Universiteit Leiden
The Netherlands

Opleiding Informatica

Analysing Classification Of Episodes From Electronic Health Records With The Use Of Text Mining

Wouter Ebing

Supervisors:
Suzan Verberne & Margot de Waal

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

23/07/2021

Abstract

Electronic Health Record (EHR) episodes are records kept by General Practitioners (GPs) regarding consultations with patients. This research focused on analyzing aspects of the classifiability of EHR episodes. This was done by predicting an episodes classification code, originating from the International Classification of Primary Care (ICPC), using the episodes descriptions as features. The metrics obtained using labelled train and test sets for this situation were analyzed and the most interesting codes were looked at individually. It turned out that the classifier used was able to predict ICPC-codes with a success-rate of 0.65. When looking at the most frequent codes, the classifier generally had a higher success-rate. One exception is the code *A97*, a code uses general or not specified illness(es), of which only about 11% of the times it is predicted it is correct. This is due this code not seeming to have distinguished words in its description. Most cases where the model was not able to predict an episodes ICPC-code correctly, originate from an episode description that contains insufficient information. In addition to this the difference between EHR software packages used to store EHRs have been investigated. The most frequent EHR software had the lowest success-rate, the least frequent the highest. This could be caused due to different distributions of ICPC-codes in EHR software types, since the most occurring EHR software had the highest amount of different unique ICPC-codes. Certain ICPC-codes also seem to occur more in certain software, relatively to other software. The results of this research show that episodes descriptions are not always informative about their ICPC-code and certain EHR software packages are used more often to store episodes with certain codes.

Contents

1 Introduction

One of the motivators to store data digitally, is that it can be shared more easily between individuals and organisations. One of the fields where this is visible is that of General Practitioners (GPs). Patients have the right to view records kept by medical professionals [Gen19], which can be supplied more easily if they are stored digitally using Electronic Health Record (EHR) software. An additional benefit of storing these records in this way is that they are able to be analyzed in large quantities. This research will analyze aspects of the classification of these records within these EHR software. This is a point of interest, since the classification is performed manually by the GP and there is room for interpretation within labelling decisions, potentially leading to inconsistencies in the data.

1.1 Electronic Health Records

Electronic Health Records (EHRs) are records regarding patients who have visited a general practitioner. These records are stored within systems using Electronic Health Record software packages (EHR software). Their content is divided in so called episodes, that contain data from a (group of) consults between a general practitioner and their patient. These episodes are classified by attaching a code describing the reason for the consult, such as a symptom, diagnoses, or treatment. The codes used originate from a system called International Classification of Primary Care. In addition to these codes the episodes can also contain a textual description filled in by the general practitioner.

1.2 International Classification of Primary Care

The International Classification of Primary Care (ICPC) is a classification system using codes that describe symptoms, diagnoses and treatments in the domains of general practice/family practice and primary care [Won20]. As described before these codes are assigned by the general practitioner filling in an episode for a EHR. Every code has a unique meaning, which is often similar to the content in the descriptions of episodes. However, this is not a enforced rule, which makes some descriptions less descriptive about the ICPC-code assigned then other cases. For example, a description containing the text “Migraine” indicates that the episodes ICPC code will be the code for migraine, while a description with the text “Alcohol” could indicate multiple different codes relating to alcohol problems.

1.3 Research Questions

This research will aim to answer three different questions that are related to ICPC-codes and descriptions from episodes in Electronic Health Records. The questions are the following:

How does a text-based classification model perform when classifying EHR episodes? This first question is intended to give insight into how descriptive episode descriptions are about their episodes ICPC-code. A text-based model is trained and tested on data labelled by GP’s, which produces a set of metrics usable to perform evaluations. The features for this model are episodes descriptions, the target values are episodes ICPC-code. Answering this question could help understand how descriptions are used when filling out information in EHRs.

How does classification quality differ between individual ICPC-codes in EHR episodes? The second question builds on the first one by looking at a specific subset of ICPC-codes. The training and testing for this question will be performed similarly to the previous one. The ICPC-codes further researched individually for this question are the six most frequent codes found in the used data. Twenty predictions where the classifier was confident and incorrect about its prediction will also be observed. Looking at these frequent codes and incorrect predictions could show unique code properties or smaller trends that are less obvious when looking at the full data.

To what extent does EHR software differ in class frequency in episodes and their classifiability? Lastly, this question aims to find differences between different EHR software. The text-based model to evaluate classifiability will be trained in the same way as the other research questions, but will now be tested three different times on the different EHR software occurring in the data. This tells us about amount of descriptions in these systems that are informative about their episodes ICPC-code.

2 Background

2.1 Data Origin

The data that is used in this research has been obtained from the Extramural LUMC Academic Network (ELAN). The Elan is a collaboration between caregivers within the regions of northern "Zuid-Holland" and the "Haaglanden", the Leiden University's Medical Center's (LUMC) Public Health and Primary Care department, and the LUMC Campus Den Haag. The Electronic Health Records in the data orignate from a range of general practices from the regions of the "Haaglanden", northern "Zuid-Holland", and the city of "Zoetermeer".

2.2 EHRs and ICPC-codes

As described in the intro section about Electronic Health Records 1.1, EHRs are records regarding patients who have visited a general practitioner. These records contain data in the form of episodes that describe (groups of) consults with general practitioners. They are classified using a system called International Classification of Primary Care, which is supplemented with a textual description. ICPC-codes follow a certain format which can be used to understand their meaning. Using an example code *L15.2* as an example, it can be read by splitting it up into three different parts:

L - The first symbol of the code is always a letter ranging from A-Z or a "-", this symbol describes the type or the location of the problem and can be called a chapter.

15 - The numbers after the letter indicate what the code is describing:
01-29: Symptoms/conditions
30-69: Treatments
70-99: Diagnoses
This can be called a title.

.2 - The number after the dot, the subtitle, indicates a specification within the section of the first numbers. Not all combinations of chapters and titles can have subtitles.

The assigned ICPC-code needs to be the one that fits the best, which is why a code does not always have a subtitle. A code should describe the subject as specific as possible, but should not cause false security by being too specific in uncertain situations.

2.3 Related Work

Improving the quality of EHR recording in primary care: a data quality feedback tool
This paper written back in 2016 had the aim of developing a tool to evaluate EHR recording quality. They were successful in stimulation quality improvements by highlighting differences between general practices and EHR software packages [vdBKTV+17]. Relating to this research, I will try to find differences between certain EHR software packages. I will not be creating a tool for this however, my finding will be achieved by exploring the data myself. My results might also not impact improvements directly, but could inspire future work that will.

Selecting relevant features from the electronic health record for clinical code prediction In a different paper published in 2017, a group of researches also investigated EHRs with the purpose of predicting code. Their intent was to use feature selection in order to reduce information from EHRs to representations with a consistent quality and less information overlap [SCL+17]. As with my research, their intention was to predict codes for classifying EHR data. In contrast however, they wanted to improve the data that can be used for prediction instead of analyzing results from predicting. The codes their prediction was focused on was also a different system than the International Classification of Primary Care, since their records originated from hospital data instead of general practices.

De-identification of Norwegian Health Record Notes: An Experimental Approach In 2013 a paper was published in Norway about the de-identification of Norgwegian Health Record Notes, which could be compared to EHR episode descriptions. The research aimed to develop a tool that could automatize the de-identification of these notes, in order to replace the process of manual de-identification [BS13]. As was discovered during the data exploration of my research, the data in EHR episode descriptions appeared to include personal information. This created an extra step for the research, writing a tool to anonymize the data. This step could be seen as a very basic version of the Norwegian research, since in my research it was not the main focus.

Do GPs know their patients with cancer? Assessing the quality of cancer registration in Dutch primary care: a cross-sectional validation study A paper from 2016 focused on Dutch primary care, specifically the registration of cancer diagnoses in EHRs. The research aimed to use cross-sectional validation to link data from EHRs to data from the Netherlands Cancer Registry [SRS+16]. Although different than my research in the way of classification, cross-sectional validation instead of a classification model, both use data from EHRs that relate to classification of medical data.

3 Data

The data used was obtained within two different related tables. These tables contained at least the following data:

Table 1	277,558	Patients	Patient-ID, Region, EHR software
Table 2	2,154,137	Episodes	Patient-ID, ICPC-code, description, Location-ID, date

Although more features could have been extracted from the tables, only the ones stated above are processed in this paper. In this section these features are explored separately via quantitative and/or qualitative observations.

3.1 ICPC-codes

Due to the data containing 2,154,137 episodes, there are as many ICPC-codes available. The use of specific ICPC-codes within classification is not evenly distributed, since there are some codes that occur thousands of times, while some occur only around ten times.
As explained in 2.2, a sub-set of ICPC-codes can contain subtitles. By leaving these subtitles out of the picture, the amount of different classes is reduced. The amount of unique ICPC-codes in the data can be found in table 1. For the experiments in this research, subtitles will be removed from episodes.

Subtitle	**Unique ICPC-codes**
With	1269
Without	705

Table 1: Amount of unique ICPC-codes with and without subtitles.

3.2 Descriptions

As with the ICPC-codes, there are 2,154,137 descriptions from episodes in the data. Usually these descriptions contain brief clarifications on the ICPC-code chosen. The information given in the episode description can be the definition of the assigned code, but also symptoms or injuries related to it. However, after going over a section of the descriptions not all seem to be related to the ICPC-codes from their episodes. An example of this would the following, based on an actual description from the data: “mailadres zoon: example@hotmail.nl”. Although how much a description is related to its assigned code is not easily measured, there are 89 empty descriptions for which any usability can be ruled out.
A selection of the episode descriptions also appeared to contain private information like names, email-addresses and phone-numbers. Fictive examples of descriptions containing this kind of information can be seen in table 2. The examples shown do not contain medical data, however there were entries where a description contained both personal and medical information. This added an extra step of anonymising descriptions to the research, which will be explained in the methods section 4.1.2.

ICPC	Description
Z16	Contactpersoon zoon 0642546873
A97	email: b.jansen@bnp.nl
W90	bevallen van zoon Peter

Table 2: Fictive examples of episode descriptions containing personal information.

In order to get a grasp of the lengths of descriptions, a graph has been generated of the amount of words per description. This graph can be found in figure 1

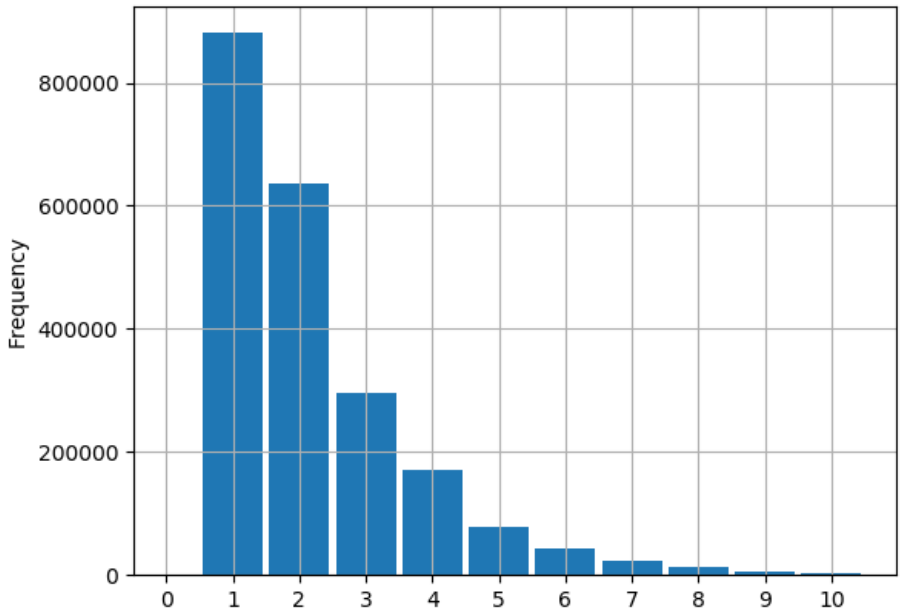


Figure 1: A density plot of the amount of words in episode descriptions.

The average amount of words per description is 2.30 and the median is two words. There are 5,654 descriptions with more than ten words with the longest description containing 48 words.

3.3 Additional EHR features

3.3.1 EHR Software types

The EHR-data contained the kind of EHR software a patients EHR is stored in. Even though these are not the only types used in medical care, the data contained the following EHR's: "Medicom", "Microhis", and "Promedico-VDF". Table 3 contains statistics relating the episodes to the EHR software. One observation deducted from it is that Medicom is the most prominent software type in the data.

	Medicom	Microhis	Promedico-VDF
Total episodes	1,716,027	58,644	20,008
Unique ICPC-codes	704	655	599
Unique patients	236,629	8,882	3,362

Table 3: Episode and ICPC-code differences between EHR software

3.3.2 Dates

The episodes in the data almost all contain a diagnosis date from around 2019. This means that the data does not inherently represent the use of EHRs from before this time period. The diagnoses that do relate to before 2019 are often known chronic diseases for which there was at least one contact moment between patient and general practitioner in 2019. There are however some episodes that contain dates that are not from 2019, but in most cases these from far outside 2019. Examples of these dates are "0001-01-0" and "2103-10-16".

3.3.3 Regions

The locations of general practitioners from the data are from three different regions: the city of "Zoetermeer" , the region "Haaglanden", and northern "Zuid Holland" (NZH). Table 4 relates these regions to ICPC-statistics.

	Zoetermeer	Haaglanden	NZH
Amount of episodes	852,854	632,968	441,719
Unique ICPC's	701	701	699

Table 4: Entries with a certain region

4 Methods

Since the focus of this research is an already collected data set, this section will describe the actions performed on the data. These actions can be split into three different chronological categories:

1. Prepare the data
2. Generate a classification model
3. Perform experiments

4.1 Preparing the data

In preparation of using the data for a classification model and other processing, it was necessary to address observations found during exploration of the data.

4.1.1 Removing subtitles

It was found that the distribution of ICPC-codes within episodes is largely long-tail. In order to make some classifications more frequent, the subtitles from all codes were removed. In this way ICPC-codes with the same chapter and title but different subtitles, could be grouped together.

4.1.2 Anonymising descriptions

During scanning through a section of the descriptions, peoples personal information was found. The decision was made to try and track down as much of this information by writing a filter. This way the personal information could be redacted to make the descriptions more anonymous. How the filter searched for the types of information is described in their respective following sections.

Names The filter found names by comparing words in the descriptions to a list of first names from Netwerk Naamkunde (NN) [Naa]. A few names were removed from the list of NN, for example when they were similar to the name of a syndrome. These cases were found by looking at words the filter detected as names to find false flags. There are more names still within the data, but due to the noise in the text and the variety in human names these have not been extracted.

Phone-numbers and email-addresses These have been found by looking for specified formats of these types of data. For phone numbers both mobile and national numbers are searched for via regular expressions. These expressions can be described as:

- *Phone-numbers:* 10 numbers with possibly a space or dash after the first two, three, or four, e.g. 06-12345678 and 0252-123456.
- *Email-addresses:* A string of letters followed by a "@", then another string, then a ".", then another string, e.g. example@mail.com.

4.2 The classification model

In order to generate a classification model process its metrics and results, python has been used in combination with the library Scikit-learn [PVG+11]. Scikit-learn has a variety of classification models that are suitable for text classification. The type of classifier used here is a Support Vector Machine classifier with Stochastic Gradient Descent (SGD) training [Sci20b]. This model will be given descriptions from episodes and will be tasked to output an ICPC-code that best suits the description. Before the descriptions are able to be inputted as features they are put through a Count Vectorizer [Sci20a] that uses raw term counts, the number of times that a term occurs in a description. This Vectorizer was set to learn up to 1000 of the most used words of which it counts the occurrence in descriptions.

4.3 The Experiments

The experiments performed on the data have been split up into three different steps, in order to grasp the capabilities and limits of using a text classification model to classify descriptions under ICPC-codes:

- Classifier performance evaluation
- Classifier prediction analysis
- EHR software comparisons

Splitting the experiments in this way, allows for a logical order of analysis: Select and analyze the performance of a certain model, analyze its classification decisions, check for possible influence of an external property.

4.3.1 Classifier performance evaluation

In order to get a general idea of the performance of the classifier, the f1-score, recall, precision, and accuracy of the model will be generated on different ratios of training and test sets alongside the support. These metrics will be generated individually per code, from which micro and macro averages of them can be calculated. The previously mentioned terms describe the following:

- *Recall:* For a target class, the number of true positive assignments divided by the number of true cases.
- *Precision:* For a target class, the number of true positive assignments divided by the number of times it was assigned.
- *F1-score:* A measure balancing both recall and precision by calculating their harmonic mean.
- *Support:* The frequency of a specific value in the test set.
- *Accuracy:* The average precision of all target values, weighted by their support.
- *Micro average:* Calculate the average of each data point. In this scenario each data point is equally weighted.

- *Macro average:* Calculate the average of each target class, then take the average of these averages. In this scenario each target class is equally weighted.

In this report we will not focus on the macro averages. The used data contains a lot of target classes with low frequencies and low results, but also the inverse. This results in macro averages that are neither informative about the full data, nor individual target values. A confusion matrix for the twenty most-occurring codes will also be visualized to possibly find pitfalls in certain classifications.

4.3.2 Classifier prediction analysis

To get insight into decision the model makes regarding classification, the weights it assigns to words for the six most occurring ICPC-codes will be put into tables. In addition too that, examples will be given of discrepancies within results of the model: situations where the model is fully confident about a wrong prediction.

4.3.3 EHR software comparisons

Lastly, I make different comparisons between EHR software, to find possibly interesting observations. These comparisons will exist of calculating a chi-squared table to find outlier ICPC-code frequencies between EHR software and testing the model on episodes per software. The chi-squared table will only be shown for the 25 most occurring codes, since showing the whole table would be too long and these contain examples of outlier frequencies. The ten outlier ICPC-codes per software that have the largest relative frequency differences will also be shown via figures. These figures will show the percentage of episodes classified by these these codes per EHR software type.

5 Results

5.1 Anonymizing descriptions

Before the results of the experiments could be generated, first the filter used to anonymize the descriptions was executed. This filter replaced the information found by general strings. The filter found the following private information and replaced each with:

Type	Amount found	Replaced with
Names	319	[Naam]
Phone-numbers	2,179	[telefoonnummer]
Email-addresses	67	[email]

5.2 Classifier performance evaluation

The experiments performed in this section are used to grasp the general performance of the used model.

5.2.1 Classification Quality – Averages

After a trained model was applied to a labelled test set a report was generated regarding the classification results. Such a classification report consist of a list of every possible target-value, which is every ICPC-code present in the test set. Each of these codes are then followed by their f1-score, precision, recall and support. The report is ended with a summary of the macro and weighted average of all classes for the before mentioned metrics, together with the accuracy of the model. In table 5 the weighted averages are shown for different ratio's of train and test sets. in these experiments the training and test splits have been obtained using clustered sampling, where clusters are based on locations. This means that when splitting the data into train and test set, episodes for a certain location would either be all put in the train or all in the test set.

Train:Test	F1-score	Precision	Recall	Support
10:90	0.659	0.738	0.652	1,734,787
25:75	0.659	0.744	0.649	1,445,656
50:50	0.660	0.743	0.651	963,771
80:20	0.659	0.746	0.647	385,509

Table 5: Weighted average metrics with different Train:Test ratio's.

As can be derived from the results, expanding the training set does not seem to influence the results of the model too much. With a larger training set the precision increases by a bit and the recall gets lowered by a small amount. This could be because the more frequent ICPC-codes become more prominent in larger sets, which makes the model assign them more than less frequent codes.

Since the clustered-sampling used with the previous show examples could result in train and test sets where certain ICPC-codes only occur in one of them, another way of sampling has been

ICPC	F1-score	Precision	Recall	Support
A97	0.20	0.11	0.94	12022
K86	0.92	0.89	0.97	8520
S87	0.86	0.78	0.93	8346
R96	0.91	0.92	0.90	6655
R74	0.89	0.89	0.89	6377
R05	0.95	0.95	0.95	6006
A04	0.90	0.92	0.88	4568
T90	0.94	0.99	0.89	4336
U71	0.95	0.96	0.95	4135
T93	0.91	0.98	0.85	4079
L99	0.69	0.80	0.61	4016
L17	0.72	0.69	0.76	3744
S74	0.87	0.92	0.82	3743
L04	0.78	0.92	0.82	3743
S99	0.76	0.84	0.69	3465

Table 7: Classification report of the fifteen most occurring codes using 80:20 train:test split with stratified ICPC-code sampling.

tested. In the table from figure 6 training and testing the model using stratified sampling is shown compared to the clustered-sampling. Both experiments had a train:test ratio of 80:20.

Sampling	F1-score	Precision	Recall	Support
Clustered by location	0.659	0.746	0.647	385,509
Stratisfied by codes	0.661	0.745	0.650	385,508

Table 6: Weighted average metrics with clustered and stratified sampling.

Although the precision and recall differ slightly between the two sampling techniques, these differences are very small These results do no seem to indicate a significant difference in performance.

5.2.2 Classification Quality – Most frequent codes

Table 7 contains a section of the classification report for the 15 most occurring codes in the test-set with a 80:20 train:test-split bases on stratified ICPC-code sampling.
Most scores seen in this table are very high, which is likely due to the classifier having a lot of descriptions for these higly frequent codes to train on. A interesting observation from this table was code *A97*. It has one of the highest recalls in this list, but the lowest precision. The code *A97* has the definition "Geen ziekte" or translated "No sickness", which is a very broad definition. A code with a broad meaning works well with the low precision, since that means it is assigned in a lot of cases were it was not that code. A code like *R96* which means "Astma" or translated "Asthma" has a high precision, this could be because the words in indicating this code are not often used in descriptions of other codes. This trend will be discussed further in section 5.3.1, where the words are shown that indicate the ICPC-codes *A97* and *R96* among four other highly frequent codes.

5.2.3 Confusion Matrix

As with the classification report, after the model is ran on a train and a test set it can generate a confusion matrix. This matrix is a two dimensional array of the shape (N,N), where N = the amount of unique ICPC-codes in the test set. In figure 8 a confusion matrix is shown for the twenty most occurring codes in a test set. The train and the test set for this model were split into 80% and 20% respectively. The codes shown on the y-axis are the target-values, the codes shown on the x-axis are the values predicted as the classification.

	A04	A97	F91	H81	K86	L04	L17	L81	L98	L99	R05	R74	R96	S74	S87	S99	T90	T91	T93	U71
A04	4,043	326	0	0	1	3	1	0	0	0	2	7	0	0	0	0	1	7	0	0
A97	5	11,418	2	0	14	1	4	2	1	0	1	1	2	1	0	0	3	9	5	9
F91	1	561	2,448	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
H81	0	68	0	3,190	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K86	0	121	1	0	8,258	2	0	0	0	0	2	1	0	0	0	0	2	0	6	0
L04	6	519	0	0	2	2,744	0	20	0	22	5	3	0	0	0	0	0	0	0	0
L17	1	515	0	0	0	0	2,874	93	19	36	0	0	0	8	4	0	0	0	0	0
L81	0	362	0	0	0	65	198	1,839	0	17	1	0	0	1	0	0	0	0	0	0
L98	0	664	0	0	0	0	105	2	2,375	16	0	0	0	0	0	0	0	1	0	0
L99	1	1056	0	0	0	7	64	48	3	2,395	0	1	0	0	0	0	0	0	0	0
R05	4	108	0	0	0	5	0	0	0	0	5,699	64	36	0	0	0	0	0	0	0
R74	2	328	0	2	0	1	0	0	0	0	59	5,654	39	1	0	0	0	0	0	4
R96	0	379	0	0	1	2	0	0	0	0	40	87	5,954	0	2	0	0	0	0	0
S74	0	392	0	0	0	1	22	0	0	0	0	0	0	3,168	61	2	0	0	0	0
S87	0	233	0	0	0	0	1	0	0	0	0	2	1	29	7,819	21	0	0	0	0
S99	0	678	0	0	0	2	7	0	0	0	0	0	0	3	9	2,308	0	0	0	0
T90	1	412	0	0	9	0	4	0	0	0	0	0	0	0	0	0	3,866	0	1	2
T91	6	296	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,868	0	0
T93	0	547	0	0	8	0	0	0	0	0	1	0	0	0	0	0	1	0	3,472	0
U71	0	115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3,906

Table 8: Confusion matrix of predictions for the twenty most occurring codes.

The first exceptional code here is *A97*. As discussed when looking at table 7, this code was assigned in a lot of predictions where it was not the target-value. This is visualized in the confusion matrix from table 8, with the whole column of *A97* having relatively high values. Another observation would be that codes with the same chapter-letter also seem to get confused in a significant way. A reason for this would be the fact that chapters define certain medical information, such as a location in the body. This could result in overlapping words between codes with similar chapters. One outstanding code in these situations however is *L17* meaning "Voet/teen symptomen/klachten" or translated "Feet/toe symptoms/complaints". One reason for its prominence in the confusing matrix could be because the meaning contains the words *symptoms/complaints*. These are very general words, which could more easily cause confusion to the model during classification. Another reason can be found by looking at the definition of *L81*, the code that *L17* is most often wrongly assigned to in the confusion matrix. *L81* has the definition of: "Ander letsel bewegingsapparaat" or in english "Other injury musculoskeletal system". Part of the injuries that fall under this category might be located at feet/toes, which could lead to confusion while classifying descriptions describing such injuries.

5.3 Classifier prediction analysis

After charting some positive and negative properties of the performance of a trained model, it was important to better understand classification decisions made by the model. In this section this aspect is explored in two parts: word-weights of classes and discrepancies between predictions and

targets. For these experiments the model has been trained on 80% of the data set. In 5.3.2 has the model has also been tested on the remaining 20%.

5.3.1 Class word-weights

In order to decide which ICPC-code to predict, the model could assign and update weights to certain words. In this research the model was allowed to assign weights to 1000 unique words it encountered. Each of these words were assigned a weight for each unique ICPC-code in the training set, indicating if a specific word was (un)likely to fit with a specific ICPC-code. In table 9 the six most occurring ICPC-codes are shown with a maximum of ten words that are most likely to indicate that code.

A97 - Geen ziekte

Word	Weight
geen	33
bvo	25
preventief	24
nvzb	23
niet	22
dossier	19
vaccinatie	19
contact	19
preventief	17
archief	16

K86 - Essentiële hypertensie zonder orgaanbeschadiging

Word	Weight
hypertensie	272
bloeddruk	23
hypertens	17
ht	15
hoge	12
beschad	10
org	9
verhoogde	8
rr	7
cvrp	5

S87 - Constitutioneel eczeem

Word	Weight
eczeem	279
constitutioneel	3
atopisch	2
dermatitis	1
kat	1
laesie	1
last	1
huisstofmijt	1
heeft	1

R96 - Astma

Word	Weight
astma	191
hyperreactiviteit	103
bhr	51
asthma	48
astmatische	30
bronchitis	7
kat	6
reactie	4
luchtwegen	4
status	4

R74 - Acute infectie bovenste luchtwegen

Word	Weight
blwi	181
verkouden	97
lwi	93
verkoudheid	80
luchtweginfectie	62
pharyngitis	57
bovenste	55
infectie	46
keelontsteking	39
infect	33

R05 - Hoesten

Word	Weight
hoesten	228
hoest	50
kriebelhoest	40
st	1
thorax	1
20	1
ziek	1
een	1

Table 9: The six most occuring ICPC-codes with words that imply them as a prediction.

First, the list of important features for the code *A97* in these tables stands out. The other five codes all contained at least one word with a weight above 150. This means there was at least one word for these codes that indicates that specific code very well. This was not the case for *A97*, it only has words with lower weights. In addition to that, most of its words are quite general words

like "geen" ('none'). Another observation is that the code *S87* only has one word with a significant positive weight. Although the word with the second highest weight "consititutioneel" occurs in the meaning of *S87*, it does not strongly indicate this code with a weight of 3. This could indicate that this word is not often used in descriptions from episodes classified by this code. Lastly, these words show how within the descriptions certain abbreviations are common practice. Although maybe not immediately clear to outsiders, these abbreviations such as "bhr" (meaning "bronchiale hyperreactiviteit" or "bronchial hyperreactivity") are better known within medical circles.

5.3.2 Discrepancies

Due to the fact that the model we used was a probabilistic classifier, it was possible to check how certain the model was with specific predictions. In the table 10 there are listed twenty predictions that are different than their target values. These twenty predictions were randomly sampled from a list of predictions of which the model was entirely sure about its classification. The predictions are supplemented with the additional information to give context to the incorrect classification. The table contains the original ICPC with its definition (labelled "Target") and the predicted ICPC with its definition (labelled "Predicted"). In addition the original description is also provided (labelled "classifier input"').

ICPC		**Description**		
Predicted	*Target*	*Predicted*	*Target*	*Classifier input*
L15	L91	Knie symptomen/klachten	Andere artrose/verwante aandoening(en)	Artrose binnenmeniscus linker knie
P15	P16	Chronisch alcoholmisbruik*	Acuut alcohol misbruik/intoxicatie	alcohol abuses
L16	L77	Enkel symptomen/klachten	Verstuiking/distorsie enkel	inversie trauma enkel
S88	S06	Contact eczeem/ander eczeem*	Lokale roodheid/erytheem huid	Dermatitis/eczeem
X72	X15	Candidiasis urogenitale vrouw	Andere symptomen/klachten vagina	wrsl Candida vulvovaginitis
A12	S88	Allergie/allergische reactie*	Contact eczeem/ander eczeem*	zonne allergie
S03	Y76	Wratten	Condylomata acuminata man	genitale wrat
U02	Y06	Frequente mictie/aandrang	Symptomen/klachten prostaat	Mictie problemen
L09	L92	Arm symptomen/klachten	Schoudersyndroom/PHS	pijn li arm
L12	S04	Hand/vinger symptomen/klachten	Lokale zwelling/papel/knobbel huid/subcutis	lokale zwelling hand
F13	F15	Afwijkend gevoel aan oog	Afwijkend aspect oog	Oog
L03	L02	Lage-rugpijn zonder uitstraling [ex. L86]	Rug symptomen/klachten	Rugpijn
L11	L72	Pols symptomen/klachten	Fractuur radius/ulna	st na pols # li
N93	K76	Carpale tunnelsyndroom	Andere/chronische ischemische hartziekte*	druk op borst + verhoogd CT calcium: sec preventie
U71	U70	Cystitis/urineweginfectie*	Acute pyelonephritis/pyelitis	uwi
U02	P12	Frequente mictie/aandrang	Enuresis [ex. U04]	in broek plassen
X15	X84	Andere symptomen/klachten vagina	Vaginitis/vulvitis nao*	vaginale infect
R81	R95	Pneumoni	Emfyseem/COPD	pneumonie bij exac COPD
K99	L14	Andere ziekte(n) hartvaatstelsel*	Been/dijbeen symptomen/klachten	atypische klachten been: veneuze insufficientie?
L99	H77	Andere ziekte(n) bewegingsapparaat*	Perforatie trommelvlies [ex. H71]	tv perforatie

Table 10: Twenty predictions were the model was confident and incorrect. The column "classifier input" is the episode descritpion that was used by the model to make the classification prediction.

Analysing the mistakes made by the classifier in these cases, it should be concluded that the model is flawed when descriptions are not clear enough. The following types of unclear descriptions seem to occur in these above mentioned examples, followed by examples from table 10:

1. The description is ambiguous, it could imply more than one classification. For example, "alcohol abuses" could indicate either *P15* ("Chronisch alcoholmisbruik" or "Chronic alcohol abuse") and *P16* ("Acuut alcohol misbruik/intoxicatie" or " Acute alcohol abuse/intoxication").

2. The description does not contain enough information, either being too brief or not containing medical information. For example, the description "Oog" only contains the word "eye", without specifying a certain condition, treatment, or symptom.

3. One or multiple words put the model on the wrong track. For example, the word "CT" in "druk op borst + verhoogd CT calcium: sec preventie" led the model to the code *N93* ("Carpale tunnelsyndroom" or "carpal tunnel syndrome"), since "CT" could be an abbreviation for its definition.

5.4 EHR software comparisons

Although the classification model used within these methods only took an episodes description as a feature, the episodes did contain other additional information. One bit of information listed was the type of EHR software used by the general practitioner, the type of system used to enter the episode into a digital network. Since our data contained three different types of software, these were the ones analyzed. These three software are:

- Medicom
- Microhis
- Promedico-VDF

One important factor to keep in mind with the methods in this section, is the ratio of the different software types within the data. As shown in table 3, most episodes originate from the Medicom EHR software. This will also be taken into account when analyzing the results of the following experiments.

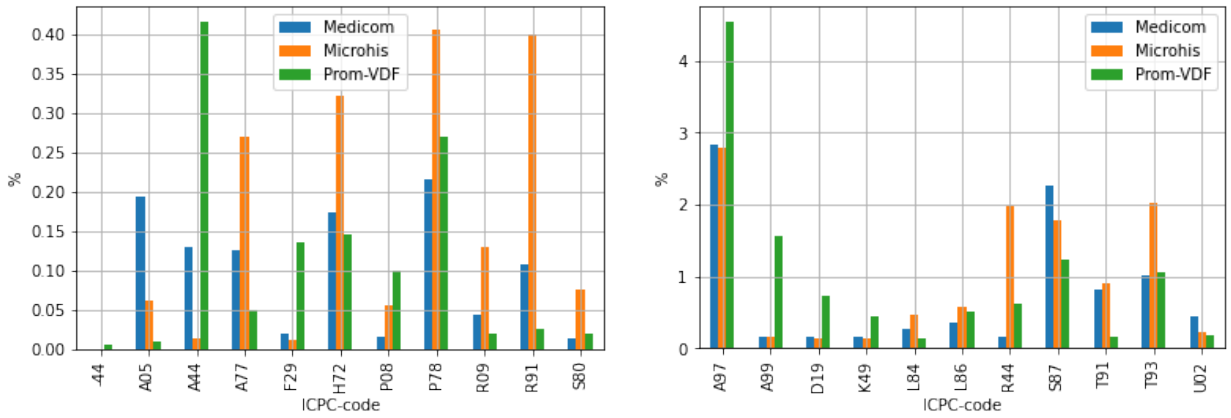
5.4.1 Code Support

While exploring the data, it was discovered that the different EHR software did not all contain the same amount of unique ICPC-codes. We performed a chi-squared test to investigate the ratio of ICPC-codes in the different software. In figure 11 a part of the chi-squared table is shown for the 25 most occurring codes. The values shown in the EHR software columns represent the difference (either positive or negative) between the expected and actual support of a code.

ICPC	Medicom	Microhis	Prom-VDF	Support
A97	1.8	0.8	198.6	51,157
K86	1.1	9.0	21.1	40,331
S87	5.7	55.2	88.9	40,009
R96	0.1	10.4	6.5	30,769
R74	1.9	65.6	1.2	29,833
R05	2.8	51.0	11.1	28,450
A04	3.2	52.5	18.1	21,301
T90	0.3	6.0	0.5	20,745
U71	2.8	42.6	19.0	19,480
L99	1.8	15.3	33.4	19,152
T93	18.6	537.4	0.1	18,756
L04	3.6	61.6	17.1	17,670
L17	1.0	23.4	1.3	17,486
S74	0.9	12.8	5.8	16,601
S99	0.6	0.2	66.1	16,549
L81	0.6	2.0	22.7	16,458
H81	0.6	24.9	2.0	15,901
L98	0.2	0.9	30.8	15,743
F91	2.2	9.4	70.2	15,673
L15	0.0	0.8	1.0	14,896
T91	0.4	6.2	107.8	14,554
L08	0.0	5.3	10.6	14,176
A80	0.5	2.8	11.6	14,048
S03	0.0	3.2	1.8	13,675
L02	2.5	56.3	2.9	13,329

Table 11: Part of the chi-squared table for the 25 most occurring codes.

Since there are 705 different ICPC-codes the chi-squared table is not the optimal way to view outlier ratios. However, it can already be seen that *Microhis* and *Promedico-VDF* seem to have more outstanding ratio's with certain codes. This is partly, because *Medicom* is the largest set of ICPC-codes. If there were 100 more of a certain code in the largest set, this would be less visible than if there were 100 more in one of the other sets. In the graphs from figure 2 there are a total of 22 codes shown. Each of the colored bars shows the percentage of episodes with that respective EHR software, that have that specific ICPC-code. These codes have been selected by selecting the ten codes with the highest value in the chi-squared table per software. This resulted in 22 codes instead of 30, since some of them overlap software types.



(a) ICPC-codes with lower proportions

(b) ICPC-codes with higher proportions

Figure 2: Percentage of episodes with certain ICPC-codes

A few of the ICPC-codes to highlight are the following:

-44: Meaning "Immunisatie/preventieve medicatie" or "Immunisation/preventive medication". This code seems to only be used within *Promedico-VDF*.

R44: Meaning "Influenzavaccinatie" or "Influenza vaccination". This code gets used more within *Microhis*. This could be because vaccinations for influenza are given more often in locations using this software.

A99: Meaning "Andere gegeneraliseerde/niet gespecificeerde ziekte(n)" or "Other generalised/not specified illness(es)". This code seems to be used a lot more within *Promedico-VDF*. Since within that EHR software the code *A97* is also used often, there could be a correlation between more general codes and that software.

5.4.2 Classification Quality

To find if EHR software types could have an influence on the performance of the model, classification reports were generated in a different way. For his method the data was first split on software type, which where then all split in 80% training data and 20% testing data. For the training sets, all three software were combined to have one standard training set. The test sets were kept separated and were tested independently of each other. This resulted in three different classification reports, one for each type of software. The weighted average of metrics from these reports can be found in table 12.

	F1-score	**Precision**	**Recall**	**Accuracy**
Medicom	0.658	0.736	0.649	0.649
Microhis	0.721	0.786	0.716	0.716
Promedico-VDF	0.744	0.793	0.754	0.754

Table 12: Weighted average results with 80:20 train/test-split per EHR software type

As seen in the table there is a difference in performance between the different software predictions. Since *Medicom* has more unique ICPC-codes (table 3). Therefore *Medicom* could have more codes in its test set that are less frequent with low prediction metrics. If we compare these classification results to those from testing different sampling techniques in table 6, we can see that the result for *Medicom* are very similar to those. This could be explained by the fact that *Medicom* makes up most of the episodes from the data as seen in table 3.

6 Discussion

6.1 Methods

It is important to acknowledge that the methods used in this research were not the only ways to answer the questions. One of the important parts of most experiments was the classification model. This model consisted of a Support Vector Machine (SVM) with a Count Vectorizer. These SVM and Vectorizer used certain settings, for example the dictionary size of the Count Vectorizer was set to 1,000. Certain settings like this size were not tested extensively with multiple variants, which means there is still room for optimization in these settings as a larger dictionary size could possibly lead to better classification results. Aside from the settings, the type of classification model and the Count Vectorizer could be swapped out as well. As an example, the Vectorizer could be replaced by td-idf weights, which could exclude non-descriptive words that are used indecently of ICPC-classifications in episode describtions.

6.2 Results

What can be learned from the results in this research is that a large part of episodes contain descriptions that are informative about the ICPC-code assigned to it, but there are episodes where this is not the case. The limitations of this research are that we are not able to assign specific numbers to these statistics, because there are certain factors that were not taken into account. We can say that it is likely that almost 65% of records in the data have informative descriptions, since the classifier is able to predict a correct ICPC-code using the information within them. However, we can not predict such a number for non-informative descriptions, since the model might not understand information from certain descriptions. This could be either due to the lack of training data for certain ICPC-codes, or the amounts of words chosen for the model to understand.

6.3 Future Work

The original focus of this research was on exploring the data, then preparing a classification model and finding discrepancies between the models ICPC-code predictions for an episode and the actual codes. The idea was to use these discrepancies to find better suited ICPC-codes for episodes, but experiments proved unsuccessful in finding episodes where it was certain the model assigned a better ICPC-code. The results from table 10 showed that the model created was not suited for finding better classifications. These were all predictions where the model was about a different ICPC-code, but none of the predictions could be proven to be better suited classifications. Thus without a metric to indicate potential episodes that could be classified better, it was decided to shift to analyzing the prediction process and results. If someone would want to research discrepancies to achieve a model that can ICPC-codes that better fit descriptions, they would need to find a metric that indicates this property.

7 Conclusion

How does a text-based classification model perform when classifying EHR episodes? The Support Vector Machine used in this research has an average accuracy of 0.65 for all codes from the data using the optimal settings from the ones tested. For most codes a higher frequency in the data resulted in it being classified more often. Increasing the size of the training set from the minimal size tested makes a minor impact on classification results.

How does classification quality differentiate between individual ICPC-codes in EHR episodes? Higher frequency in the data in general indicates better precision and recall, due to more training data being available. An exception being the code *A97*, which has the lowest classification performance of the top 15-most frequent codes (see table 7). Codes with similar chapters are confused in classification more often than codes with different chapters (see figure 8), a strong example being *L17*. The classification quality of an ICPC-code is also higher if in its description specific words are used that are not used for other codes. These words will often be directly related to the meaning of the code.

To what extent does EHR software differ in class frequency in episodes and their classifiability? Certain codes occurred in higher proportions within the three different software types compared to each other. In this research the least frequent EHR software *Promedico-VDF* was able to be classified correctly the best with an accuracy of 0.75. This was followed by the software *Microhis* with an accuracy of 0.72. Lastly there was *Medicom* with an accuracy of 0.65.

References

[BS13] Roar Bjurstrøm and Jaspreet Singh. De-identification of norwegian health record notes: An experimental approach. Master's thesis, Institutt for datateknikk og informasjonsvitenskap, 2013.

[Gen19] Nederlands Huisartsen Genootschap. Patiëntendossier, 2019.

[Naa] Netwerk Naamkunde. Nederlandse voornamen top 10.000. Available at http://www.naamkunde.net.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Sci20a] Scikit. Countvectorizer, 2020. Available at https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.

[Sci20b] Scikit. Sgdclassifier, 2020. Available at https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.

[SCL+17] Elyne Scheurwegs, Boris Cule, Kim Luyckx, Léon Luyten, and Walter Daelemans. Selecting relevant features from the electronic health record for clinical code prediction. *Journal of Biomedical Informatics*, 74:92–103, 2017.

[SRS+16] Annet Sollie, Jessika Roskam, Rolf H Sijmons, Mattijs E Numans, and Charles W Helsper. Do gps know their patients with cancer? assessing the quality of cancer registration in dutch primary care: a cross-sectional validation study. *BMJ Open*, 6(9), 2016.

[vdBKTV+17] S. van der Bij, N. Khan, P. Ten Veen, D. H. de Bakker, and R. A. Verheij. Improving the quality of ehr recording in primary care: a data quality feedback tool. *Journal of the American Medical Informatics Association*, 24,1, 2017. Available at https://www.ncbi.nlm.nih.gov.

[Won20] Wonca. International classification (wicc), 2020. Available at https://www.globalfamilydoctor.com.