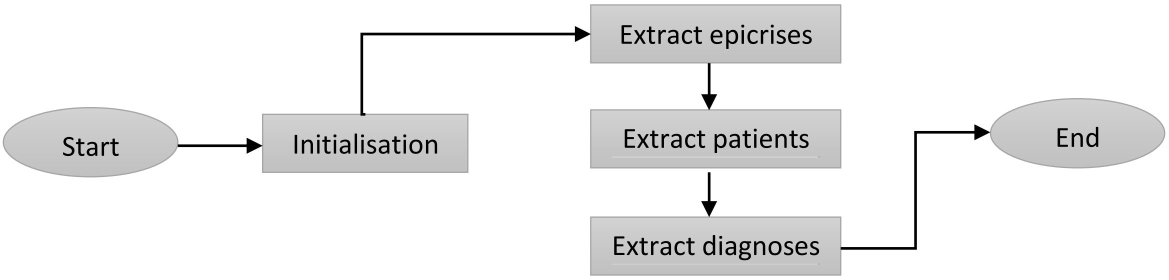


Figure 1: DataFilesCreator flow diagram

As can be seen from Figure 1, the data file creator takes care of linearly extracting all of the data corresponding to the specified three text files separately. The preceding initialisation consists of reading parameters, generating corresponding file names and ensuring the specified data file directory exists, as can be seen from the DataFilesCreator class.

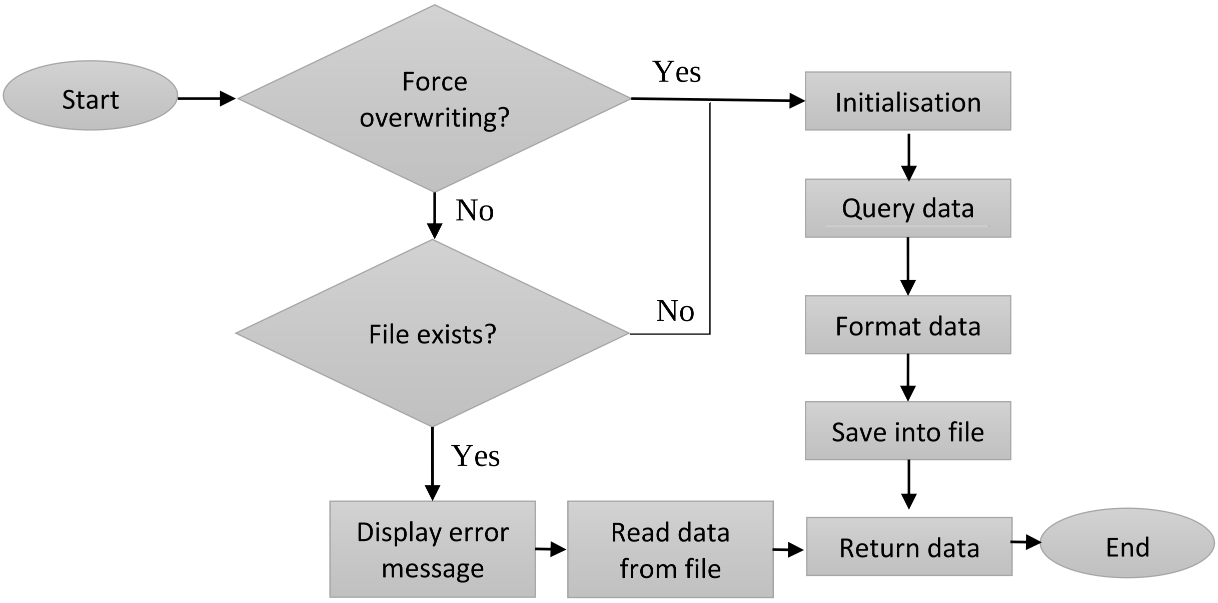


Figure 2: The data extraction process flow diagram

As depicted on Figure 2, for each of the three files, the data extraction process consists of initial checks of whether the data needs to be extracted at all – if extraction is not necessary because the user chooses for it to not be overwritten and it already exists in the specified file, the data is read and returned from the file, as described in Section 4.2.2. If it is extracted from the database, however, the process consists of self-explanatory steps – initialisation, data querying, data formatting, saving it into a text file and returning it to the user.

In order to get the data out of the database, SQL queries are needed to be constructed. However, as the queries changed a lot during research and needed to accommodate complex uses, a class capable of constructing SQL queries was made. The QueryConstructor class is a simple implementation for that purpose which allows to provide the table the data is being queried from, the fields that are to be queried, the where clauses, grouping and ordering clauses and limits. It is also possible to add inner joins to the query and to use unions. Values for the queries are provided as separate arguments and are only equipped to the query in the last phase, either query output or its execution.

As the databases used in this thesis are very specific, the exact data extraction code described in this section is probably not suitable for research in any other facilities than that of STACC. For that reason, the code is not meant for out of the box usage and no comprehensive user interface was built, as described in Section 4.6.

4.2.2. File Operations

As discussed in Section 3.3, data is saved into text files, and in order to assure consistency throughout this project, should also be retrieved into the same format as it was in before saving it into a file. The implementation of the code for the saving process can be seen in the DataFilesCreator class and for the reading process in the DataFileReader class.

Each patient can be described as a tuple of their identification code, their gender and their birth year. If all of the patients are listed into text files on separate lines, they can easily be retrieved back into the same format they were used as in the program – as a Python dictionary mapping patients' identification codes to the corresponding lists containing their gender and birth year, in that specific order.

The diagnoses file is only used for quick reference to the most popular diseases patients have been diagnosed with. Each row of the diagnoses file contains the diagnosis' ICD-10 code and its frequency in the database. In the program, the data is referred to through a Python dictionary mapping diagnosis codes to their corresponding frequencies.

The most important file used in the research conducted in this thesis is the epicrises file. The epicrises file lists all of the patients, and for each patient, the chronological list of the main diagnoses they have been attributed. For each diagnose, its date, ICD-10 code and the epicrisis and diagnosis types are saved. In order to save space and improve data reading speed, the dates are referenced as the number of days passed from the start of the query period, 1 January 2012. The diagnosis type is defined as either 1 – main diagnosis, 2 – by-illness, 3 – outer cause, or 4 – complication. The epicrisis type can be "stationary" or "ambulatory", referenced correspondingly as "S" and "A".

When the text file is read, the data is formatted as a Python dictionary mapping patients' anonymised patient identification codes to the corresponding lists containing their diagnoses' times, diagnosis codes and the epicrisis and diagnosis types. This list is essentially the same as a disease trajectory – only with additional information.

4.3. Data Analysis

4.3.1. Matrix Creation and Retrieval

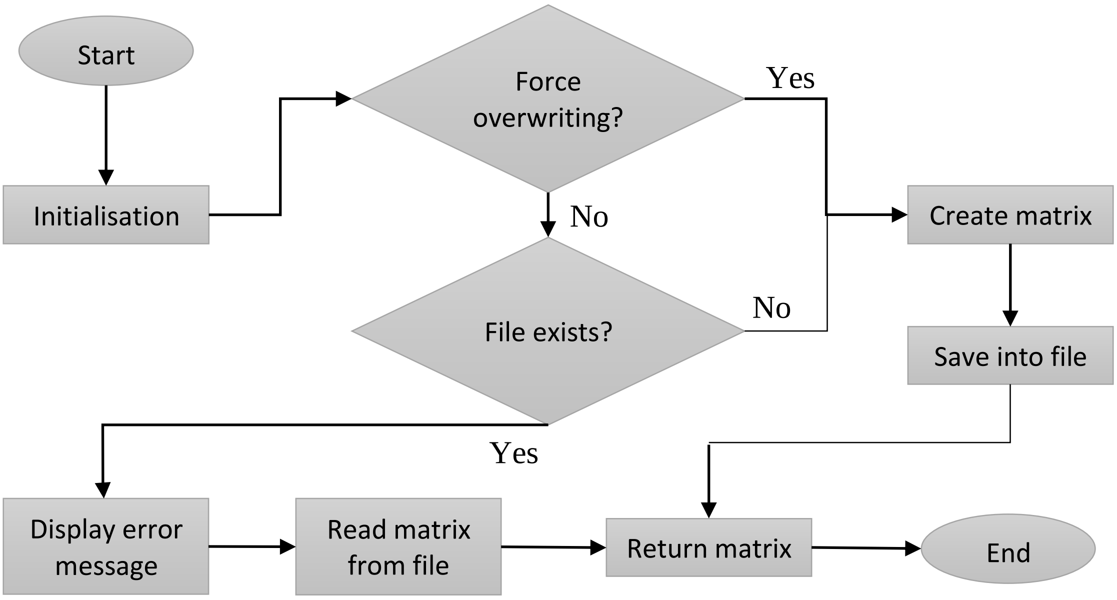


Figure 3: The abstract matrix creation and retrieval process flow diagram

The general matrix creation and reading mechanism is illustrated in Figure 3. It is used in several analyses in this thesis, for example in disease frequency analysis described in Section 4.3.2 and comorbidity analysis described in Section 4.4. The only parts of the flow diagram differing from the general process in those specific variants are the initialisation phase, and the format of the matrix used in its creation, saving and reading phases.

Similarly to the data extraction process flow diagram depicted in Figure 2, the matrix creation process first relies on initial checks whether the file the matrix is saved into for later usage needs to be created or rewritten before returning the corresponding data, or the data could just be read and returned immediately without any changes to the file containing the data in the matrix.

4.3.2. Disease Frequency

For disease frequency analysis, compound graphs consisting of 4 different population groups are plotted. The population groups are split based on the patients' gender and whether their diagnoses originate from stationary or ambulatory epicrises.

The DiseaseAgeMatrixCreator is responsible for formatting the data for plotting. Disease frequency is formatted as a Python dictionary mapping ICD-10 diagnose code chapters to Python dictionaries mapping age periods to the frequencies of diagnoses belonging to that chapter counted in people of that age period. In this thesis, 1 year is used as the age period and patient ages range from 0 to 100.

Population groups might differ by size based on the patients' gender or age group. Relative disease frequencies can be plotted when the initial absolute disease frequencies are divided by the corresponding population amount value. National population pyramid matrices are constructed using the NationalPyramidMatrixCreator which functions very similarly to the DiseaseAgeMatrix creator, creating Python dictionaries mapping patients' genders to Python dictionaries mapping age periods to the amount of patients belonging to that age group.

Disease frequencies are depicted as stacked plots using the Matplotlib Python library from the SciPy Stack [19]. In order for the plots to be comparable to similar research conducted in Denmark, exactly the same colours were chosen to denote ICD-10 chapters.

4.4. Comorbidity Analysis

4.4.1. Overall Multimorbidity

Multimorbidity is not explicitly the topic of this thesis. However, to further illustrate the need to investigate the co-occurrence of diseases, a graph can be plotted to depict the amount of co-occurring discrete diagnoses patients have suffered from during the 2-year-period used in this thesis.

The method used to depict the frequency of multimorbid patients is very similar to the one used in plotting disease frequencies. The MultimorbidityMatrixCreator returns Python dictionaries mapping different amounts of different co-occurring diagnoses to Python dictionaries mapping age periods to the actual percentages of patients in that age group that have suffered from that many medical disorders during the 2-year-period used in this thesis.

As with other plots in this thesis, the overall multimorbidity frequency plots are graphed as stacked plots using the Matplotlib Python library from the SciPy Stack [19]. As in similar analysis conducted in Scotland [34], 8 different levels of co-occurrence counts were included in this thesis, each depicted darker than the one before it in increasing order.

4.4.2. Contingency Tables

As discussed in Section 2.5, in order to investigate comorbidity, contingency tables first have to be constructed for each existing pair of co-occurring ICD-10 diagnosis code in the dataset. The ComorbidityMatrixCreator is intended for creating 2x2 contingency tables.

The ComorbidityMatrixCreator converts a Python dictionary mapping patients' anonymised identification codes to their diagnosis trajectories to a NxN Python dictionary mapping all of the pairs of diagnoses to their corresponding 2x2 lists containing the amounts of diagnoses in each of the 4 groups in a 2x2 contingency table, as discussed in Section 2.5. The co-occurrences of diseases X and Y are counted as well as their separate appearances in the matrix. All of the values in the contingency table are derived from those.

ICD-10 chapters XXI and XIX concerning "factors influencing health status and contact with health services" and "injury, poisoning and certain other consequences of external causes" were not used in the comorbidity analysis.

4.4.3. Scoring

All NxN of the 2x2 contingency tables are scored, using Fisher's exact test, as described in Section 2.6.2. The scipy.stats.fisher_exact function in the SciPy Stack [19] is used to calculate the p-value of Fisher's exact test. To save time, the cut-off is applied straight away, before even saving the value into the scoring file. If the value needs to be filtered out, the actual value is substituted with 0. The implementation of the ScoreCalculator allows the usage of other contingency scoring methods, as well. The MatrixSorter sorts all of the scores calculated in descending order so they would be conveniently readable for human eyes.

4.4.4. Heat Maps

Heat maps of the NxN matrices of scored contingency tables are plotted using the SciPy, pandas and Matplotlib libraries in the SciPy Stack [19]. The code in the heat map plotter was constructed by basing it on DeBoever's tutorial in IPython [35]. As Fisher's test returns values between 0 and 1, 0 was chosen to be depicted as blue and 1 as red on the heat maps.

In the implementation, the NxN matrix is first converted into a pandas library structure. It is then hierarchically clustered, using the scipy.cluster.hierarchy.linkage function. The heat map is then plotted in a file, labels added to two sides of the plot and dendrograms attached to the other sides of the plot. A colormap is attached into one of its corners.

4.5. Automatic Analyser

4.5.1. Necessity

Database queries usually do not take much time in practice but as there are millions of epicrises in the database, as described in Section 3.1, and as joining large tables takes even more time, database queries can be very time-consuming. Saving data into text files solves that problem because that way, it does not have to be queried multiple times – only when the initial test data needs to be changed – but slicing the data into different groups based on patients' biological background still takes a lot of time. Moreover, as there are almost a million patients taken into account in the analysis, conducting Fisher's exact test as described in Section 2.6. can take hours because of factorial calculations on contingency table matrices with more than 10,000 diagnoses in both rows and columns.

The research in this thesis consists of several different tests on several different sliced data sets. In order to save time testing, an automatic analyser was built to integrate all of the previously described data extractors, readers, writers, analysers and plotters, as the process of typing in commands one by one is aggravating and time-consuming for a human and recalculating the same initial states in order to perform different types of tests on the same test sets is redundant.

4.5.2. Design

The main points taken into consideration while constructing the automatic analyser were quick development and fast execution. As such, there are some hacks used in the code but overall, its purpose of providing quicker research into comorbidity was served. The code can be reused but needs to be refactored before applying to critical systems.

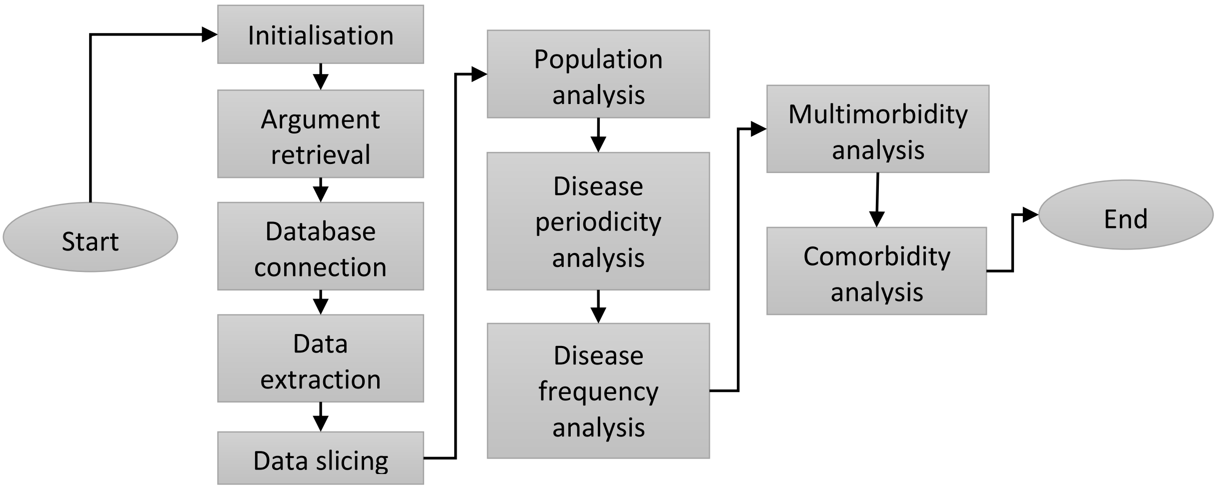


Figure 4: The automatic analyser process flow diagram

In the initialisation phase, all of the arguments and parameters used in all of the analyses in this thesis are set, using code implemented in the Arguments class. Then, the user is connected to the research database based on the username and password entered from the command line, using the code presented in the Database class. Data is then extracted from either the database or text files, as described in Section 4.2, and the dataset gathered is split into multiple parts based on the parameters set in the initialisation phase. From there on, as can be seen from Figure 4, all of the different analyses are carried out – the population analysis, disease periodicity analysis for future research purposes, disease frequency analysis, multimorbidity analysis and comorbidity analysis.

4.5.3. User Interface

The code written during the research conducted in this thesis is intended for very specific tasks. Currently, this kind of analysis can mostly only be used by data scientists. Further research, as described in Section 6.2, is needed, in order to make the results usable for actual doctors. This, however, falls out of the scope of this thesis.

As data scientist are used to code, no effort was made to develop a graphical user interface. For programmers, a command line interface is usually enough. Should the need for a graphical user interface arise, each user of this code can manipulate it to fit for their purposes.

Several functions are left for the user to modify in the code. Initial test sets and their usage is defined in the code and can not be modified from the command line interface. SQL queries, as discussed in Section 4.2.1, are very implementation-specific and should be modified, should the code be used. The code also somewhat relies on the implication that only the years 2012 and 2013 are analysed. If data from other years is included in the analysis, code should be modified.

4.6. Guidelines for Usage

In order for the implementation of the research conducted in this thesis to be installed on another system, certain software needs to be installed. As described in Section 4.1, the implementation uses Python 2.7 [26], SciPy [19] and MySQLdb [32]. The system used in this thesis ran on the Linux operating system, however, Python and the specified Python libraries are multi-platform and can be used elsewhere. Access to the server and database used in this thesis can, however, not be granted for public access, so data to be used in similar research has to be gathered by other means.

The code itself is research-specific only meant for data miners who already know how to program, and thus, no graphical user interface or simple usability instructions were constructed – in order to incorporate it into other research, the code itself has to be refactored and reformatted. If the code was abstract enough for it to suit for every type of similar research, it would be too general, would not be able to provide out of the box usage, and would need to be modified to fit for the purpose of other research anyway. The development of personalised medicine is still at the very beginning of its course and a lot of future research, some of it described in Section 6.2, has to be performed before it can be used in real time by actual healthcare personnel.

The code itself should provide the necessary information to use it, to learn how it was built, and if need be, refactor it for usage elsewhere.

5. Results

Chapter 5 describes the results of the comorbidity analysis conducted on the epicrises gathered from the Estonian E-Health Foundation's system. This chapter provides an abstract analysis and interpretation of the outcome of this research. The reasoning behind the results outlined in this chapter is in no way conclusive, however, it attempts to cover all of the main points found from the results.

5.1. Data Analysis

5.1.1. Absolute Frequency of Diseases

An absolute frequency plot retrieved from the analysis can be seen from Figure 6 in Appendix A.2.1. It would, of course, be possible to dissect it in many ways, however, the main findings seem to be that in the Estonian population:

1. most childbirths occur when the mother is 20 to 40 years old;
2. neoplasms appear to occur earlier and more frequently in women than in men;
3. diseases of the genitourinary system also appear to occur earlier and more frequently in women than in men;
4. diseases of the circulatory system and diseases of the musculoskeletal system and connective tissue appear to be the main medical issues for older people;
5. younger people seem to suffer more from diseases of the respiratory system.

The same analysis conducted in Denmark [10] can be seen from Figure 7 in Appendix A.2.2. As these results are formatted in the same way, they can be compared to each other. The results seem to match quite well, however, in Denmark, diseases of the blood and blood-forming organs and certain disorders involving the immune mechanisms seem to be a greater issue for older people than in Estonia. Conclusions 2, 3 and 4 do not seem to be as apparent in Denmark, either.

5.1.2. Relative Frequency of Diseases

As can be seen from the absolute frequency of diseases from Figure 6 in Appendix A.2.1, there seem to appear more diseases in women than in men. Also, there is a gap in the graphs for 20-year-olds and 70-year-olds. This does not mean that women are more prone to illnesses or that some specific marginal age groups are safer from diseases than others. The effect is caused by differences in the Estonian population, either between men and women, or between different age groups.

The same relationship can, indeed, be seen from the Estonian national population pyramid depicted in Figure 5 in Appendix A.1. There are less 20-year-olds because of a crisis in population growth, and there are less 70-year-olds because of World War II.

When divided by the national population pyramid of the patients taken into account in this analysis, a relative disease frequency plot is constructed. This can be seen in Figure 8 in Appendix A.2.3. The main findings seem to be that in the Estonian population:

1. as a person ages, the amount of disorders they are diagnosed with rises rapidly;
2. the relationships between different age groups in the same disease category mostly demonstrate stability;
3. old age is usually accompanied with
    a. diseases of the circulatory system,
    b. diseases of the blood and blood-forming organs and certain disorders involving the immune mechanisms (as was seen to be the case in Denmark in Section 5.1.1),
    c. diseases of the musculoskeletal system and connective tissue,
    d. neoplasms;
4. middle-aged and older people suffer the most from diseases of the musculoskeletal system and connective tissue – men earlier than women;
5. women suffer more and earlier from diseases of the genitourinary system;
6. women also suffer earlier from neoplasms, however, relatively more men are diagnosed with neoplasms in old age.
7. younger people suffer more from diseases of the respiratory system;
8. injury, poisoning and certain other consequences of external causes mostly affect young people, especially children. For women, these issues disappear when they become adults, however, men continue to get into accidents. These disorders start to reappear in old age;
9. people are also diagnosed with many different disorders before becoming an adult, probably because of regular health check-ups;
10. during younger ears, people go to check-ups more often than in old age, however, in old age, more disorders are discovered and diagnosed than in younger years;

5.2. Comorbidity Analysis

5.2.1. Overall Multimorbidity

It can be seen from Figure 9 in Appendix A.3 that in the Estonian population, most of the people suffer from multiple disorders. It is worth noting that as age increases, the amount of multimorbidity in diagnoses from stationary epicrises increases, too. However, ambulatory epicrises can consist of injuries, too, so some of the younger people have an increased chance of multimorbidity as well.

5.2.2. Comorbid Diseases

As can be seen from the results of the comorbidity analysis, most often, disorders of the same ICD-10 diagnose class are clustered together. Diseases of the respiratory system often appear together, and viral infections cluster with these diagnoses.

Another frequently occurring disease cluster is associated with overweight and old age. This does not necessarily mean that old people are overweight but it means that there is a connection between the disorders overweight people and older people usually suffer from. Overweight often causes asthma or endocrine, nutritional and metabolic diseases, which in turn cause diseases of the circulatory system. The latter are often also associated with old age and as such, other types of diseases often occurring in old age, such as disorders of the eye, appear in the same cluster as well.

The comorbidity plots are accompanied with this thesis in Appendix A.4, and the complete sorted lists of the analysis' results are provided in Appendix C.

6. Conclusions

6.1. Summary

Personalised medicine is certainly a goal to be achieved. It could provide much more effective treatment to patients, lower its costs and improve the overall healthcare of the society. It is, however, a complex area which still needs a great deal of research.

One way to make way towards personalised medicine is to analyse the comorbidity of diseases subject to different parameters in the population. This thesis focuses on the analysis of co-occurring diseases in the Estonian population in the years 2012-2013 based on epicrises gathered by the Estonian E-Health Foundation's system.

Disease comorbidity can effectively be investigated by using 2x2 contingency tables. Fisher's test turned out to be a very good measure to find associations between ICD-10 diagnose codes, and interesting pairs can be filtered using a logarithmic measure. The results are best depicted as heat maps, however, sorted association lists work, too.

Results from this thesis identify overall relationships of ICD-10 diagnose classes in the Estonian population based on the disease frequency analysis. The comorbidity analysis resulted in a large number of co-associations, as well, illustrated in the plots in Appendix A.4, and in the sorted comorbidity lists in Appendix C.

6.2. Further Work

In science, there is always more work to be done. This research contributes to the development of personalised medicine, however, to achieve substantial improvement of personalised medicine systems, extensive further analysis is needed. There are many fields left to be explored.

First of all, this thesis analysed comorbidity only based on the patients' gender and age groups. In future analysis, more biological background information could be integrated into similar comorbidity analysis to gain insight into other causal epidemiological association rules between illnesses patients have suffered from and other risk factors. For example, people who smoke, could be more prone to coronary artery disease [36].

Diseases' characteristics could also be factored in. This thesis does not separate contagious and noncontagious diseases but in theory, their comorbidity could be different, especially when the epicrises are looked at geographically. Chronic and non-chronic illnesses could have differences on how long it takes until the effect of comorbidity wears down.

Inverse comorbidity, discussed in Section 2.3, could be further analysed in the same data set to find relationships similar to the example of some CNS disorders causing a lesser probability of certain cancers [13]. Inverse comorbidity could also be analysed in relation to different patient subgroups acquired through data slicing.

Temporal disease trajectories could also be investigated, as has been done in Denmark [10]. Epidemiological knowledge on which diseases usually follow which could lead to better methods for disease prediction and prevention in the future. This kind of research can, however, only be conducted when the dataset under investigation is large enough, as discussed in Section 2.4.

Diagnoses that appear together in patients could often appear at the same time periods. Thus, by investigating when diseases have been diagnosed the most, similarities could be found between diseases after performing cluster analysis. This kind of disease periodicity analysis could be used to identify epidemics, and if applied in real time, prevent them.

Comorbidity and other types of analysis could be used to investigate possible associations between drugs and illnesses. It could be used to find which drugs are the most effective for which population groups to counter which illnesses, and could also better illustrate how the treatment process affects patients. Relationships between medications and their side-effects could be better countered if it were known which population groups are most probable to suffer from them.

It has been predicted that by 2020, about 5 million individuals in the world will have had their full genome sequenced [37]. Furthermore, while the cost for genome sequencing was once measured in millions of dollars per genome, by now, it has dropped to only thousands of dollars and will continue to drop as technology evolves [3]. The rapid decline in prices will create a golden opportunity to investigate the relationship between the human genome and illnesses.

One of the final goals of personalised medicine would be to be able to predict diseases that could occur in the future. Episode mining, while using as much data about the patients as possible, would provide insight into what could happen with specific patients as time moves on. Temporal predictions for the time period in which a disease is most probable to occur for a certain patient could also be made to make the predictions more precise.

All of the above and more could be used in research towards effective personalised medicine. The author intends to move towards these aims in the future, and to contribute to the field as much as he can.

References

[1] K. Steinhausen and S. Berghams, “Key issues affecting the development and implementation of personalised medicine: a foresight exercise,” *Drug Discovery Today: Therapeutic Strategies,* vol. 10, no. 4, pp. e189-e194, 2013.

[2] *Analüüs personaalmeditsiini rakendamise võimalustest Eestis,* Eesti Arengufond, 2013.

[3] F. F. Costa, "Big data in biomedicine," *Drug Discovery Today,* vol. 19, no. 4, pp. 433-440, 2014.

[4] V. Brin and A.-E. Kaasik, “Migreeni ja epilepsia komorbiidsus. Haigusjuht ja kommentaar,” *Eesti Arst,* vol. 82, no. 8, pp. 578-579, 2003.

[5] Estonian E-Health Foundation, “Health Information System,” [Online]. Available: http://www.e-tervis.ee/index.php/en/health-information-system. [Accessed 4 May 2015].

[6] M. Maripuu and R. Tapfer, *Tervise infosüsteemi edastatavate dokumentide andmekoosseisud ning nende säilitamise tingimused ja kord,* Riigi Teataja, 2008.

[7] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury and M. Roland, “Defining Comorbidity: Implications for Understanding Health and Health Services,” *Annals of Family Medicine,* vol. 7, no. 4, p. 357–363, 2009.

[8] P. Foudeh and N. Salim, “Information Extraction from Handwritten Medical Records and Assigning ICD-10 Codes,” in *ICCS 10 - 8th International Conference on Conceptual Structures,* Kuching, Sarawak, 2010.

[9] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, L. J. Jensen and S. Brunak, “Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts,” *PLoS Comput Biol,* vol. 7, no. 8, 2011.

[10] A. B. Jensen, P. L. Moseley, T. I. Oprea, S. G. Ellesøe, R. Eriksson, H. Schmock, P. B. Jensen, L. J. Jensen and S. Brunak, “Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients,” *Nature Communications,* vol. 5, no. 4022, 2014.

[11] P. J. Foster and Y. Jiang, “Epidemiology of myopia,” *Eye,* vol. 28, pp. 202-208, 2014.

[12] R. Vos, M. van der Akken, J. Boesten, C. Robertson and J. Metsemakers, “Trajectories of multimorbidity: exploring patterns of multimorbidity in patients with more than ten chronic health problems in life course,” *BMC Family Practice,* vol. 16, no. 2, 2015.

[13] R. Tabarés-Seisdedos and J. L. Rubenstein, “Inverse cancer comorbidity: a serendipitous opportunity to gain insight into CNS disorders,” *Nature Reviews Neuroscience,* vol. 14, pp. 293-304, 2013.

[14] B. S. Everett, The Analysis of Contingency Tables, Second Edition, Florida, USA: CRC Press LLC, 2000.

[15] B. M. Bennett and P. Hsu, “On the power function of the exact test for the 2x2 contingency table,” *Biometrika,* vol. 47, no. 3-4, pp. 363-398, 1960.

[16] E. Schechtman, “Odds Ratio, Relative Risk, Absolute Risk Reduction, and the Number Needed to Treat — Which of These Should We Use?,” *Value in Health,* vol. 5, no. 5, pp. 431-436, 2002.

[17] D. C. Howell, “Chi-Square Test: Analysis of Contingency Tables,” in *International Encyclopedia of Statistical Science,* Springer-Verlag Berlin Heidelberg, 2014, pp. 250-252.

[18] L. Wilkinson and M. Friendly, “The History of the Cluster Heat Map,” *The American Statistician,* vol. 63, no. 2, pp. 179-184, 2009.

[19] SciPy developers, “Scientific Computing Tools for Python - SciPy.org,” [Online]. Available: http://www.scipy.org/about.html. [Accessed 9 May 2015].

[20] P. D'haeseleer, “How does gene expression clustering work?,” *Nature Biotechnology,* vol. 23, pp. 1499-1501, 2005.

[21] M. A. Hernán, D. Clayton and N. Keiding, “The Simpson’s paradox unraveled,” *International Journal of Epidemiology,* vol. 40, pp. 780-785, 2011.

[22] C. J. Cates, “Simpson's paradox and calculation of number needed to treat from meta-analysis,” *BMC Medical Research Methodology,* vol. 2, no. 1, 2002.

[23] G. Rücker and M. Schumacher, “Simpson's paradox visualized: The example of the Rosiglitazone meta-analysis,” *BMC Medical Research Methodology,* vol. 8, no. 34, 2008.

[24] R. Reintjes, A. de Boer, W. van Pelt and J. Mintjes-de Groot, “Simpson’s Paradox: An Example from Hospital Epidemiology,” *Epidemiology,* vol. 11, no. 1, pp. 81-83, 2000.

[25] World Health Organization, “ICD-10 Version:2015,” [Online]. Available: http://apps.who.int/classifications/icd10/browse/2015/en#/H40-H42. [Accessed 9 May 2015].

[26] Python Software Foundation, “About Python,” [Online]. Available: https://www.python.org/about/. [Accessed 9 May 2015].

[27] JetBrains s.r.o., “Python IDE &amp Django IDE for Web developers : JetBrains PyCharm,” [Online]. Available: https://www.jetbrains.com/pycharm/. [Accessed 9 May 2015].

[28] Debian wiki team, “SSH - Debian Wiki,” SPI, 27 December 2014. [Online]. Available: https://wiki.debian.org/SSH. [Accessed 13 May 2015].

[29] “Virtual Machines & Multiple Operating Systems: VMware Player Pro | United States,” VMware, Inc, [Online]. Available: http://www.vmware.com/products/player. [Accessed 13 May 2015].

[30] “Debian -- The Universal Operating System,” SPI, 30 April 2014. [Online]. Available: https://www.debian.org/. [Accessed 13 May 2015].

[31] Oracle Corporation, “MySQL :: About MySQL,” [Online]. Available: http://www.mysql.com/about/. [Accessed 9 May 2015].

[32] A. Dustman, “MySQLdb User's Guide,” [Online]. Available: http://mysql-python.sourceforge.net/MySQLdb.html. [Accessed 9 May 2015].

[33] Oracle Corporation, “MySQL :: MySQL Workbench,” [Online]. Available: https://www.mysql.com/products/workbench/. [Accessed 9 May 2015].

[34] K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke ja B. Guthrie, „Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study,“ *The Lancet,* kd. 380, nr 9836, pp. 7-9, 2012.

[35] C. DeBoever, “Hierarchical Clustering, Heatmaps, and Gridspec,” 16 October 2013. [Online]. Available: http://nbviewer.ipython.org/github/ucsd-scientific-python/user-group/blob/master/presentations/20131016/hierarchical_clustering_heatmaps_gridspec.ipynb. [Accessed 14 May 2015].

[36] X. L. Wang, A. S. Sim, R. F. Badenhop, R. M. Mccredie ja D. E. L. Wilcken, „A smoking−dependent risk of coronary artery disease associated with a polymorphism of the endothelial nitric oxide synthase gene,“ *Nature Medicine,* kd. 2, pp. 41-46, 1996.

[37] E. J. Topol, “The big medical data miss: challenges in establishing an open medical resource,” *Nature Reviews Genetics,* vol. 16, pp. 253-254, 2015.

[38] Statistics Estonia, “Rahvastikupüramiid,” [Online]. Available: http://www.stat.ee/public/rahvastikupyramiid/. [Accessed 14 May 2015].

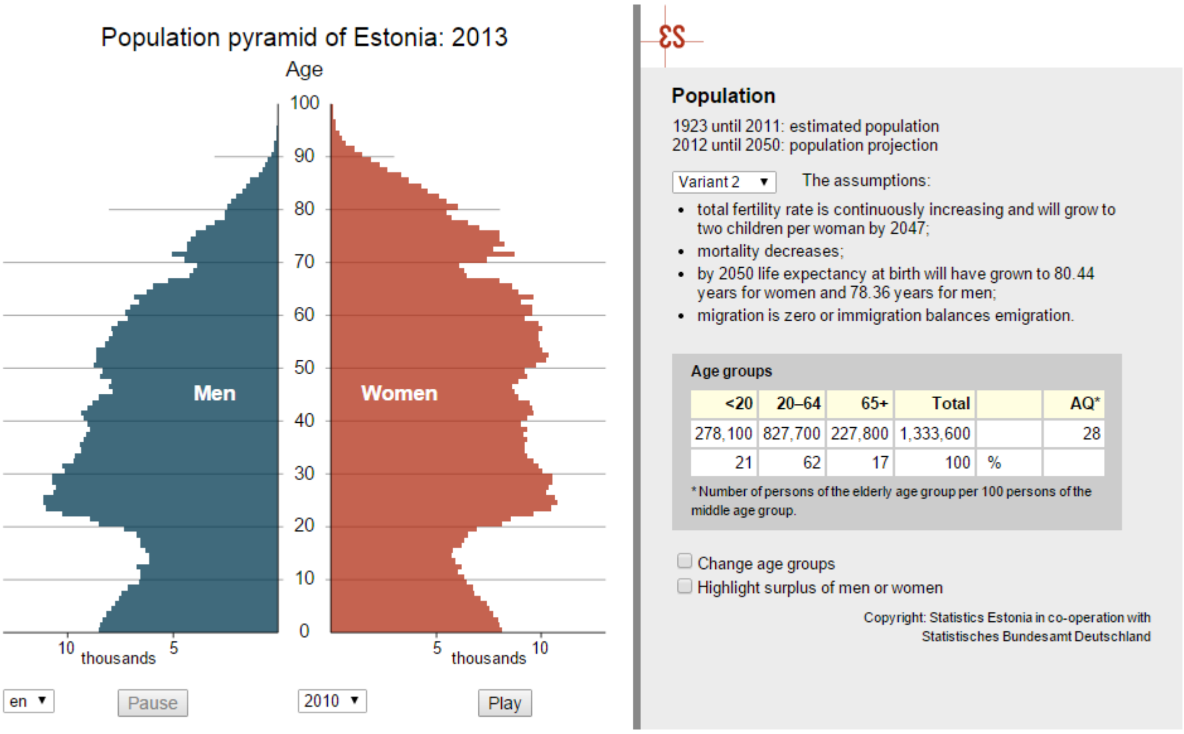
[39] S. Galea, M. Riddle and G. A. Kaplan, “Causal thinking and complex system approaches in epidemiology,” *International Journal of Epidemiology,* vol. 39, no. 1, pp. 97-106, 2009.

[40] WHO, „International Classification of Diseases (ICD),“ [Võrgumaterjal]. Available: http://www.who.int/classifications/icd/en/. [Kasutatud 14 May 2015].

Appendices

A. Graphs

A.1. Population Pyramid of Estonia 2013



Population

1923 until 2011: estimated population
2012 until 2050: population projection

Variant 2 ▾ The assumptions:

- total fertility rate is continuously increasing and will grow to two children per woman by 2047;
- mortality decreases;
- by 2050 life expectancy at birth will have grown to 80.44 years for women and 78.36 years for men;
- migration is zero or immigration balances emigration.

Age groups

<20	20–64	65+	Total		AQ*
278,100	827,700	227,800	1,333,600		28
21	62	17	100	%	

* Number of persons of the elderly age group per 100 persons of the middle age group.

□ Change age groups
□ Highlight surplus of men or women

Copyright: Statistics Estonia in co-operation with Statistisches Bundesamt Deutschland

Figure 5: The population pyramid projection of Estonia in 2013 [38]

A.2. Disease Frequency

A.2.1. Absolute Disease Frequency in Estonia

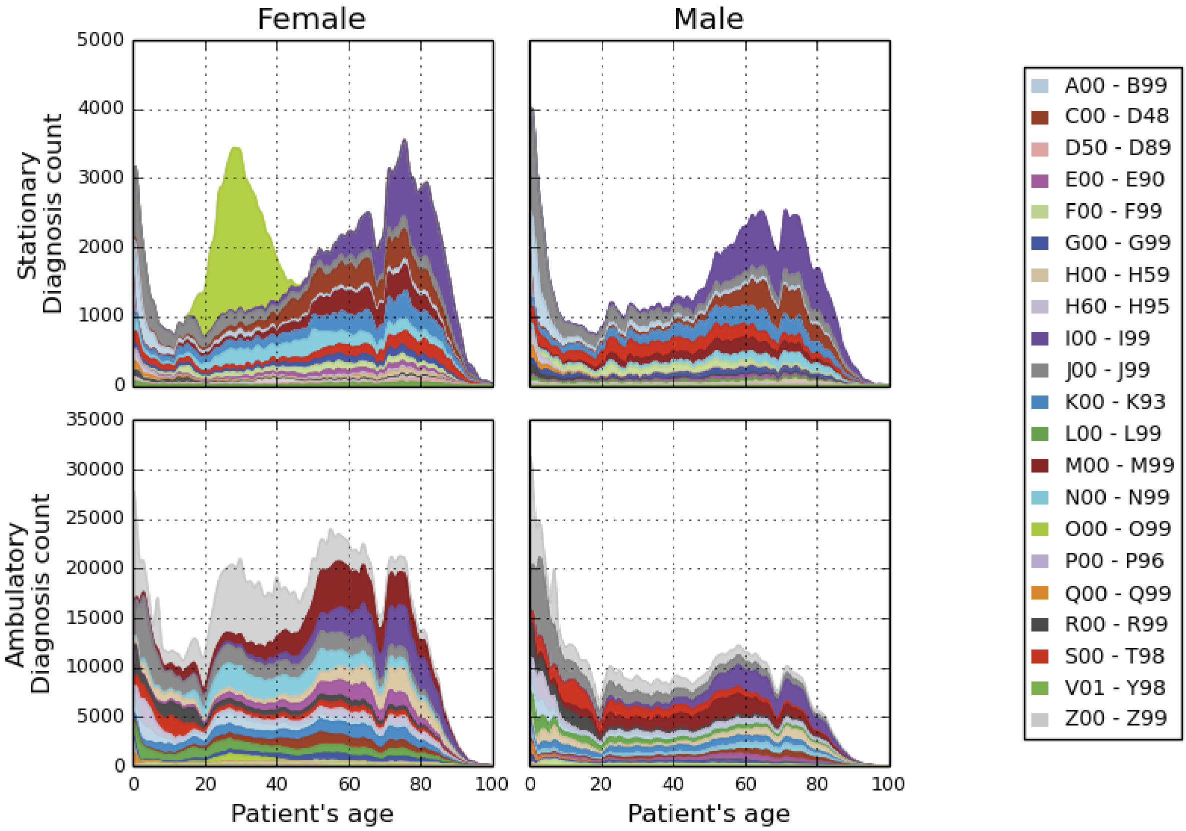


Figure 6: Absolute disease frequency in the Estonian population in 2012-2013

A.2.2. Absolute Disease Frequency in Denmark 1996-2010

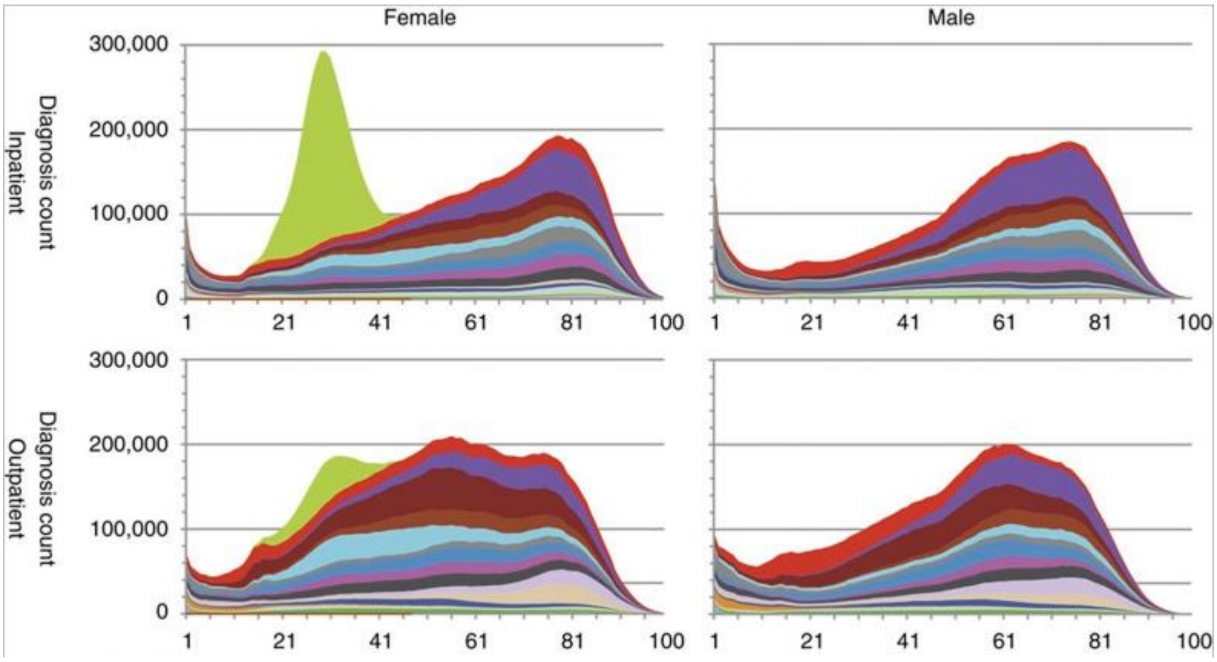


Figure 7: Absolute disease frequency in Denmark in 1996-2010 [10]

A.2.3. Relative Disease Frequency

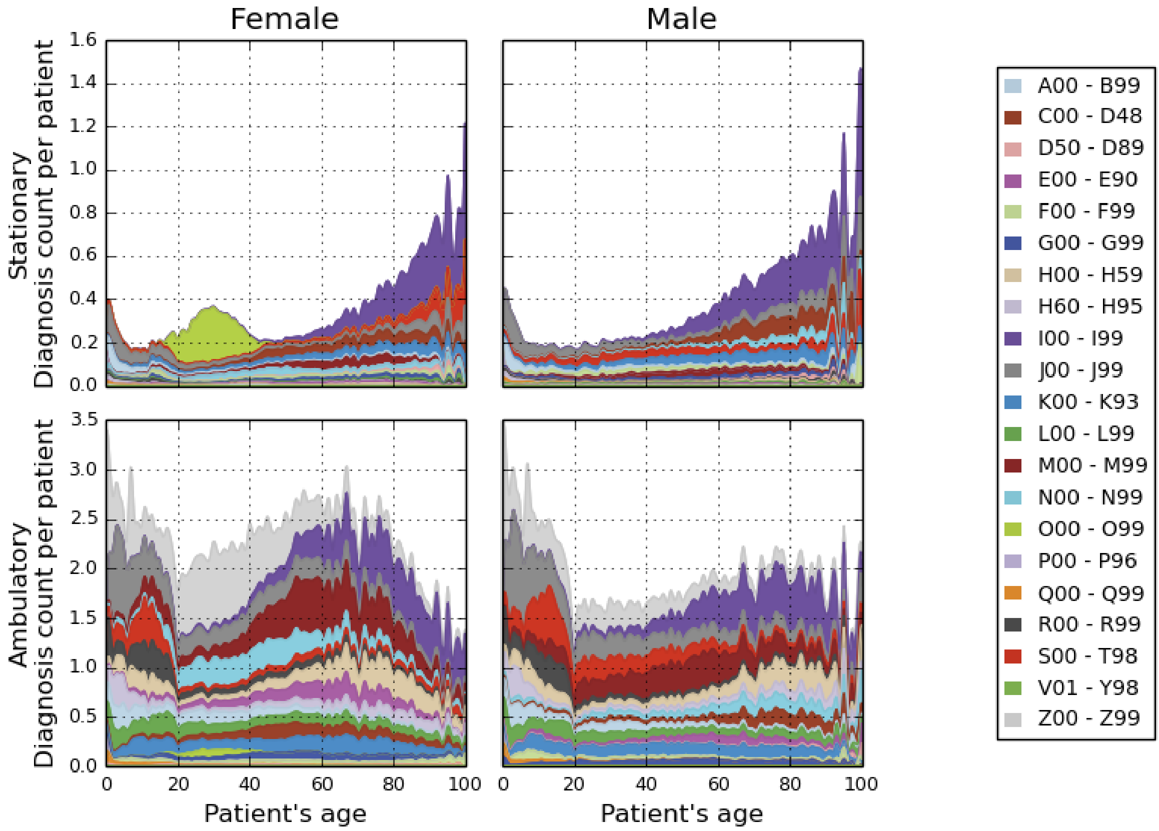


Figure 8: Relative disease frequency in the Estonian population in 2012-2013

A.3. Overall Multimorbidity

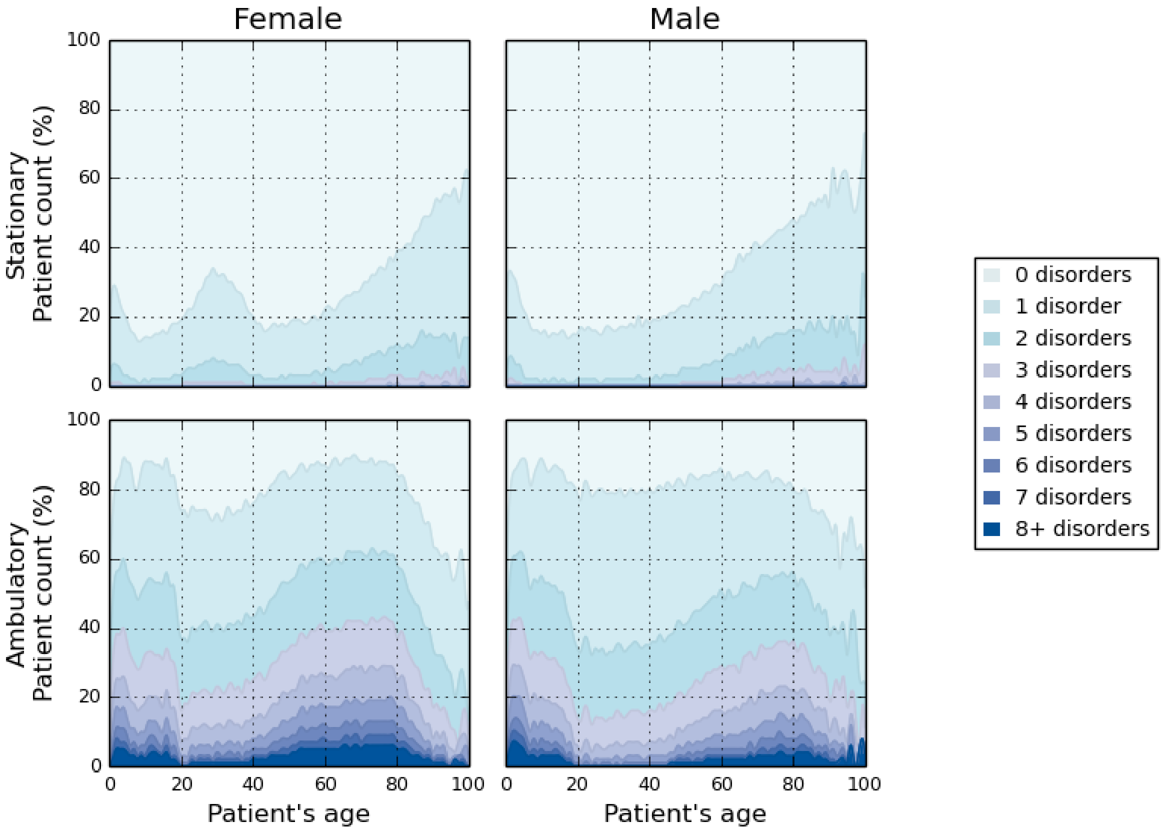


Figure 9: Overall frequency of diseases in the Estonian population in 2012-2013

A.4. Comorbidity

The comorbidity plots are accompanied separately.

B. Code

The actual code used in the implementation phase of this thesis is accompanied separately.

C. Sorted Comorbidity Lists

The lists are accompanied separately.

Non-exclusive licence to reproduce thesis and make thesis public

I, **Ants-Oskar Mäesalu** (date of birth: 13 September 1994),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:
   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and
   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   "Disease Comorbidity Analysis",

   supervised by **Jaak Vilo**,
2. I am aware of the fact that the author retains these rights.
3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 14 May 2015