UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science curriculum

Ainika Adamson

Assessment of the suitability of the Estonian Health Record data for the prediction of ischemic stroke

Master's Thesis (30 EAP)

Supervisors: Toomas Haller, PhD

Kaur Alasoo, PhD

Tartu 2021

Assessment of the suitability of the Estonian Health Record data for the prediction of ischemic stroke

Abstract: An increase in the instances of cardiovascular diseases has elevated the need for better and more efficient prediction models for ischemic stroke as well. Therefore, it is vitally important to assess the Estonian Health Record laboratory data to find out its suitability for ischemic stroke prediction models. To that effect five different approaches and three methods were utilized in three tiers in this Thesis. The potential of binary statement of measurement facts, as well as the actual analysis results, calculated z-scores and medical reference values were evaluated as the input for prediction models. It was found that the binary statement of measurements itself contained enough information for a competitive prediction model. However, several analytes were identified that had increased the quality of the prediction outcomes and therefore should be studied further.

Keywords: Data usability assessment, laboratory analyses, ischemic stroke

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics, P175 Informatics, systems theory

Terviseandmete sobivushinnang insuldi ennustusmudelites kasutamiseks

Lühikokkuvõte: Südame-veresoonkonna haiguste, sealhulgas isheemilise insuldi kasv rahvastikus on põhjustanud suurema vajaduse paremateks ennustusmeetmeteks. Seetõttu on oluline hinnata terviseandmete sobivust insuldi ennustusmudelites kasutamiseks. Käesolevas töös kasutati peamiselt uriini ja vere laborianalüüside tulemusi ennustuste konstrueerimiseks. Erinevate lähenemiste hindamiseks uuriti insuldi sõltuvust soost, vanusest ning terviseandmetest. Sealjuures hinnati binaarsete mõõtmisfaktide, meditsiiniliste referentsväärtuste, laborianalüüside absoluutväärtuste ja z-skooride mõju mudelile. Tulemuseks leiti, et binaarne mõõtmisfakt on mudelil võrdväärne sisend analüüsitulemuste absoluutväärtustele. Mainitud tulemus viitab asjaolule, et pelgalt fakt, et mõnd analüüsi teostati, on piisavalt hea alus isheemilise insuldi ennustamiseks. Siiski ilmnesid kõikide mudelite puhul ühtsed kõige paremini korreleeruvad analüüdid, nagu näiteks uriini pH ja hemogrammi üksikparameetrid, mille individuaalset potentsiaali ennustusmudelites oleks tarvis edasi uurida.

Võtmesõnad: Andmete sobivushinnang, isheemiline insult, laborianalüüsid

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika; P175 Informaatika, süsteemiteooria

Table of Contents

List of abbreviations

LIS – Laboratory Information System

HIS – Hospital Information System

CVD – Cardiovascular disease

CHD – Coronary heart disease

HL-7 – Health Level 7

RF – Random forest

BLG – Binary logistic regression

KNN - K-Nearest Neighbors

EHR – Estonian health records

CPM – Clinical prediction models

STACC – Software technology and Applications Competence Centre

ICD-10 - 10th revision of the International Statistical Classification of Diseases and Related Health Problems

1 Introduction

The number of ischemic stroke cases has grown alongside the average lifespan [1]. This drives increasing need for better stroke prevention mechanisms. The main purpose of this Thesis was to investigate, with the help of machine learning methods if and how the clinical laboratory data from the Estonian Health Records (EHR) could be useful for stroke prediction.

The ischemic stroke was selected as the object of study due to its prevalence in the population and the amount of data available for it. In order to investigate the disease, two high level research questions were posed:

1. **How to pre-process the Estonian Health Records' data to support predicting ischemic stroke?**
2. **What machine learning methods selected for comparison provide perspective insight to ischemic stroke prediction?**

The Estonian Health Records dataset had to be cleaned and structured for prediction modelling. The cleaning process varied for different approaches, since the actual laboratory results, analyte names and medical reference values were of great importance. First the data were pre-processed in order to maximize the number of clean analysis results. Secondly, the data were cleaned again to maximize the number of entries with clean reference values. The pre-processing is explained in detail in Chapter 4.

To tackle the higher-level aims, subsequent, more specific questions were presented in Chapter 3 to understand the specific nature of the data and the work at hand. More precisely, the binary statement of measurement facts' (whether a given measurement was conducted or not for a specific person) predictive value was compared to the predictive value of the laboratory results. Whether or not the medical reference values provided by laboratories or calculated z-scores (based on analysis results) offer better basis for stroke prediction models was investigated as well. Finally, the age and sex related differences were explored.

To dive into the second research question three predictive models were chosen (Chapter 3.4): Logistic Regression, K-Nearest Neighbors and Random Forest. These methods were selected to represent different machine learning approaches. All these models were

compared using the pre-processed data to help understand whether EHR data were suitable for stroke prevention and which of the selected methods performed best.

As a baseline stroke was predicted based on only sex and the year of birth. Next, the binary analyte presence variable was added. The predictions were made solely on the statement that an analyte was measured. And finally, the z-scores, absolute values and analyte groups were added as input to the models.

In conclusion, this Thesis set out to pave the road for ischemic stroke predictions based on EHR data – to find out the bottlenecks and determine best approaches and parameters for further investigations.

2 Background

This thesis will incorporate several fields in a cross-functional analysis. The purpose of this chapter is to offer a theoretical overview of the most relevant topics regarding the assessment.

2.1 Electronic Health Records

Digilugu is the Central e-Health database for laboratory measurements and a medical data sharing platform for physicians in Estonia. It is mandatory for all Health Information Systems (HIS) and Laboratory Information Systems (LIS) to send case summaries and analysis results to *Digilugu*. Regardless of the fact that the Health Level (HL) 7 standard is adopted for data exchange; some information is rather loosely structured, causing a great variety in data quality for the perspective of data mining. [2]

In addition, every laboratory information system provides their results according to their specific laboratory protocols with their own reference values and specifications. Medical reference values are affected by several variables, most of all by the specific laboratory methods used. Since universal reference values do not exist understanding the measurement results require additional knowledge of the technical context. [3]

2.2 Ischemic stroke

Cardiovascular diseases (CVD) are disorders that mainly concern blood vessels and heart. CVDs include a wide variety of common diseases such as coronary heart diseases (CHD). Currently, they are the leading cause of deaths, in 2017 around 27.75% of deaths were attributed to the heart and circulatory diseases, such as strokes. [4]

One in four people experience stroke in their lifetime [5] providing incentive to investigate the disease trajectories. There are three main types of stroke: ischemic, hemorrhagic and transient ischemic. About 87% of strokes are ischemic, [6, 7], which are often caused by blockages in the cardiovascular network. An ischemic stroke that has been detected within a few hours could be treatable with medication, thus offering incentive to find more efficient detection criteria [6].

Hemorrhagic stroke occurs due to the rupture or leakage of artery in the brain. The leaked blood causes pressure on brain damaging them in the process. High blood pressure is often the cause for hemorrhagic stroke. In case of transient ischemic stroke, the blood

flow to the brain is blocked only for a short time [6]. In acute ischemic stroke, a reduction of oxygen and glucose supply to brain is caused by a sudden decline in cerebral blood flow. This leads to extensive cell death and neurological functions' alterations. Even more, the immunological changes are not limited to brain, but causing the inflammatory response in other organs as well. [7]

The diagnosis process of acute ischemic stroke is not always straight-forward, because similar symptoms characterize multiple medical conditions resulting often in misdiagnosis and harmful medical prescriptions [5]. There are multiple risk factors for strokes, among others obesity and high blood pressure, diabetes, smoking and kidney diseases [8][9]. Due to the high prevalence it is becoming more imminent to predict who is most at risk.

To tackle aforementioned problem, the clinical prediction models (CPM), that use environment and patients' characteristics to calculate disease risk have been developed and are in use in Europe as well as USA. However, these models require human input, that at best can offer more precise scores, but at worst bias and steer the model in the wrong direction. [4] Machine learning algorithms offer possibilities to potentially outperform the developed CPMs, because such methods mostly do not require additional human input. [4]

2.3 Common analyses

There are at least 22 medical laboratories in Estonia that send electronic health data to the central *Digilugu*. [10] All those laboratories measure most common analytes (fig. 11,12) such as creatinine and complete blood count called hemogram, however they may use different methods, reagents and equipment for analysis. According to *Digilugu*, at least 7 different types of instruments are used for creatinine measurements, and different methods may result in various results [11]. Therefore, the laboratory results should be used together with incorporating the knowledge of the method and instrument used.

The analyte names mostly consist of various letter and symbol combinations provided by the laboratory information system; the most frequent symbols are explained in Table 1.

Table 1. Analyte markings and descriptions

Analyte code	Description
S_P-CK-MBm	Creatinine kinase [Mass/volume] in Serum or Plasma
S_P-Crea	Creatinine [Moles/volume] in Serum or Plasma
S_P-ALAT	Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma
S_P-ASAT	Aspartate aminotransferase [Enzymatic activity/volume] in Serum or Plasma
S_P_CRP	C reactive protein [Mass/volume] in Serum or Plasma
S_P_Urea	Urea [Moles/volume] in Serum or Plasma
S_P_K	Potassium [Moles/volume] in Serum or Plasma
S_P_Na	Sodium [Moles/volume] in Serum or Plasma
S_P-NT-proBNP	Natriuretic peptide prohormone [Mass/volume] in Serum or Plasma
B-Hct	Hematocrit [Volume Fraction] of Blood by Automated count
B-Mono_%	Monocytes/100 leukocytes in Blood by Automated count
S_P-Alb	Albumin [Mass/volume] in Serum or Plasma
S_P-cTnT-hs	Troponin T.cardiac [Mass/volume] in Serum or Plasma
B-Baso%	Basophils/100 leukocytes in Blood by Automated count
B_Eo%	Eosinophils/100 leukocytes in Blood by Automated count
B-Hb	Hemoglobin [Mass/volume] in Blood
B-Lymph%	Lymphocytes/100 leukocytes in Blood by Automated count
B-Mono#	Monocytes [#/volume] in Blood by Automated count
B-Neut#_	Neutrophils [#/volume] in Blood by Automated count
B-Neut%	Neutrophils/100 leukocytes in Blood by Automated count
B-MCH/V	Erythrocyte mean corpuscular hemoglobin [Entitic mass]/volume by Automated count
B-RDW-CV/SD	Erythrocyte distribution width [Ratio/Volume] by Automated count
S_P-TSH	Thyrotropin [Units/volume] in Serum or Plasma
S-P-PSA	Prostate specific Ag [Mass/volume] in Serum or Plasma
S-P-Chol	Cholesterol [Moles/volume] in Serum or Plasma
B-HbA1c	Hemoglobin A1c/Hemoglobin. total in Blood
B-Baso#	Basophils [#/volume] in Blood by Automated count
INR	INR in Blood by Coagulation assay
Fs_fP-Gluc	Fasting glucose [Moles/volume] in Serum or Plasma
U-Crea	Creatinine [Moles/volume] in Urine

Analyte tests are often ordered as panels and packets, provided by LIS-s to meet the specific needs of health care professionals. For example, the 12 analytes marked gray in Table 1 belong to a common panel "hemogram" and are often ordered together. The first symbol in the analyte code generally marks the material of specimen, such as serum, plasma or full blood, indicating the nature of the measurement. [10,2]

Different laboratory analyses are performed on various specimen materials, such as blood, urine, stool and multiple bodily fluids and skin crafts. The blood specimens are mostly divided into three groups: plasma, serum and whole blood. [12] Most of the specimen material used in current work is of blood and urine origin.

2.4 Machine learning algorithms

Machine learning (ML) is a method for data analysis that learns and improves through experience [13]. ML algorithms have already found some applications in medicine and the accuracy and precision of the machine learning techniques used to solve problems have improved [14]. Therefore, its applications in stroke prediction could be investigated as well.

For this Thesis, three machine learning methods were chosen that are explained in detail in the following chapters.

2.4.1 **Binary logistic regression**

Logistic regression is a statistical model used for machine learning in classification. It uses logistic function for modelling a binary dependent variable. In a binary logistic regression (BLR) there are two categories, and the model calculates probability of the potential output for the classification task. [15]

The algorithm is straightforward to implement and does not require high computational power. Training time is relatively short and the algorithm can easily be extended to multiclass classification, e.g. multinomial logistic regression. Thereby, it is a good choice for benchmarking. [16]

On the other hand, BLR is prone to over-fitting because its predictions are based on independent features. Logistic regression requires moderate or no multi-collinearity between independent variables and so the repetition of information could lead to skewed results. [17]

The algorithm is sensitive to both the outliers and noise, thus it is sometimes underperforming on medical data in comparison to more complex models. [18]

2.4.2 **K-Nearest Neighbours**

K-Nearest Neighbours (KNN) algorithm is used in machine learning for classification and regression problems. The class of a data point is predicted by the classes of its neighbouring data points and the predicted class is the most common one among the k-nearest neighbours. [19]

The neighbouring data points are determined by calculating the distance between them. There are multiple possibilities to choose from, Euclidian and Manhattan distance functions being the most common. The main parameters in this algorithm are the (k) number of neighbours and the chosen distance function. [20]

KNN is a good choice for large scale data as it is relatively fast. It "learns" from the training data at the time of making predictions. In addition, it is relatively intuitive to implement and KNN can be used on linear and non-linear data**.** [21]

However, KNN is sensitive to outliers and it cannot handle missing values. Thus, usually some kind of imputation for missing values is required. Features need to have the same scale to work most accurately. Imbalanced data cause problems that may result in skewing the outcome and large number of variables may cause the model to struggle. To overcome this, feature scaling is usually performed. [19]

2.4.3 **Random Forest**

Random forest (RF) is an ensemble learning method using a combination of decision trees. It can be used to tackle regression as well as classification tasks and it works well with continuous and categorical variables alike. [22]

The classification is based on the most popular prediction results of the decision trees in the forest. Thus, one of the most important parameters in random forest is the number of trees in it. Another important RF parameter is the split criteria, defining the decision point for dividing a node into multiple sub-nodes. There are multiple splitting criteria to choose from; Gini index and information gain being some of the most frequent. Gini index for example is equal to one minus the sum of the squared probabilities of each class. [22]

There are different parameters that can be fine-tuned to achieve better prediction performance. As an example, a maximum depth of a tree and a minimum number of samples required to form a leaf node help to avoid over-fitting. [23]

As opposed to KNN, RF uses rules for decision making, not distances. Random Forest can overcome missing data, noise and outliers, making it a good fit for medical datasets. It can also handle lots of variables and is considered a dimensionality reduction method. In addition, RF does not need feature scaling. This makes it a suitable model for datasets with a wide information variance. [24]

However, since RF creates multiple decision trees, it can require a lot of computational resources making it overly time-consuming on very large datasets. RF cannot make predictions beyond the range of training data on regression models. [22,25]

2.5 Terminology

Most important definitions and terminology used in this Thesis are described briefly. **Clinical parameters**, such as cholesterol or vitamin D values are referred as ***analytes***.

Usually more than one ***analyte*** is measured in a ***measurement***. Thereby, a measurement refers to a specific time and place where the specimen material was collected for a set of analytes. **One individual may have multiple measurements of various analytes over time**,

3 Methodology

The laboratory data originate from the Estonian electronic health records, coordinated by the Estonian Health and Welfare Information Systems Centre. However, this data repository consolidates medical information from multiple independent laboratory information systems and the universal standardized LOINC coding system for analytes has not been fully introduced for all systems. [26]

Four out of the five data files used in this work were provided by the Software Technology and Applications Competence Centre (STACC [27]). The data were anonymized having only unique identifiers for each subject. One data file was provided by Estonian Biobank [28] containing only anonymized information as well. The data analysis was performed with Python 3.9 using pandas, numpy and scikit-learn [29].

The principal task of this Thesis was to assess the suitability of clinical laboratory data for ischemic stroke prediction and potentially demonstrate its use in applicable stroke prediction models. Thus, to dissect the question **how to pre-process laboratory data to support predicting ischemic stroke**, the most vital queries are as following:

- What are the main problems and inconsistencies in the raw data?
- Does the laboratory data support predictions on individual level or measurement-based levels?
- Do medical reference values provided by LIS's offer necessary insight for predictive models?
- How much predictive value do the analysis results contain compared to the binary statement of measured analytes?
- How much predictive value do the analyte concentrations contain compared to just the year of birth and sex?

In order to resolve the second research question of **which methods selected for comparison provide perspective insight to ischemic stroke prediction** three machine learning methods were chosen.

In order to provide answers to aforementioned questions, a system of five different approaches described in Chapter 3.1, three tiers described in Chapter 3.2 and three different methods described in Chapter 3.3 was designed.

3.1 Selected approaches for the analysis

First of all, the data were pre-processed (Chapter 4) for **five different approaches (Figure 1)**. These approaches are the binary statement of measurements, the binary measurements with equal controls and cases, the medical reference values, the absolute analysis values and the calculated z-scores.

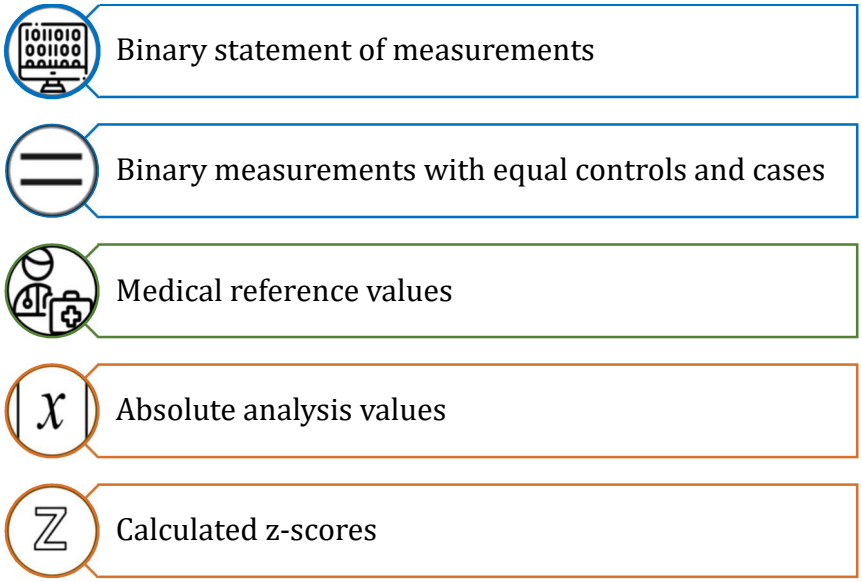

Figure 1. Approaches selected for the analysis

Used medical data contained many inconsistencies, such as missing reference values or units. For example, 23% of the laboratory results were without reference values. A few contained only one number with no indication whether it marked lower or upper bound. Moreover, some reference values had only markings "negative" or "positive" offering no meaningful information.

Therefore, it was important to determine the cleaning goals beforehand, because there were multiple options for using these data. Since the purpose of this assessment was to investigate the possibilities of clinical data applications for ischemic stroke prediction it was important to test distinct approaches in order to determine best prediction outcomes.

3.1.1 Binary data approaches

The following two approaches correspond to the first two (blue) approaches in Figure 1. Both approaches contain binary statement of measurements and no analysis result values. The cleaning process for this approach can be found in Chapter 4, in Figure 8 in color blue.

The first approach included all eligible measurements from stroke and control groups, regardless of their imbalance in numbers. Thus, there were several analytes that were measured more often for stroke patients than for control groups. The data in this approach used only the fact of measurement, coded as 1 or 0.

In the second approach the bias in control and stroke cases was adjusted by equalizing the number of each individual test within controls and strokes. To reach that goal, the number of measurements for each analyte was calculated for control groups and strokes. The lesser number was chosen, and both groups were equalized to that value. This resulted in balanced control and case groups, meaning that the control and case groups contained an equal amount of same analytes as calculated per person.

3.1.2 **Medical reference values approach**

A separate "Medical reference values" approach was developed to make use of medical reference values provided by LIS'. The purpose was to investigate their value since being provided by the laboratories, the medical reference values should contain most accurate information about the specifics of performed tests. Such approach would not over-generalize tests by assuming analysis results with same analyte names are necessarily comparable.

In this approach the analysis result values were classified as one of the following and given as inputs for the prediction models:

- Below medical reference value
- Within medical reference value
- Above medical reference value
- Negative
- Positive

For this approach a separate pre-processing logic (discussed in Chapter 4, Figure 8, in color green) was applied to the raw data in order to maximize the amount of clean reference values provided by LIS'.

3.1.3 Absolute value and z-score approaches

The following two approaches correspond to last two (brown) approaches in Figure 1. The purpose of these approaches was to assess the potential of absolute result values and z-scores as model inputs.

Here only the absolute continuous analysis values were used as input for the model. There was no adjustment for sex nor age, because the absolute values should carry the information within if some of these parameters are of importance. In this case, medical reference values were not taken into account.

For the last approach – the z-score approach, z-scores were calculated for the absolute values with respect to age and sex. The distribution of laboratory result values within each analyte were taken into account to determine whether multiple measurement units were used within one analyte results. In case of such varying measurements, the z-scores were calculated for each varying unit separately.

3.2 Selected tiers for the analysis

The aforementioned approaches were applied in three different tiers to determine if the best predictions occurred for each individual or for each measurement (Figure 2). In the first two tiers, the data were organized for each measurement, such that each row represented specific time for measurements with its values. In the third tier data were organized such that each row represented an individual. As can be seen in Figure 2, all tiers had to have at least 10 analysis results per row.

These different tiers were compared to investigate whether the prediction models work well with individual-based data even if they contain medical information from different time periods.

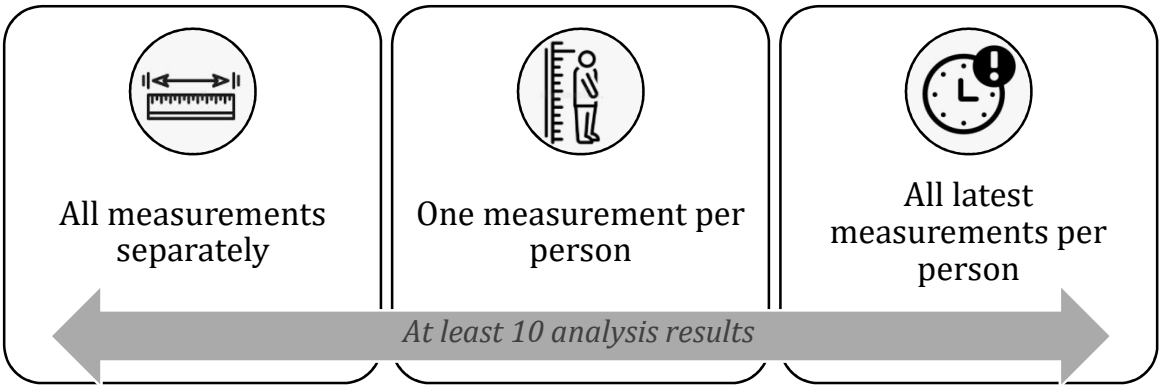

Figure 2. Selected tiers for the analysis

3.2.1 **Measurement-based data tiers**

In the first two tiers the ischemic stroke was not predicted for the individual patient, but for each measurement. Thus, the first two tiers were measurement-based, meaning the model predicted stroke not for certain individual, but for a specific measurement.

For the first tier, all measurements with at least 10 analytes were retained. For the second tier all the measurements for each person were pooled together and a random measurement of each person was chosen.

Table 2 illustrates how measurement-level data were organized for the z-score approach input. Each person can occur multiple times in the dataset with several measurements.

Table 2. Example of measurements-based data for prediction models

	Measurement Time	birth year	S-ALAT-BK	P-CLUC-BK	B-HB	S-TSH-BK	B-MHG	…	Stroke
patient1_13_01_18	13.01.2018	1943	NaN	0.3	NaN	1.2	0.12	…	1
patient1_20_02_17	20.02.2017	1943	0.8	1.34	NaN	0.6	NaN	…	1
patient2_05_07_19	05.07.2019	1939	NaN	2.1	1.7	1.1	0.4	…	0

An example of binary measurement statement data in “all measurements separately” tier can be seen in Table 3. As mentioned in Chapter 3.1.1 the “Binary measurement statement” approaches did not contain any analyte values.

Table 3. Example of binary measurement statement data for the prediction models

	Measurement Time	birth year	S-ALAT-BK	P-CLUC-BK	B-HB	S-TSH-BK	B-MHG	…	Stroke
patient1_13_01_18	13.01.2018	1943	0	1	0	1	1	…	1
patient1_20_02_17	20.02.2017	1943	1	1	0	1	0	…	1
patient2_05_07_19	05.07.2019	1939	0	1	1	1	1	…	0

3.2.2 **Individual-based data tier**

For the third tier, the predictions were individual-based, meaning the stroke was predicted for each patient basing the predictions on the latest measurement of each analyte (Table 4).

Table 4. Example of individual-level data for prediction models

	birth year	S-ALAT-BK	P-CLUC-BK	B-HB	S-TSH-BK	B-MHG	…	Stroke
patient1	1943	0.8	0.3	NaN	1.2	0.12	…	1
patient2	1939	NaN	2.1	1.7	1.1	0.4	…	0

Comparing Table 3 to Table 4, it can be observed that there is only one row per each person, all measurements have been aggregated together and only the latest ones retained. Such aggregation resulted in situations where one test result may have originated from year 2017 and another one from 2018 if there were no more subsequent measurements of those specific tests.

3.3 Selected methods for analysis

Three machine learning methods (Figure 3) were chosen to assess the applicability of EHR data for ischemic stroke predictions.

Figure 3. Selected methods for data assessment

Logistic regression was chosen to analyze whether linear models would be applicable to the data at hand. It was also selected to be a benchmark for other methods.

KNN was chosen as the second model, with Euclidian distance selected as the distance function for all approaches and number of neighbors was determined by examining different number of possible values (Chapter 3.4).

Random Forest was chosen as the third because it has been proven relatively tolerant to noise in the data. The number of trees in a forest was selected by examining different options on a separate dataset (Chapter 3.4) and other parameters were set to default to have the most comparable results

3.4 Feature engineering

The purpose of this thesis was not to fine-tune the used methods, but to generalize and make higher-level conclusions about the data at hand to pave the road for further investigations into the matter. However, the most basic parameters for selected methods still had to be chosen. The number of trees in Random Forest had to be determined. For this, the accuracy and precision were measured with different number of estimators. For KNN, the effect of different number of neighbors on accuracy and precision were measured.

For such estimations 10% of previously unused datasets were used to validate the best parameters.

3.4.1 **Parameters for Random Forest**

From Figure 4 it seems that the model accuracy was stable from 30 to 190 trees. The median precision of model (Figure 5) increased slightly from 10 to 170 trees. Thus, for the overall measurements 170 estimators were chosen for all approaches.

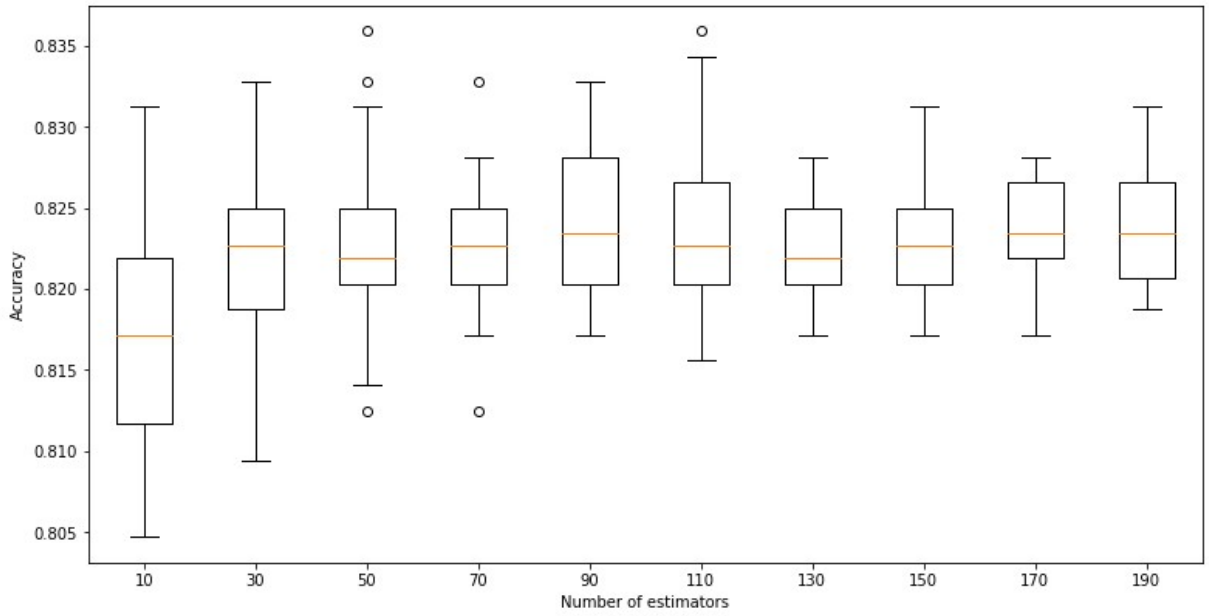

Figure 4. Random Forest accuracy depending on the number of trees

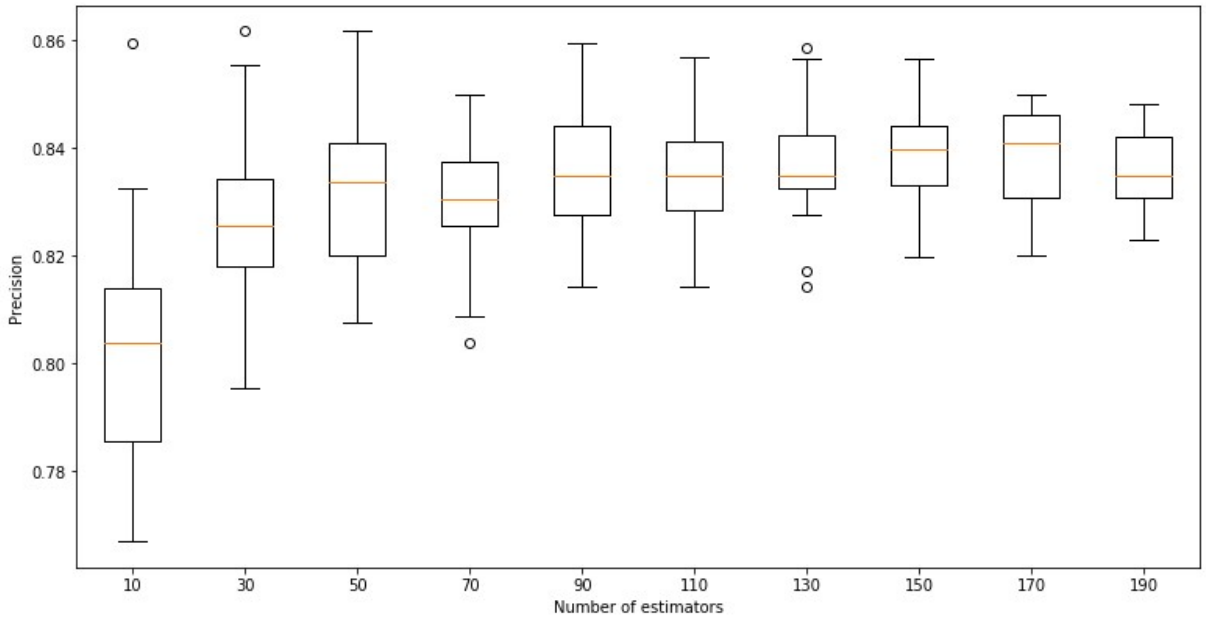

Figure 5. Random Forest precision depending on the number of trees

3.4.2 **Parameters for K-Nearest Neighbors**

For K-Nearest Neighbors the effect of different neighbors on accuracy and precision were evaluated. From Figures 6 and 7 it can be observed that both the accuracy and precision become stable at $k = 50$, thus for all approaches 50 neighbors were chosen.

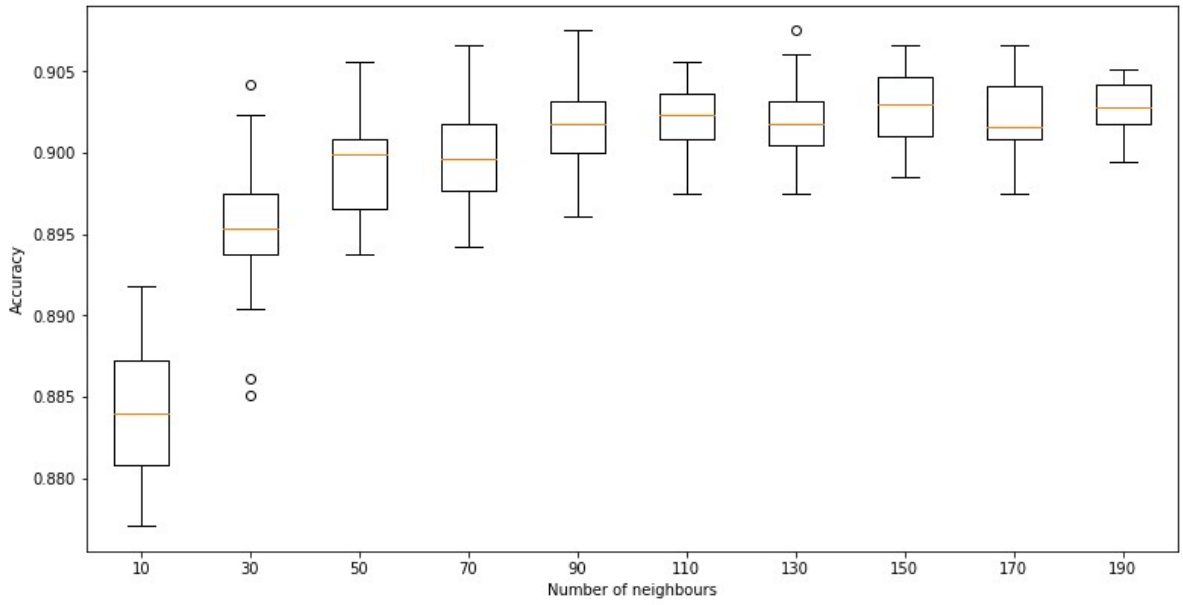

Figure 6. KNN accuracy based on the number of neighbors

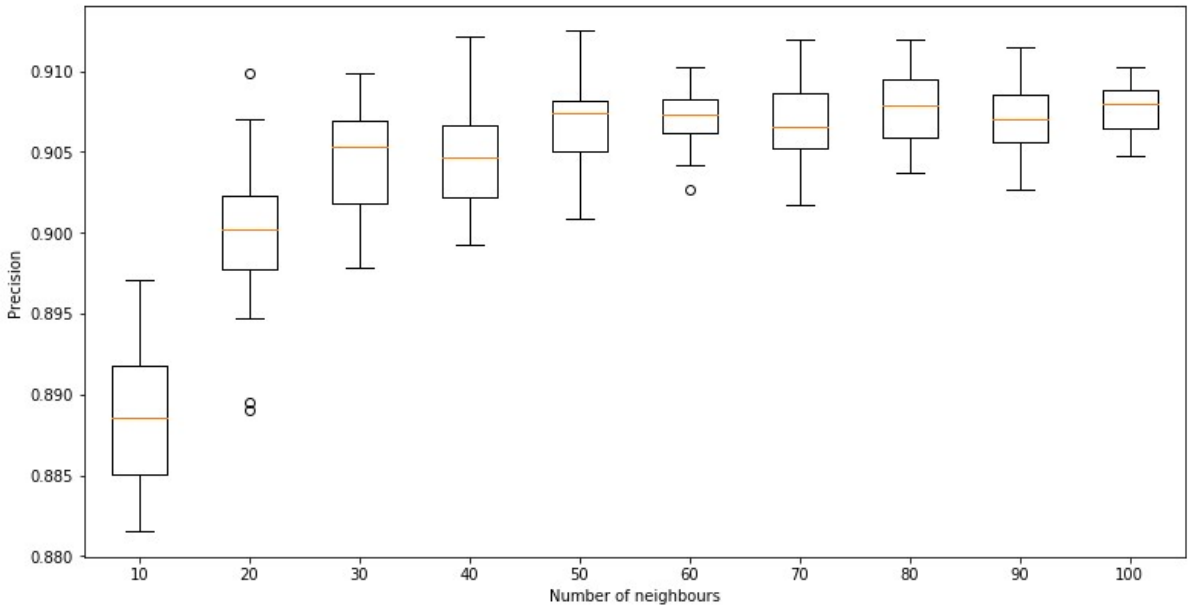

Figure 7. KNN precision based on the number of neighbors

4 Pre-processing

The pre-processing steps and an overview of the processed data were presented in this chapter. The pre-processing steps were visualized and explained and then the resulted data examined.

4.1 Pre-processing steps

To fully analyze the possibilities with the clinical dataset, **the applicability of medical reference values for predictive models needed to be assessed**. Another competing approach was to disregard medical reference values and use the **calculated z-scores for numerical analyte values or the analysis absolute values** as the model input. However, since the reference values of the raw data were not in a standardized format, different cleaning methods were utilized in order to have the best cleaned datasets for all approaches.

Figure 8 represents the main steps for pre-processing workflow. There were four datasets from STACC and one from the Estonian Biobank:

- Analyte value file
- Cohort data file
- Stroke data file
- Patient general information data file
- Estonian Biobank validation file

The analyte value file contained information such as patient id, measurement time, stroke episode, analyte class (wider class of the test), LOINC, elabor_t_lyhend (source laboratory provided abbreviation for the analysis), analysis_name, parameter_name (main part of the abbreviation), parameter unit, test value and reference value (used in laboratory). However, the LOINC codes, elabor_lyhend and analys_names did not match in several cases. Thus, upon investigating the result values, it was decided to use LOINC codes as well as elabor_t_lyhend codes for analyte names. The names were cleaned of additional symbols. In some cases, the e-labor names were missing, but LOINC codes existed, thus allowing to additionally import more than 100 000 analyte values.

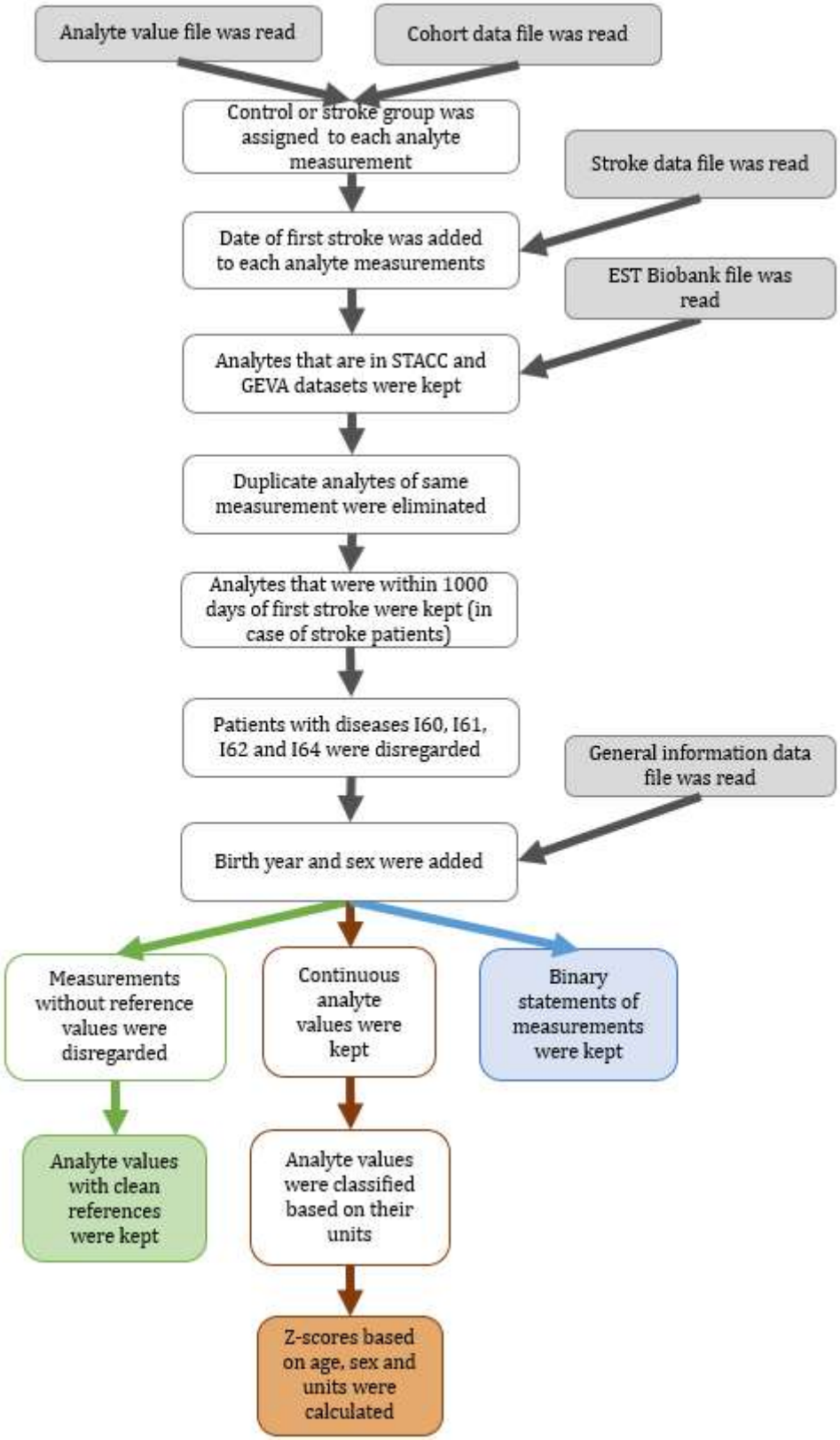

Figure 8. Pre-processing steps

The cohort data file contained information about whether each patient should belong to the stroke or control group. However, the data did not fully match the previous files 'stroke episode" information, meaning some patients were marked in control group who had marked a certain stroke episode in the first file. To assure maximum quality only those patients were retained in the analysis group who had matching information in both data files. Finally, a stroke data file was used to assign the stroke episode date for each individual in the case group.

Since the discords between the used data files did not enforce confidence, a separate data file from Estonian Biobank was used to validate if all the control cases truly were without stroke and vice a versa. The match was 100%.

Next, the duplicate analytes of same measurement were eliminated, keeping only measurements with highest absolute value. Followingly, in case of stroke patients, the measurements within the last 1000 days before stroke were kept, because it was assumed that due to the nature of the illness, the most important measurements are from shortly prior to the episode of stroke.

The ICD-10 medical classification by World Health Organization was used to distinguish ischemic stroke from other similar diseases. Thereby, in the next step, patients with diseases (ICD-10) of I60, I61, I62 and I64 were disregarded, because such illnesses may have similar clinical portraits as ischemic stroke and interfere with the analysis. Finally, birth year and sex were added from the general information data file provided by STACC.

To accommodate previously mentioned approaches, the following pre-processing was performed in three parallel workflows. The goal for medical reference value approach was to maximize the amount of **clean medical reference values** that contain only the lower and upper bound of reference values or a mark stating that the result value should be binary (negative or positive). Thus, measurements without reference values were disregarded.

Since the medical reference values originated from each reporting laboratory information systems their consistency varied extensively. For example, a common creatinine analysis had 28 different reference values for adults only. Thus, regex and cleaning scripts were applied to clean the reference values. To illustrate the need for separate cleaning

approaches, let it be noted, that the final unique analyte count for this approach was 777, mostly consisting of microbiological and molecular biology analytes.

For the absolute values and z-score approach the goal was to maximize the amount of clean analyte values that were continuous and contained no symbols such as “%”, “-“, “:”, etc. In total, there were 146 analytes with continuous result values. The z-scores were calculated considering sex, birth year and the measurement unit. In some cases, like B-HB that measures blood haemoglobin, different laboratories use various units and reference values. In case of H-HB part of the analytes were with units such as G/dl and g/l, suggesting a magnitude difference of ten times. Thus, the z scores were calculated separately for each measurement unit.

4.2 Data Overview

The data overview in this chapter is based on the dataset that contained absolute values and z-scores (brown color in Figure 1). Although, medical reference value approach contained more analyte identities, they were mostly molecular biology and microbiology results that have separate analyte name for each microorganism and PCR result, meaning they are not numerically well comparable to most frequent clinical analyses.

4.2.1 **Control vs cases measurements**

As can be seen on Figure 9, in case of control group most measurements were performed for people with birth year of 1936 and 1938. The figure shape seems logical since older individuals are expected to have more tests performed on them.

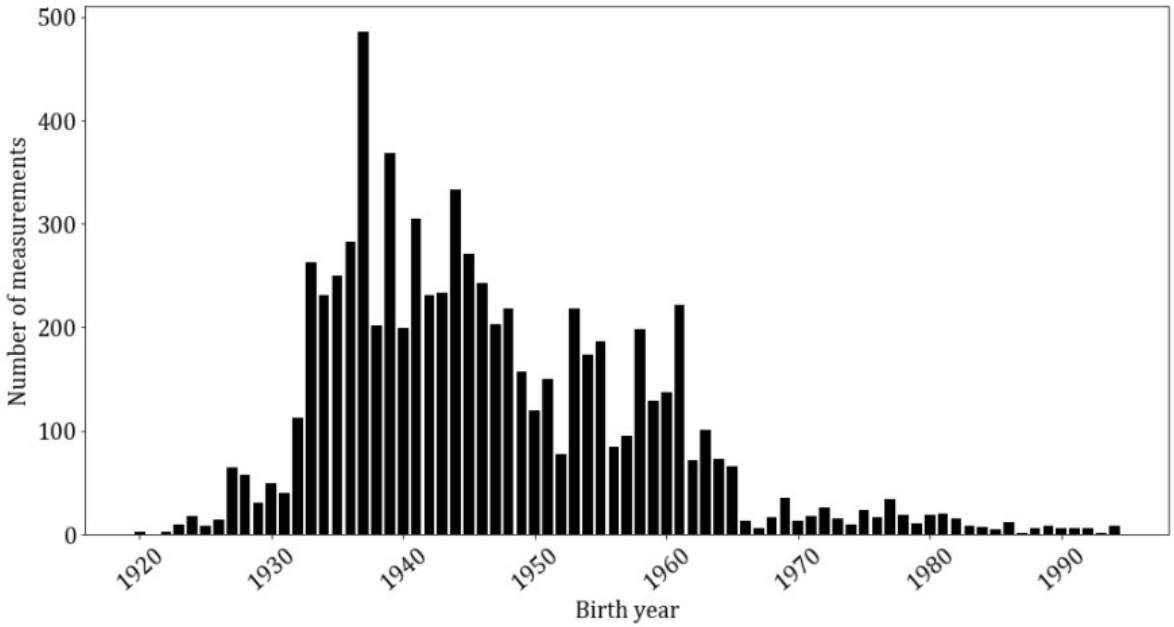

Figure 9. Number of measurements per birth year for the control group

It can be observed from Figure 10 that for the stroke cases most analytes were measured in the same age group as for the controls. For controls, there is a clear decline of measurements from 1944 to 1953, but the same cannot be stated for stroke cases. Rapid decline of measurements starts from birth year of 1950. This may be attributed to the fact that stroke itself occurs mostly in elderly people, and there are not many stroke cases in younger population.

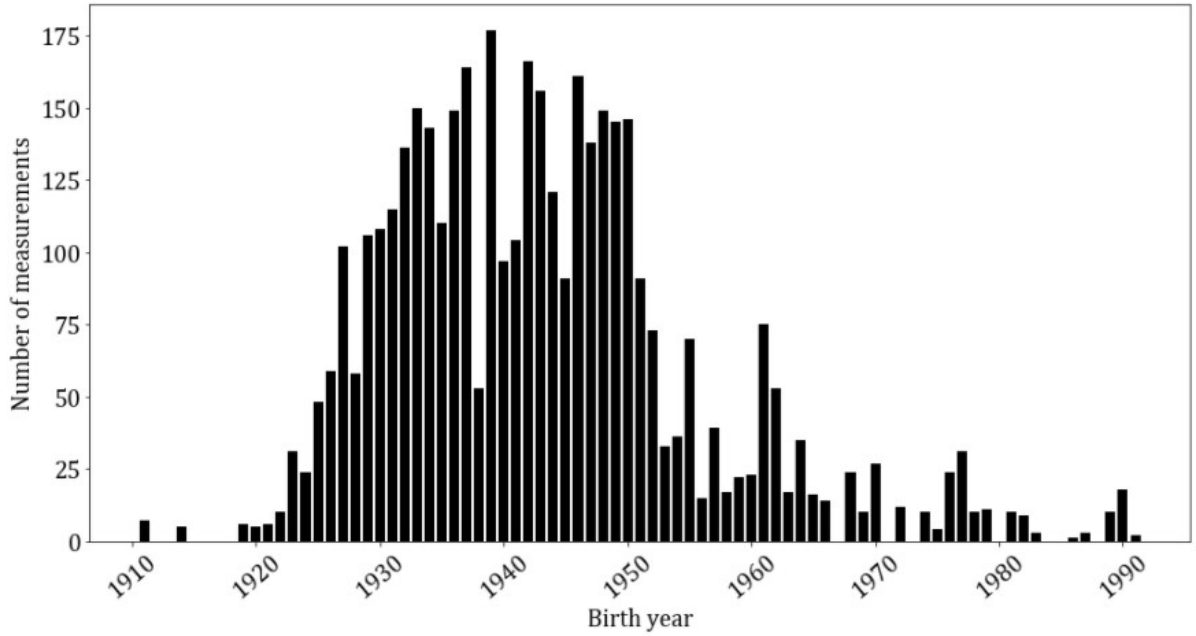

Figure 10. Number of measurements per birth year for the stroke group

In total, there were 4 094 measurements and 79 505 analytes for stroke cases and 7 347 measurements and 144 176 analytes for controls.

Figure 11 depicts the relative difference in number of measurements between stroke cases and controls. The proportion of each analyte in stroke and control groups were calculated and their respective ratios subtracted. The biggest differences were visualized. Thus, it can be seen that eGFR occurrence ratio in control group is 1.5% higher than in stroke group. B-Hct is more prevalent in stroke group as well, whereas B-MPV proportion in stroke groups is higher than in control group.

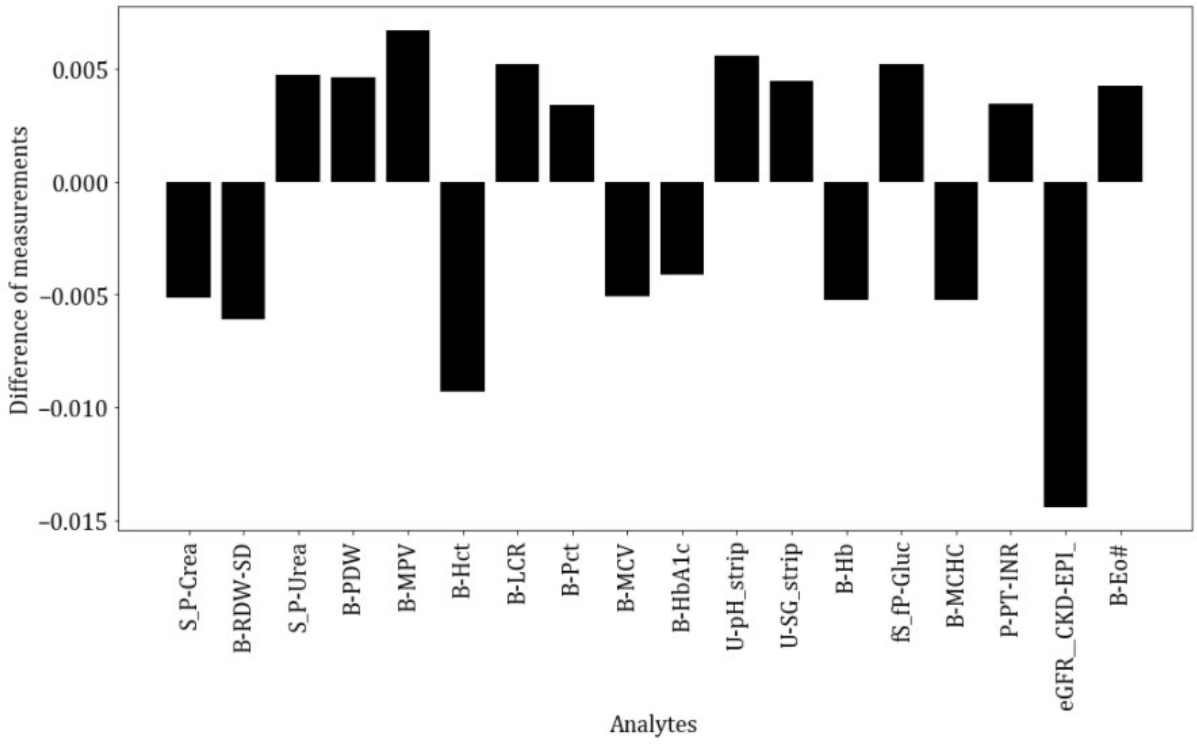

Figure 11. Biggest relative differences in number of analyte measurements between strokes and cases.

4.2.2 **Age effect**

In this chapter the differences of measurements and analytes between older and younger cohorts is discussed. The younger is defined as having birth year 1960 or later and the older cohort is defined as having birth year before 1960. Such distinction was made taking into account the distribution of analytes among different birth years.

For the older cohort, there were in total 9 948 measurements with 195 258 analytes whereas only 1493 measurements with 28 423 analytes for the younger cohort. Figure 12 presents the number of measurements per birth year for the older cohort and Figure 13 for the younger.

It can clearly be seen that the number of measurements is smaller for the younger cohort.

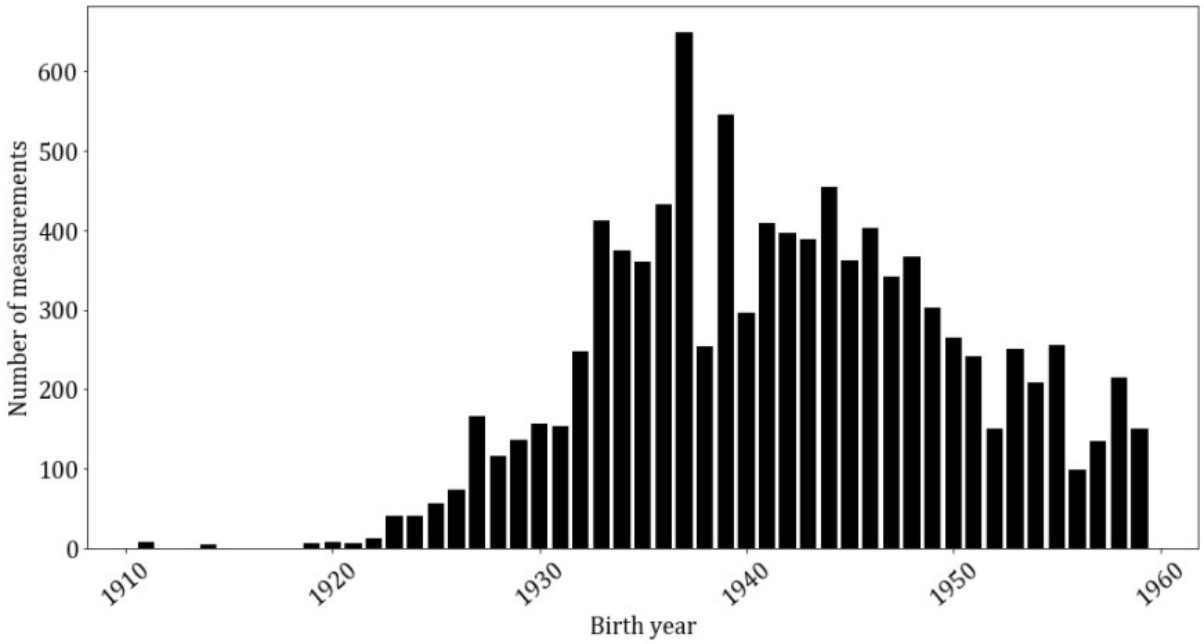

Figure 12. Number of measurements per birth year for the older cohort

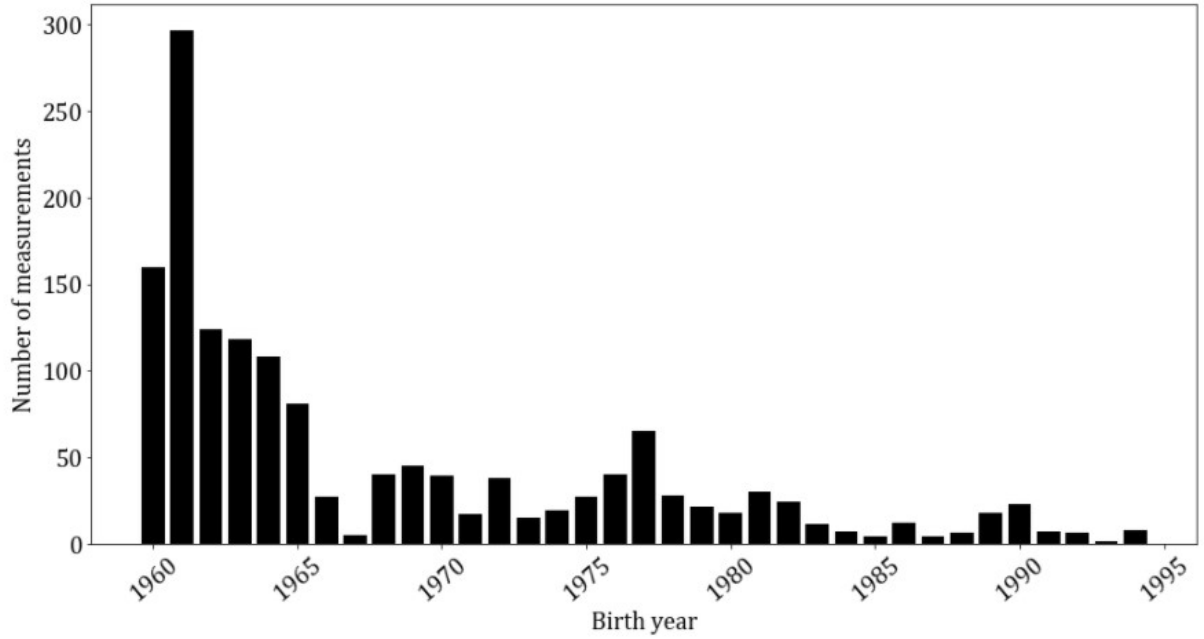

Figure 13. Number of measurements per birth year for the younger cohort

Following, the most frequent analytes among the younger and the older cohort were examined. It can be observed from Figure 14 and Figure 15 that the most frequent analytes for the older cohort were B-Hb, B-Plt, B-RBC and other parameters of hemogram, along with creatinine having up to 6000-7000 measurements for each.

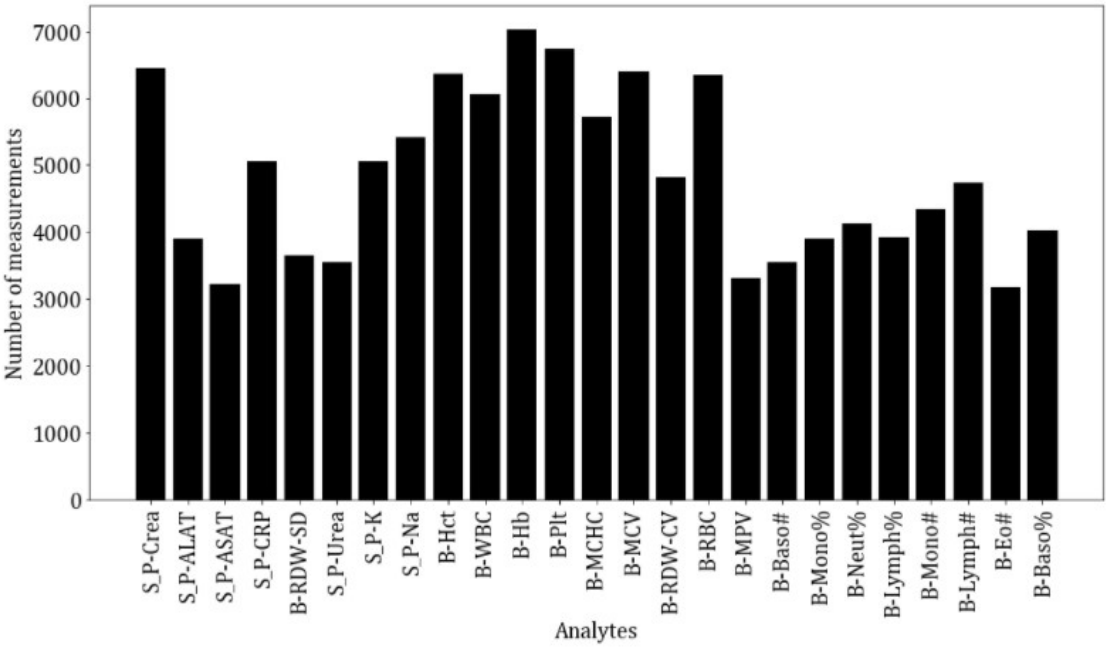

Figure 14. Most frequent analytes for the older cohort

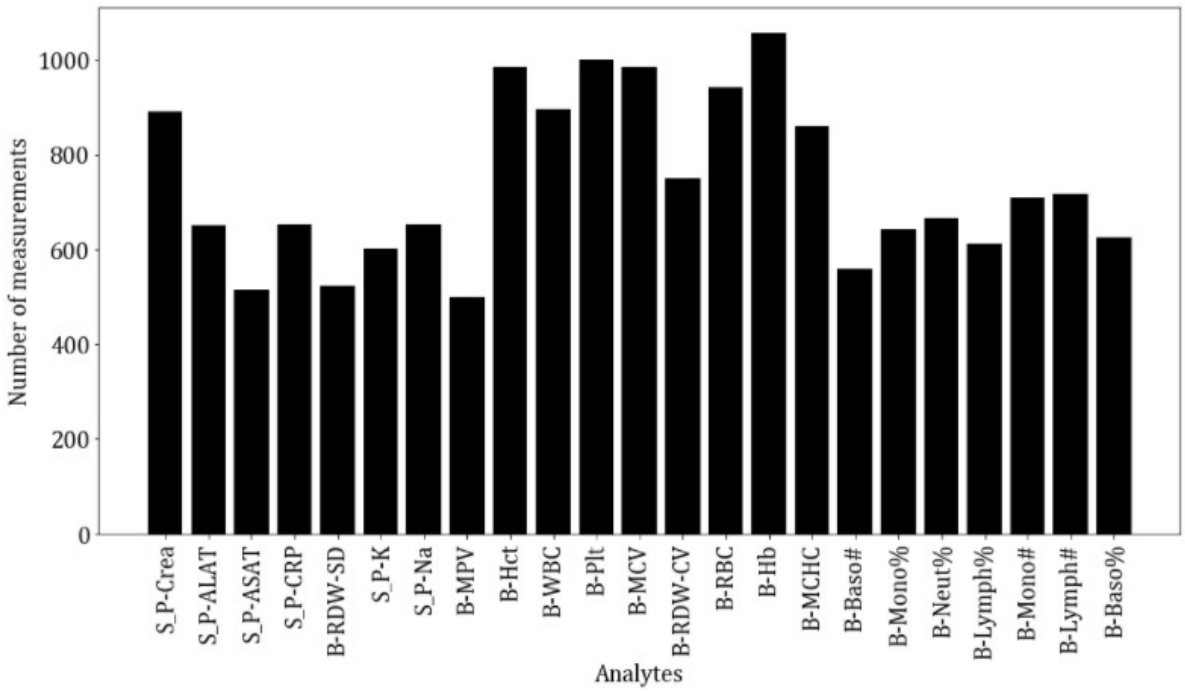

Figure 15. Most frequent analytes for the younger cohort

From Figure 16 it can be observed that the biggest differences of relative number of measurements were with S-P-K and S_P_Na and B_Mono that were often ordered for the older cohort . Whereas the parameters of complete blood count such as B-Mono# and B-Mono% were proportionally more ordered for the younger cohort.

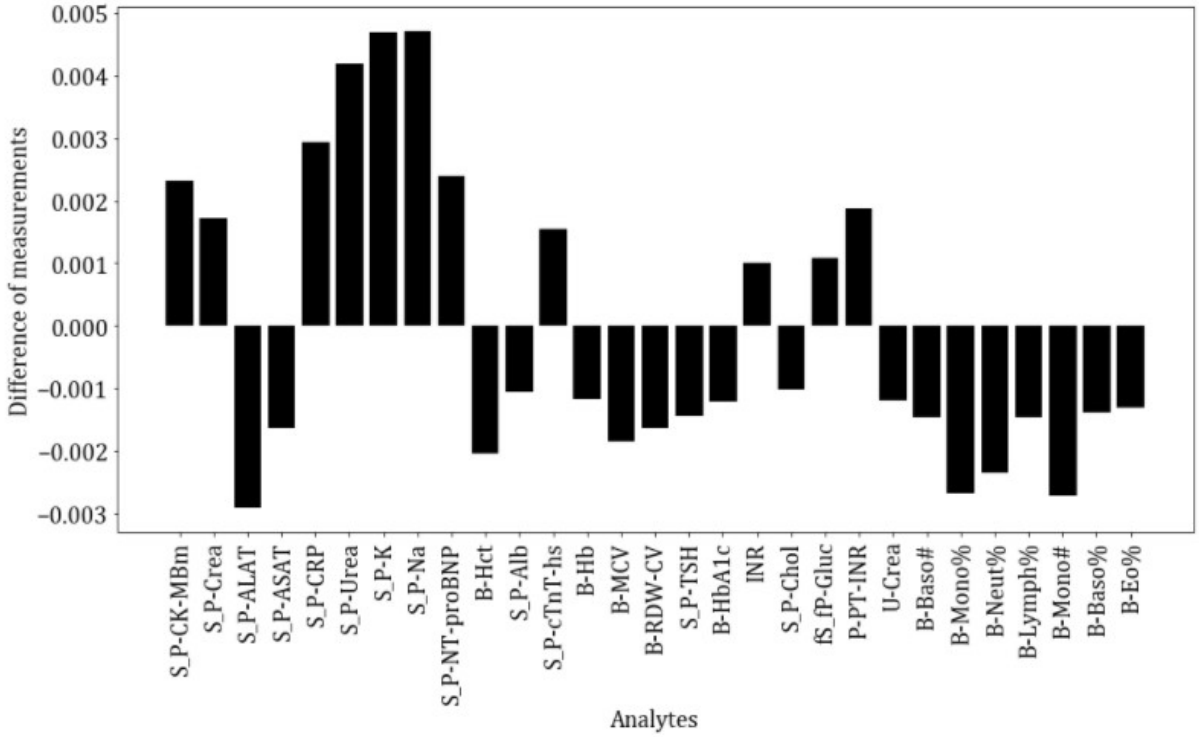

Figure 16. Differences of measurements between older and younger cohorts.

Finally, the number of clinical parameters in one measurement were examined within the older (Figure 17) and the younger (Figure 18) cohorts. The average number of analytes in one measurement for the older cohort was 19 with standard deviation of 13 and 18 for the younger population with standard deviation of 13. This indicates that on average the number of tests ordered per measurement differed little.

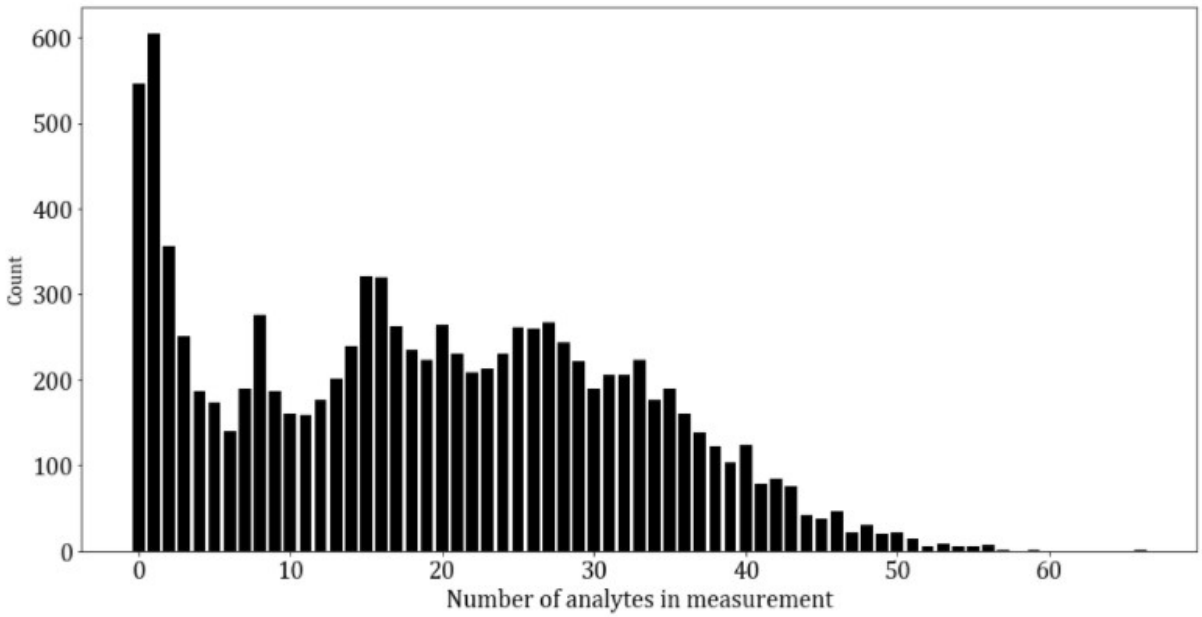

Figure 17. Number of analytes in one measurement in the older cohort

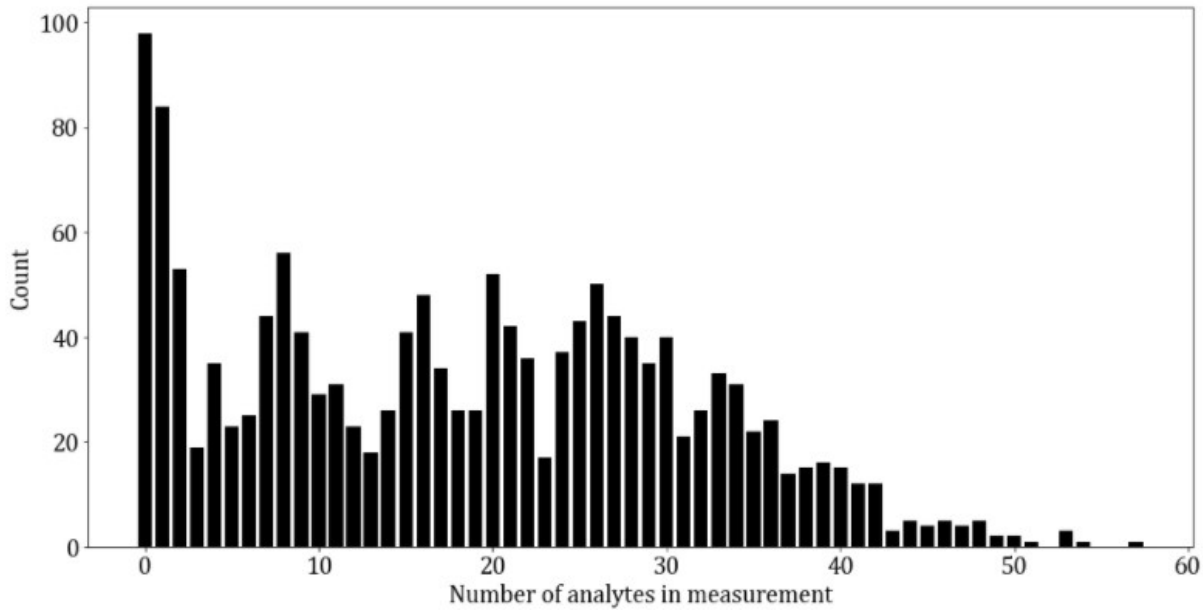

Figure 18. Number of analytes in one measurement in the younger cohort

4.2.3 **Men vs women**

In this chapter the differences in measurements and analytes between the sexes were investigated. In total there were 105184 analytes with 5204 measurements performed on men and 118 497 analytes with 6237 measurements performed on women.

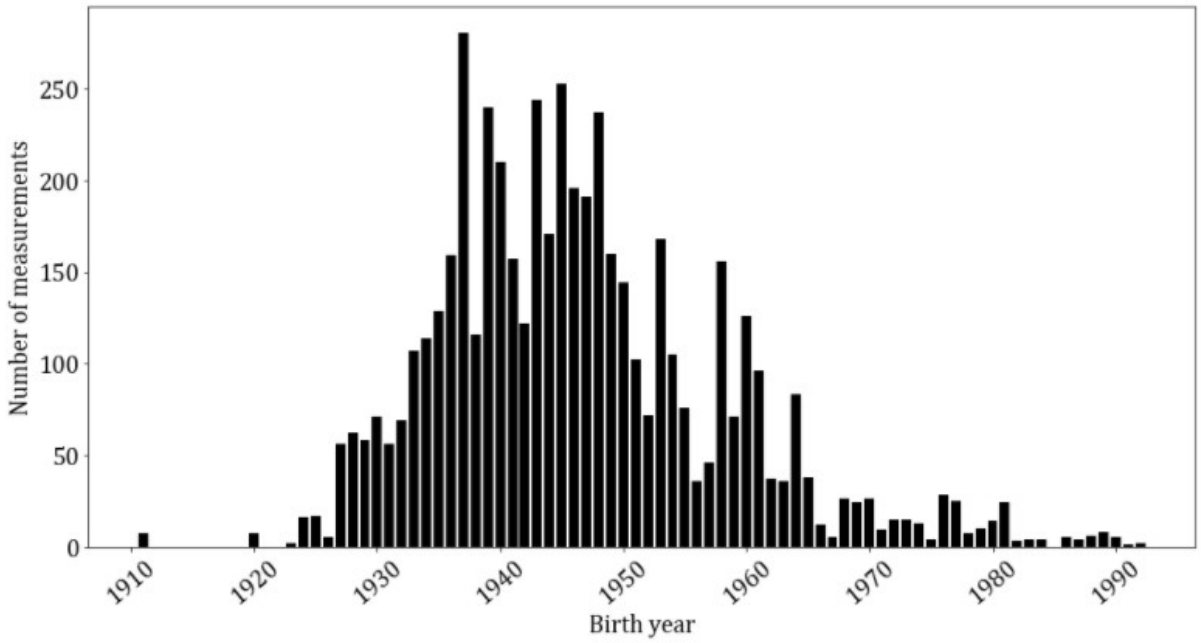

Figure 19. Number of measurements for men

It can be observed from Figure 19 that the most number of measurements were performed on those with birth year between 1940 and 1950. In case of women (Figure 20) most measurements were performed for patients with birth years between 1935 and 1945. It may due to the difference of average lifespan between men and women.

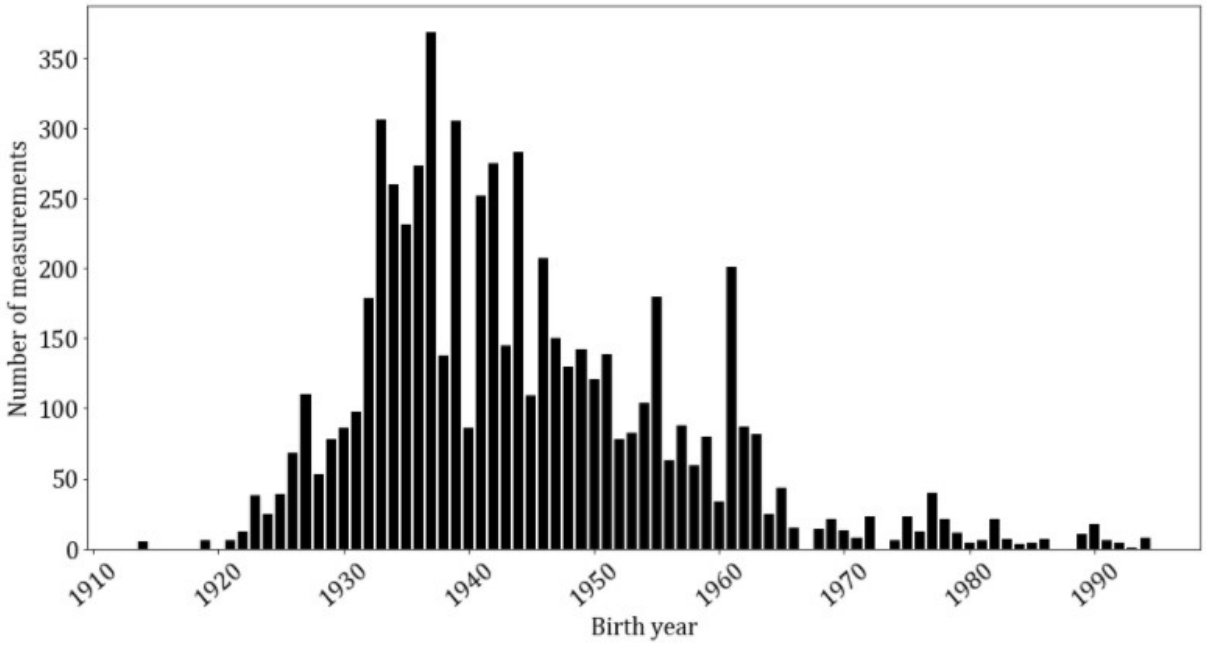

Figure 20. Number of measurements for women

Figure 21 and 22 depict most frequent analytes measured in men and women. In absolute numbers, the most frequently measured analytes were the same, being B-Hb, S,p_Crea and B-Hb.

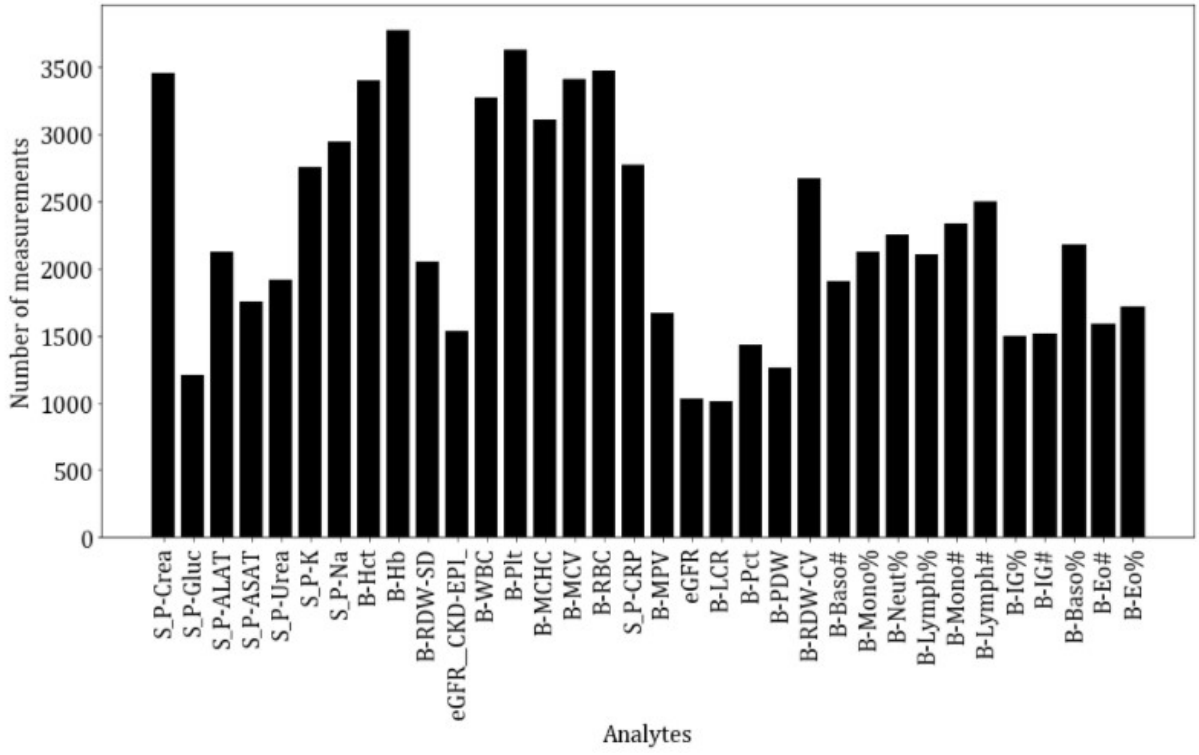

Figure 21. Most frequent analytes among men

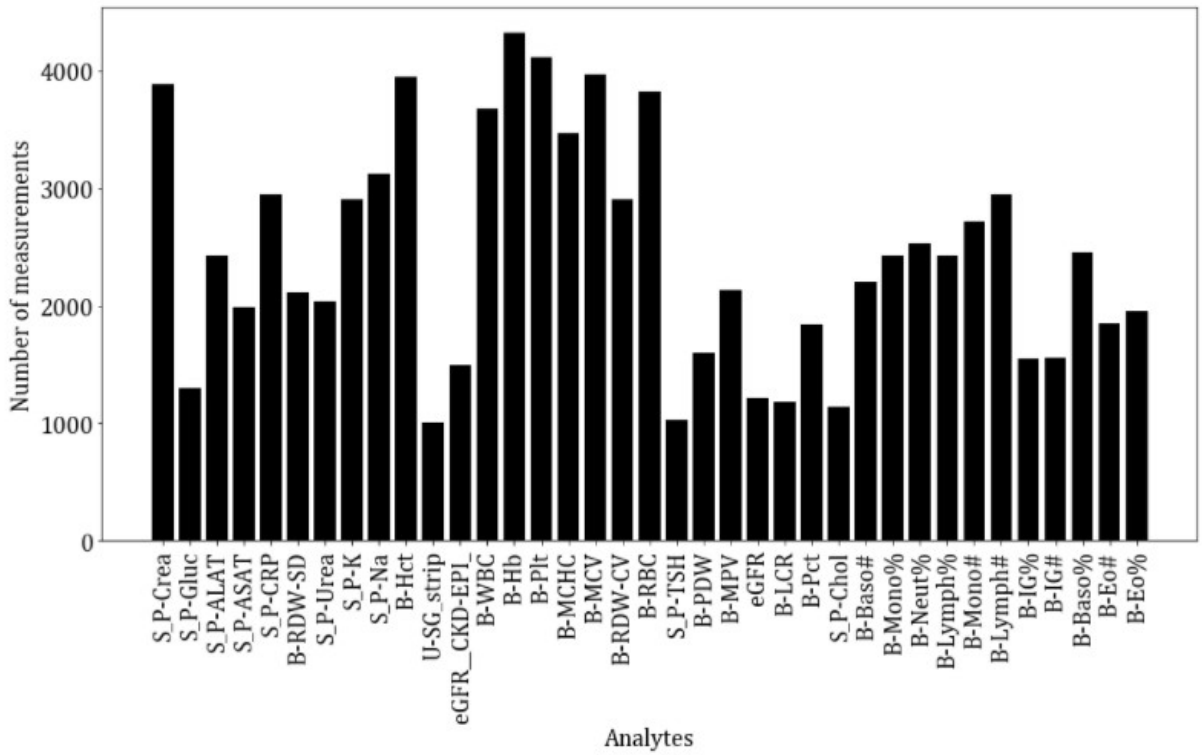

Figure 22. Most frequent analytes among women

To assess the relative difference of measurements between men and women each analyte's density in men and women was subtracted and the highest and lowest values presented in Figure 23. Thus, it can be stipulated that S-PSA and S,P-TSH had the biggest differences among men and women. However, it was to be expected, since S,P-PSA is used to estimate prostate problems and TSH is used to diagnose autoimmune diseases that are more prevalent in women.

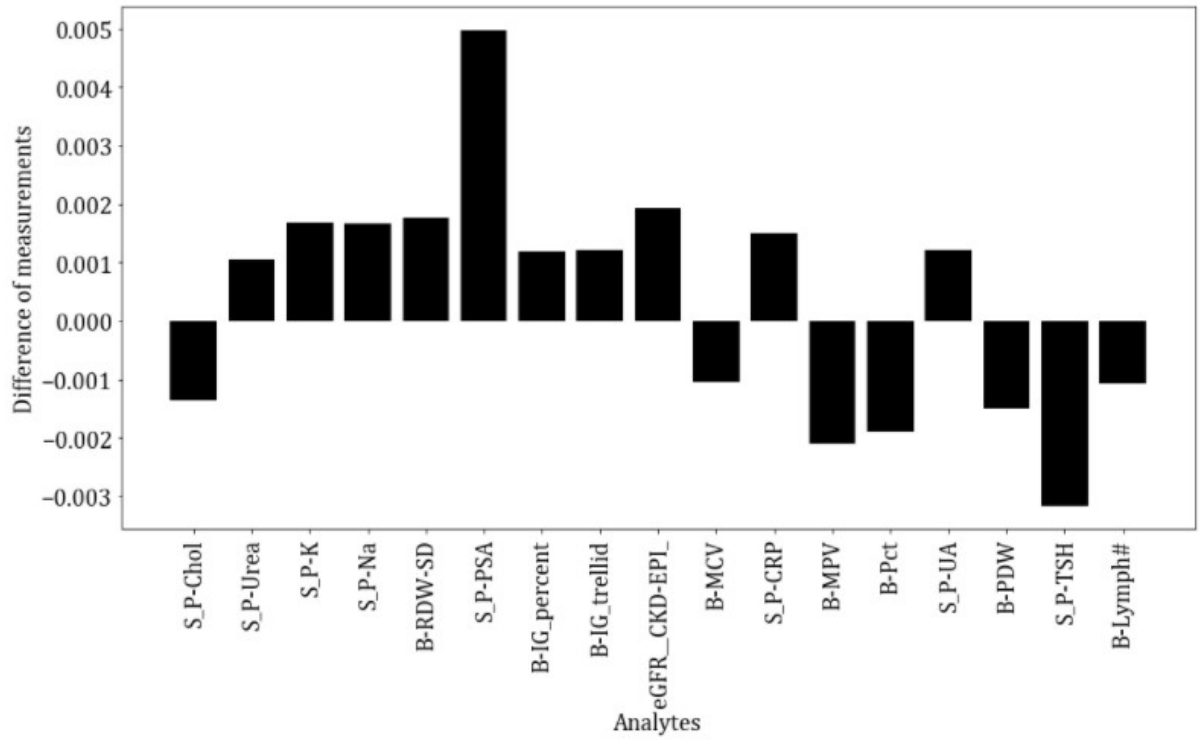

Figure 23. Relative difference of measurements between men and women

Finally, the difference in number of analytes in each measurement between men and women were investigated. Figure 24 and Figure 25 depict that the overall outline for men and women were similar, with men having an average of 20 analytes per measurement and women an average of 18 analytes per measurement.

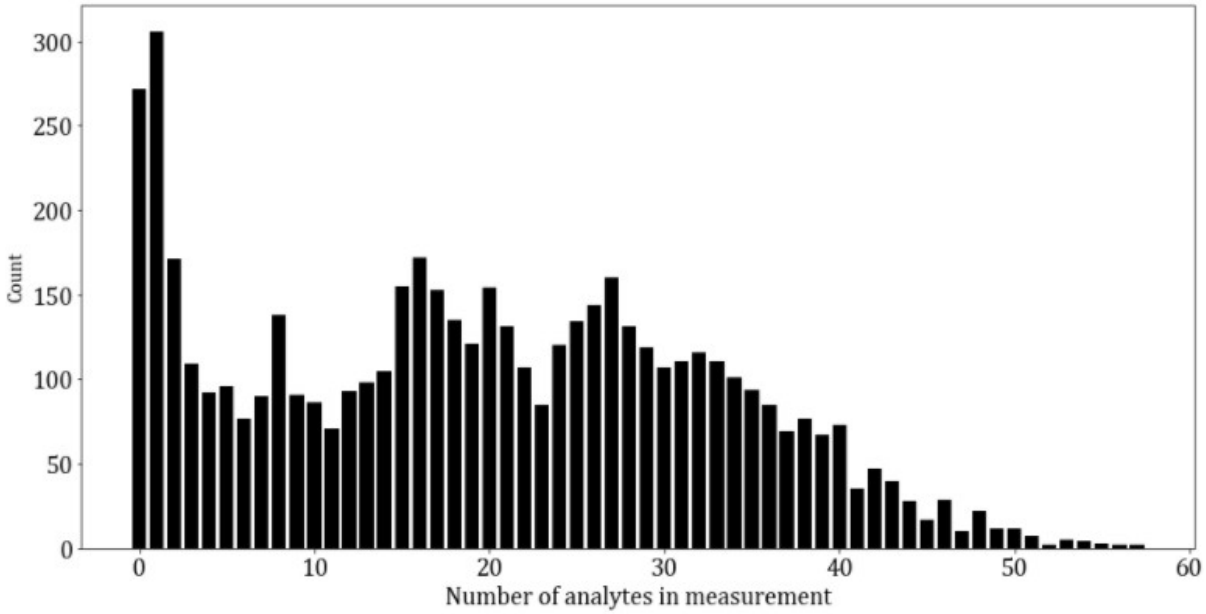

Figure 24. Number of analytes in one measurement for men

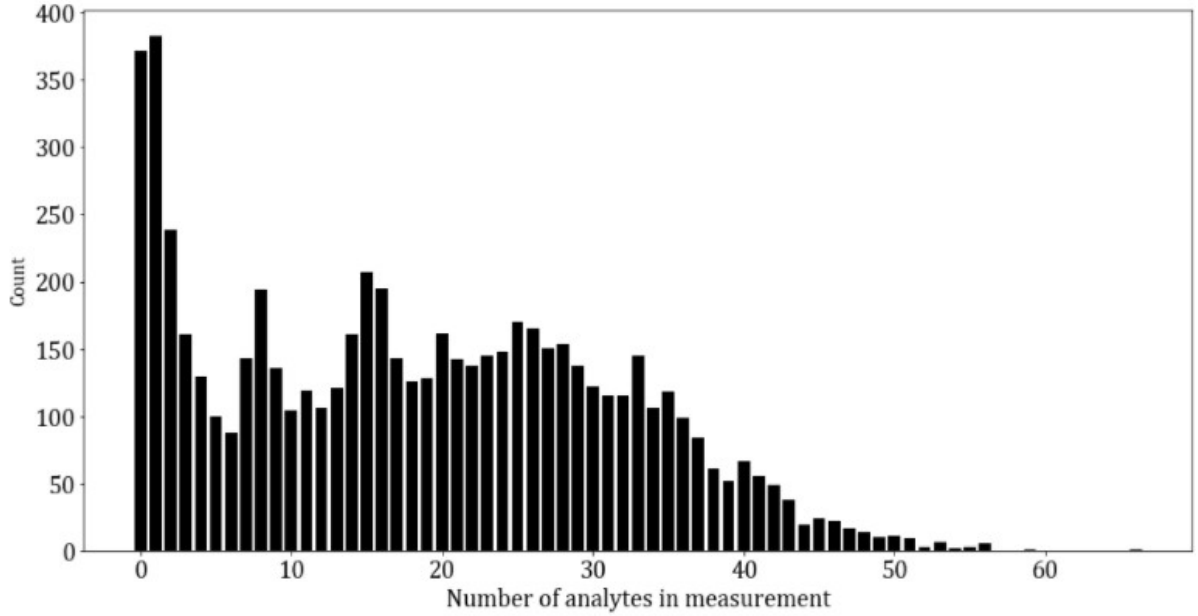

Figure 25. Number of analytes in one measurement for women

5 Results

In this chapter the results of the 5 approaches and 3 methods in 3 tiers were presented.

5.1 Baseline

The prediction results of Logistic regression, KNN and Random Forest with year of birth and sex as inputs are in Table 5. The precision, accuracy and F1-score range between 0.5 and 0.73 for all methods, thus performing better than random but still rather poorly.

Table 5. Baseline with sex and year of birth

Baseline	Precision	Accuracy	F1
Logistic regression	0.53	0.73	0.6
KNN (n= 50)	0.66	0.72	0.7
Random Forest	0.67	0.73	0.7

5.2 Results of five approaches in three tiers

The applicability of EHR data for ischemic stroke prediction was assessed in three tiers:

1) All measurements with at least 10 analytes were treated equally, regardless of how many measurements were available per person (chapter 5.2.1);
2) One random measurement with at least 10 analytes was chosen per person (chapter 5.2.2);
3) All latest analyte values were used for each person (chapter 5.2.3).

Each of the mentioned tiers was analyzed via five different approaches: the binary statement of measurements, binary measurements with equal controls and cases, medical reference values, absolute analyte values, and calculated z-scores.

All five approaches in three tiers were assessed via logistic regression, KNN and Random Forest algorithms.

5.2.1 **All measurements separately**

It can be seen from Table 6 that when all measurements with at least 10 analytes were treated equally the best combination was Random Forest on **absolute analysis values** with precision, accuracy and F1-score around 0.91. The second best combination was Random Forest on **medical reference values** with precision, accuracy and F1-score around 0.9. Random Forest is the best performing method among compared methods. In the case of calculated z-scores and binary statement of measurements, the logistic regression was worse than KNN.

Table 6. All measurements separately

Approach|Method		**All measurements separately**		
		Precision	**Accuracy**	**F1**
Binary statement of measurements	Logistic regression	0.69	0.7	0.67
	KNN (n= 50)	0.8	0.78	0.75
	Random Forest	0.88	0.88	0.87
Binary measurements with equal controls and cases	Logistic regression	0.86	0.86	0.84
	KNN (n= 50)	0.68	0.69	0.6
	Random Forest	0.86	0.86	0.85
Medical reference values	Logistic regression	0.8	0.8	0.79
	KNN (n= 50)	0.78	0.77	0.75
	Random Forest	0.9	0.9	0.9
Absolute analysis values	Logistic regression	0.8	0.8	0.8
	KNN (n= 50)	0.78	0.77	0.75
	Random Forest	0.91	0.9	0.9
Calculated z-scores	Logistic regression	0.69	0.71	0.69
	KNN (n= 50)	0.79	0.78	0.76
	Random Forest	0.89	0.88	0.88

5.2.2 **One measurement per person**

Table 7 illustrates the prediction scores of EHR data, when only one random measurement with at least 10 analytes was chosen per person. The best overall method was Random Forest, especially when the **absolute analysis values** or **z-scores** were chosen, with respective precisions of 0.91 and 0.87.

Table 7. One measurement per person

Approach|Method		**One measurement per person**		
		Precision	**Accuracy**	**F1**
Binary statement of measurements	Logistic regression	0.74	0.77	0.72
	KNN (n= 50)	0.71	0.76	0.71
	Random Forest	0.86	0.84	0.83
Binary measurements with equal controls and cases	Logistic regression	0.86	0.85	0.83
	KNN (n= 50)	0.7	0.73	0.65
	Random Forest	0.86	0.85	0.84
Medical reference values	Logistic regression	0.79	0.82	0.77
	KNN (n= 50)	0.76	0.81	0.74
	Random Forest	0.79	0.82	0.77
Absolute analysis values	Logistic regression	0.84	0.84	0.83
	KNN (n= 50)	0.8	0.79	0.77
	Random Forest	0.91	0.9	0.91
Calculated z-scores	Logistic regression	0.73	0.75	0.72
	KNN (n= 50)	0.7	0.74	0.67
	Random Forest	0.87	0.86	0.84

5.2.3 **All latest measurements per person**

For the third tier, the predictions were not measurement-based, but individual-based. For each individual the latest measurement of each analyte was chosen (Table 8). Again **Random Forest was the best method for all approaches**, however the best approach was the "Binary statement of measurements" with scores 0.92-0.93. The "absolute analysis values" approach had scores between 0.91 to 0.92.

Table 8. All latest measurements per person

Approach|Method		**All latest measurements per person**		
		Precision	**Accuracy**	**F1**
Binary statement of measurements	Logistic regression	0.85	0.86	0.85
	KNN (n= 50)	0.8	0.8	0.77
	Random Forest	0.93	0.93	0.92
Binary measurements with equal controls and cases	Logistic regression	0.78	0.79	0.77
	KNN (n= 50)	0.8	0.81	79
	Random Forest	0.88	0.88	0.87
Medical reference values	Logistic regression	0.86	0.86	0.34
	KNN (n= 50)	0.78	0.82	0.76
	Random Forest	0.86	0.87	0.83
Absolute analysis values	Logistic regression	0.84	0.85	0.84
	KNN (n= 50)	0.8	0.8	0.78
	Random Forest	0.92	0.91	0.91
Calculated z-scores	Logistic regression	0.75	0.78	0.74
	KNN (n= 50)	0.75	0.77	0.68
	Random Forest	0.89	0.89	0.88

5.3 Best correlations with ischemic stroke

The best correlating analytes with ischemic stroke were calculated for each approach to compare whether the same analytes or at least the "same type" of analytes were of importance. From Table 9 it can be observed that none of the approaches returned very well correlating analytes. The highest correlation occurred in "absolute analysis values" approach as analyte B-Hct with correlation of 0.38. In three methods, B-RDW-SD and B-RDW-CV were one of the best correlating analytes.

Table 9. Analytes with best correlations with ischemic stroke

Approach	Correlations	Analyte	Score
Binary statement of measurements	Best correlating analyte	B-RDW-SD	0.12
	2nd best correlation analyte	S_P-Crea	0.11
	3rd best correlating analyte	B-RDW-CV	0.11
Binary measurements with equal controls and cases	Best correlating analyte	B-RDW-SD	0.12
	2nd best correlation analyte	B-RDW-CV	0.11
	3rd best correlating analyte	S_P-Crea	0.10
Medical reference values	Best correlating analyte	B-MCHC_G	-0.11
	2nd best correlation analyte	B-Neut#_G	0.10
	3rd best correlating analyte	B-Lymph#_G	0.09
Absolute analysis values	Best correlating analyte	B-Hct	0.38
	2nd best correlation analyte	eGFR_CKD-EPI	0.33
	3rd best correlating analyte	B-MPV	0.26
Calculated z-scores	Best correlating analyte	B-RDW-SD	0.12
	2nd best correlation analyte	S_P-Crea	0.11
	3rd best correlating analyte	B-RDW-CV	0.11

5.4 Differences between younger and older

The best method for predicting stroke (based on previous chapters) was Random Forest with individual-based data. Thus, the differences between younger and older cohorts were investigated with the help of Random Forest. With Random Forest the prediction scores did not vary much between the sexes (Table 10) .

Table 10. Stroke prediction for younger and older population.

	Precision	Accuracy	F1	Best correlating analytes
Older	0.91	0.92	0.9	eGFR, B-Hct, B-MPV, U-pH, B-RDW-SD
Younger	0.9	0.9	0.89	eGFR, B-Hct, B-MPV, U-pH, S,P-Prot

As can be seen from the table above the best correlating analytes were mostly the same for older and younger cohort, with the exception of S,P-Prot having a better correlation in the younger cohort .

5.5 Differences between men and women

The Random Forest prediction results for men and women separately can be seen in Table 11. The prediction scores were very similar and the best correlating analytes were almost the same, with the exception of men having U-pH among the top 5 best correlating analytes. However, U-pH was in top 10 best correlating analytes among women as well.

Table 11. Stroke prediction for men and women

	Precision	Accuracy	F1	Best correlating analytes
Men	0.91	0.9	0.90	eGFR, B-Hct, B-MPV, U-pH, S,P-Prot
Women	0.91	0.9	0.9	eGFR, B-Hct, B-MPV, B-RDW-SD, S,P-Prot

6 Discussion

In this chapter the results will be discussed in the context of the main research questions.

The pre-processing part presented numerous challenges, such as inconsistent analyte names, missing units and obscure reference values. However, as shown in Chapter 5 the pre-processing resulted in success, because the prediction scores increased significantly when adding EHR data as parameters to the prediction models.

The baseline results from Chapter 5.1 assured that sex and age alone do not hold much predictive value for ischemic stroke prediction. Thereby, the search for best approach, method and tier with analytes was justified. Table 12 represents the summarized version of Table 6 to Table 8 to illustrate their differences.

Table 12. Comparison of logistic regression, KNN, RF and five approaches in three tiers

Approach|Method		Measurement-based						Individual-based		
		All measurements separately			One measurement per person			All latest measurements per person		
		Precision	Accuracy	F1	Precision	Accuracy	F1	Precision	Accuracy	F1
Binary statement of measurements	Logistic regression	0.69	0.7	0.67	0.74	0.77	0.72	0.85	0.86	0.85
	KNN (n= 50)	0.8	0.78	0.75	0.71	0.76	0.71	0.8	0.8	0.77
	Random Forest	0.88	0.88	0.87	0.86	0.84	0.83	0.93	0.93	0.92
Binary measurements with equal controls and cases	Logistic regression	0.86	0.86	0.84	0.86	0.85	0.83	0.78	0.79	0.77
	KNN (n= 50)	0.68	0.69	0.6	0.7	0.73	0.65	0.8	0.81	79
	Random Forest	0.86	0.86	0.85	0.86	0.85	0.84	0.88	0.88	0.87
Medical reference values	Logistic regression	0.8	0.8	0.79	0.79	0.82	0.77	0.86	0.86	0.84
	KNN (n= 50)	0.78	0.77	0.75	0.76	0.81	0.74	0.78	0.82	0.76
	Random Forest	0.9	0.9	0.9	0.79	0.82	0.77	0.86	0.87	0.83
Absolute analysis values	Logistic regression	0.8	0.8	0.8	0.84	0.84	0.83	0.84	0.85	0.84
	KNN (n= 50)	0.78	0.77	0.75	0.8	0.79	0.77	0.8	0.8	0.78
	Random Forest	0.91	0.9	0.9	0.91	0.9	0.91	0.92	0.91	0.91
Calculated z-scores	Logistic regression	0.69	0.71	0.69	0.73	0.75	0.72	0.75	0.78	0.74
	KNN (n= 50)	0.79	0.78	0.76	0.7	0.74	0.67	0.75	0.77	0.68
	Random Forest	0.89	0.88	0.88	0.87	0.86	0.84	0.89	0.89	0.88

6.1 Binary approaches

When binary statements of measurements were included in the models, the predictive value of the models rose considerably compared to the baseline. Previous precision of 0.67 increased to 0.93 in case of Random Forest. This indicates that having certain analytes measured implies a potential risk for stroke.

In addition, it is logical that certain tests would be often prescribed to patients at risk of stroke. If the control and stroke groups are unproportional regarding certain analytes, the models may base their predictions not on the individual specifics but rather on the descriptive statistics of controls and strokes

To counter such problem, the second binary approach with equal controls and cases was undertaken. The precision decreased in the case of logistic regression and RF, but stayed the same for KNN (at 0.8 precision). The accuracy and precision were both 0.88 indicating that the statements of measurement fact still had a major role in the prediction of ischemic stroke. The small decrease in scores compared to the first binary approach is welcomed, meaning that the bias in groups had some effect on the predictions, but not significant enough to disregard the statements completely.

For both binary approaches, the data in individual tier resulted in best outcomes. This was expected since the binary approach did not contain any result values which could theoretically blur the individual-based data.

6.2 Medical reference values approach

When assessing the predictive value of medical reference values the first tier where stroke was predicted for all measurements separately had a clearly better outcome than the other tiers. This is specific to medical reference values, since for all other approaches the third tier clearly resulted in best scores. It could be due to the fact that the reference values themselves originate from specific laboratory information systems. Thus, if hospitals use considerably different reference values, it could explain the sudden improvement in score results. Patients whose tests were performed in hospitals generally are in worse medical state than those whose tests were ordered by general physicians and thus carried out in different laboratories with distinguished medical reference values.

The continuous analysis results were classified into three categories based on lower and upper bound of the reference values. Due to such classification some of the potential value of the data was lost. Therefore, in further investigations the classifications should be more precise and detailed for each analyte.

6.3 Absolute value and z-score approaches

As in the case of binary approaches, the best data tier was individual-based and best method was Random Forest. However, since the absolute value approach resulted in better prediction scores than the z-scores, then perhaps the z-score calculations lost some relevant information, such as accompanying illnesses.

Interestingly, in case of z-scores the differences between methods were the largest compared to other approaches. The difference of accuracy between Random Forest and

logistic regression was 0.14, meaning logistic regression performed much more poorly than Random Forest. At the same time, the difference between methods in case of medical reference values approach was 0.08.

Most likely, the success of absolute analysis value approach can be attributed to the fact that real analysis values contain the test and individual-specific characteristics within.

6.4 Important clinical parameters and sub-populations

Most of the best correlating analytes were from frequently administered tests and panels, such as complete blood counts and EGFR-EPIs. Unfortunately, no clinical parameter had a better correlation with stroke than 0.38. However, it was imminent that certain analytes had more impact on model outcomes, since removing eGFR-Epi and complete count parameters from the predictions resulted in worse prediction scores.

The comparison of predictions between men and women showed that both had comparable prediction scores and the U-pH was slightly more important for men than for women.

Comparing older and younger sub-populations resulted in similar outcomes: the older and younger patients had similar prediction scores and best correlating clinical parameters. There was little data for younger population, so with larger dataset the population differences should be studied further.

6.5 Future developments

Since Random Forest performed best with all approaches, the effect of gradient boosting trees should be investigated on this data. In addition, the forest parameters should be fine-tuned and among others the minimum number of samples required for splitting a node investigated.

Furthermore, the importance of eGFR, creatinine and complete blood count parameters surfaced, indicating that their individual potential for stroke predictions should be investigated. Maybe the source of the data should be taken into account as well, considering that the laboratory instrumentation may vary considerably.

7 Conclusion

Despite multiple challenges and hardships offered by the EHR data, the finalized models performed well for ischemic stroke prediction. Thereby, Estonian Health Record data can be used to predict ischemic stroke. However, the binary statement of measurements contained enough information to create good prediction models, indicating that the tests ordered by physicians already point towards the underlying illness. This questions the usefulness of real analysis values as input for prediction models.

A decline in prediction scores was observed when the binary statements of measurements were equalized in control and case groups. This pointed to the fact that some analytes were ordered more often for the case group had an effect for the model outcome. Thus, the same logic should be applied to absolute values approach as well to determine, whether its scores would also decline.

Nevertheless, it was clear that some clinical parameters were more important to all prediction models, such as complete blood count and urine pH. Following, the potential of prediction models should be investigated with respect to mentioned analytes only and a more precise analysis targeting most interesting clinical parameters created.

The best predictions were received on data which was pre-processed with the goal to maximize clean analyte result values. The pre-processing workflow with aim to maximize medical reference values resulted in success as well, having close to 0.9 precision and accuracy scores.

From selected methods, Random Forest outperformed Logistic Regression and K-Nearest Neighbours in all investigated approaches and tiers. Using analysis result values as input, the Random Forest resulted in accuracy and precision above 0.9 Thus, the applicability of Random Forest on EHR data should be investigated further.

As a conclusion it can be stated that the pre-processing resulted in success and viable models were created that prove predicting ischemic stroke on health data is possible.

8 References

[1] Sivenius J, Torppa J, Tuomilehto J, Immonen-Räihä P, Kaarisalo M, Sarti C, Kuulasmaa K, Mähönen M, Lehtonen A, Salomaa V, Modelling the burden of stroke in Finland until 2030, *National Library of Medicine*, 2009 Oct

[2] SYNLAB Estonia OÜ, https://synlab.ee/arstile/laboriteatmik/ (May 10th, 2021)

[3] Kratz A, Ferraro M, Sluss P.-M, Lewandrowski K. B, Laboratory Reference values, *The New England Journal of Medicine*, 2004;351:1548-63

[4] Allan S, Olaiya R, Burhan R, Reviewing the use and quality of machine learning in developing clinical prediction models for cardiovascular disease, *National Library of Medicine*, 2021 March,

[5] Buck B.H, Akhtar N, Alrohimi A, Khan K, Shuaib A, Stroke mimics: incidence, aetiology, clinical features and treatment, *Annals of Medicine*, March 2021,

[6] US Centres for disease control and prevention, https://www.cdc.gov/stroke/types_of_stroke.htm (May 10th 2021)

[7] Montaner, J., Ramiro, L., Simats, A. *et al.* Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. *Nat Rev Neurol* **16,** 247–264 (2020).

[8] US National Heart, Lung, and Blood Instititue, https://www.nhlbi.nih.gov/health-topics/stroke (May 4th 2021)

[9] Donnan G, Fisher M, Macleod M, Davis S M, „Stroke“, *Lancet* 2008; 371: 1612–23

[10] TEHIK, https://elhr.digilugu.ee/data/laboridList.html, April 15th 2021)

[11] Bosshart, M., Stover, J.F., Stocker, R. *et al.* Two different hematocrit detection methods: Different methods, different results. *BMC Res Notes* **3,** 65 (2010).

[12] SYNLAB Estonia OÜ, https://synlab.ee/arstile/laboriteatmik/analuuside-koondtabel/ (April 20th, 2021)

[13] Mitchell T, Hill McGraw, Machine Learning, 1997.

[14] Sindhu V, Nivedha S, Prakash M, An empirical science research on bioinformatics in machine learning, *Journal of Mechanics of Continua and Mathematical Sciences*, February 2020,

[15] Walker S. H, Duncan D. B, Estimation of the probability of an event as a function of several independent variables, *Biometrika*, Volume 54, Issue 1-2, June 1967, Pages 167–179

[16] Sarkar, S. K. and H. Midi. “Importance of Assessing the Model Adequacy of Binary Logistic Regression.” *Journal of Applied Sciences* 10 (2010): 479-486.

[17] Habshah Midi, S.K. Sarkar & Sohel Rana (2010) Collinearity diagnostics of binary logistic regression model, *Journal of Interdisciplinary Mathematics*, 13:3, 253-267

[18] Lynam, A.L., Dennis, J.M., Owen, K.R. *et al.* Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagn Progn Res* 4**,** 6 (2020)

[19] Shouman M, Turner T, Stocker R, Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients, *International Journal of Information and Education Technology*, 2020, 2(3):220-223

[20] N. S. Altman (1992) An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression*, The American Statistician*, 46:3, 175-185

[21] Nenad Tomašev, Krisztian Buza, Hubness-aware kNN classification of high-dimensional data in presence of label noise,*Neurocomputing*, Volume 160, 2015, Pages 157-172, ISSN 0925-2312,

[22] Tin Kam Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, pp. 278-282 vol.1,

[23] Piryonesi, S. M. and T. El-Diraby. Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation Engineering, Part B: Pavements,* June 2020, Volume 146.

[24] Lee H, Lee EJ, Ham S, Lee HB, Lee JS, Kwon SU, Kim JS, Kim N, Kang DW. Machine Learning Approach to Identify Stroke Within 4.5 Hours. Stroke. 2020 Mar;51(3):860-866

[25] Hastle T, Tibshirani R, Friedman J-H, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, February 2009

[26] Ross P, Mill R, Digitaalsete dokumentide jagamise standardprofiilid tervishoius, *Eesti Arst* 2013; 92(9):516–523

[27] STACC, https://www.stacc.ee/en/ (April 27th, 2021)

[28] Prins BP, Leitsalu L, Pärna K, Fischer K, Metspalu A, Haller T, Snieder H. Advances in Genomic Discovery and Implications for Personalized Prevention and Medicine: Estonia as Example. *Journal of Personalized Medicine*. 2021; 11(5):358.

[29] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011

License

Non-exclusive license to reproduce thesis and make thesis public

I, Ainika Adamson,

1. herewith grant the University of Tartu a free permit (non-exclusive license) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   "Assessment of the suitability of the Estonian Health Record data for the prediction of ischemic stroke", supervised by Toomas Haller and Kaur Alasoo

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons license CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive license does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Ainika Adamson

13.05.2021