Universiteit Leiden
The Netherlands

Computer Science

Utilizing unstructured information from
Electronic Medical Records for the prediction of stroke

Name: A. Louwe
Date: April 29, 2020

Supervisors:
Dr. S. Verberne (LIACS)
Dr. M. van Leeuwen (LIACS)
H.J.A. van Os (LUMC)

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Stroke is one of the leading causes of disability and death in The Netherlands. Currently general practitioners estimate stroke risk using a simplistic model consisting of only five risk factors, namely age, gender, blood pressure, cholesterol level and smoking status. A more accurate prediction model is needed to recognize patients at risk of stroke, which is an essential step towards more effective stroke prevention.

Through the development of a prediction model based on information extracted from free text in electronical medical records (EMRs), this project aims to achieve earlier and more accurate recognition of persons at risk. Previous studies in the field of disease prediction have mainly focused on using structured data from EMRs, due to the ease of information extraction from this type of data. However, we expect that the free text could contain additional information valuable for stroke prediction.

We build a text mining pipeline specifically for Dutch medical texts, which includes two techniques for the extraction of predictors from text namely bag-of-words and topic modeling, and we build a logistic regression model for stroke prediction. In each step of the pipeline we prefer methods that are easily explainable to medical experts and we provide visualisations when appropriate, which are also used for the identification of possible novel risk factors for stroke. Furthermore, we study the confounding effect of age on the prediction of stroke risk.

Contents

Chapter 1

Introduction

Stroke is one of the leading causes of disability and death in the Netherlands, with an incidence of around 3 per 1000 person-years. The most frequent type of stroke, ischemic stroke, is caused by a narrowed or blocked blood vessel, which deprives an area in the brain from oxygen. The other type of stroke, hemorrhagic stroke, is caused by a ruptured or broken blood vessel that bleeds into the brain, thereby compressing the surrounding brain tissue [2]. Both types of stroke can result in irreversible brain damage or even death. In men, quality of live after experiencing a stroke has improved over the years. In women however, this is worse than in men and stroke results more often in institutionalization. Only a small proportion of all patients can be treated. Therefore, preventing stroke is key, and new strategies for early recognition of women at risk of stroke are urgently needed.

In the Netherlands, general practitioners (GPs) currently stratify their patients into risk groups using a simplistic algorithm that contains only five traditional risk factors (i.e. age, gender, hypertension, cholesterol and smoking). However, evidence on the importance of female-specific risk factors such as migraine [48], coagulation disturbances, endothelial dysfunction and reproductive disorders is mounting. These factors may interact with each other and with traditional risk factors [49]. Furthermore, psychosocial distress has been associated with stroke risk in older adults [26]. Currently, none of these other factors are taken into account for stroke prevention, possibly withholding preventive measures from thousands of persons at risk.

Machine learning methods and data mining techniques have grown to be increasingly popular within the medical domain, and have been applied to a broad range of tasks, including case detection [18] and disease prediction [23]. Currently, such data-driven prediction models do not yet exist for stroke. Whereas infection diseases often have clear symptoms and are relatively easy to predict, acute events such as stroke are much harder to predict. Furthermore, stroke risk increases with age, the incidence doubling each decade after the age of 55 years.

In The Netherlands 84% of all strokes in 2017 occurred above the age of 65 [1]. As a result, possibly predictive stroke symptoms are difficult to distinguish from the symptoms of other age-related diseases.

Electronic medical records (EMRs) are an important data source for machine learning methods. Medical institutes increasingly use EMRs to record a patient's condition, including diagnostic information, procedures performed, and treatment results. The majority of this information is structured; measurements and laboratory results are stored using numerical values and diagnoses and medication prescriptions are often stored using standardized coding systems. This information allows for large-scale analysis and does not require extensive preprocessing before it can be used as input for machine learning models.

A minority of the fields in EMRs is unstructured; typically these are the manually written consultation notes and diagnosis descriptions. This is a rich source of information. Free text accommodates the reporting of relevant information that is not suited for coding, including the expression of feelings, uncertainty, the addition of supporting evidence and recording strange collections of symptoms. Furthermore, these notes can contain non-medical information that are relevant indicators for the development of medical problems including (causes of) psychosocial distress or factors related to socioeconomic status.

Information extraction from unstructured sources requires advanced preprocessing pipelines, which are language and domain specific. Apart from the huge amount of medical terms, text from the medical domain is also extremely concise and contains a relatively large amount of misspellings [30]. Furthermore, notes in EMRs are much more prone to contain personal information about the patient compared to the coded parts of the EMR, while current de-identification and anonymization techniques still lack the required precision expected by health care institutes. Consequently, most studies only use the structured information from EMRs at the cost of information loss, which in turn can lead to more biased prediction models.

In this research, we develop a pipeline for the extraction of information from Dutch medical text and investigate the confounding effect of age in stroke prediction. We use EMR data obtained from general practitioners in The Netherlands, which we describe in more detail in Chapter 3.

1.1 Research questions

We aim to improve the prediction of stroke by using information extracted from free text in EMRs. More specifically, we aim to answer the following questions:
1) How can we develop an explainable text mining pipeline for the prediction of stroke?
2) How can we discern the confounding effect of age from this information?

1.2 Main contributions and challenges

In this work we contribute towards more accurate prediction of stroke by the development of data-driven prediction model based on unstructured data. As part of this contribution, we also develop a text mining pipeline for Dutch medical text, which can also be applied in other disease prediction or case detection studies. Furthermore we investigate methods to discern the confounding effect of age in disease prediction results.

Challenges include working with unedited natural language, which typically contains a variety of noise. Furthermore, due to intended use in diagnostication, we are challenged to develop a model that is easily interpretable by medical experts, while we also have to ensure high precision.

1.3 Thesis structure

In Chapter 2 *Related Work*, we introduce various existing methodologies and tools for extracting information from medical texts. In Chapter 3 *Datasets*, we describe the EMR dataset we analyse in this research, which we obtained from general practitioners in The Netherlands. In Chapter 4 *Analysis & Results*, we dive into the details of each step in the analysis; *Pre-processing*, *Data Selection & Analysis*, *Feature selection* and *Prediction models* respectively. In Chapter 5 *Discussion* we describe the main results, we discuss the strengths and limitations, and point to directions for future work. Finally, in Chapter 6 *Conclusion* we summarize our research and draw conclusions.

Chapter 2

Related Work

A number of studies have been performed to explore how information can be extracted from medical texts and how valuable this information can be in predictive modeling. In our previous research we explored various preprocessing methods and found text cleaning and spelling correction are essential steps for reducing noise in clinical texts [30].

2.1 Tools for information extraction

Multiple tools have been developed that allow for the identification of medical terms from text by coupling it to a medical ontology, typically the Unified Medical Language System (UMLS) ontology [7]. Good examples of such tools are HITEx [52], cTAKES [42] and Sophia [15]. These algorithms perform preprocessing followed by a number of complex matching algorithms, to match the data to the right concepts in the ontology. Although a subset of the UMLS has been translated to Dutch [34], there are no matching tools available for Dutch medical concepts.

It is also important to take into account the contextual properties of the identified medical concepts. Examples of such tools are NegEx [9] and ConText [25], which can be used for the identification of negations, when the symptom occurred in (temporality) and who experienced the symptom (experiencer). The trigger words and rules applied in ConText are also translated to Dutch in the tool ContextD [3] for the identification of similar contextual properties in Dutch notes.

2.2 Disease prediction

Studies that explore the benefit of distilling information from medical text for the purpose of predictive modeling are also limited in number. Kop et al. [28]

describe the development of a pipeline for the prediction of colorectal cancer using structured EMRs and Hoogendoorn et al. [27] show that adding predictors derived from the texts improved the accuracy of the prediction model. Similar results where found in [13], in the area of detecting acute respiratory infections.

There are also studies that only use text. In [5] a study is performed in the area of life-expectancy prediction showing that extracting information from doctor's notes is a promising technique for this task. In [40] risk of suicide among veterans is estimated, again using only information from clinical notes. Some recent studies also investigate the extraction of information from patient forum data [51, 14]. Disease prediction models based on this type of narrative data are still under development.

Although there are no machine learning studies that aim at predicting stroke, some aim at predicting cardiovascular diseases (CVD) which has some characteristics similar to stroke. In [50] four machine learning algorithms (random forest, logistic regression, gradient boosting machines, neural networks) were compared, showing that the machine learning models have an accuracy that is significantly higher that the currently established tools for CVD risks. Here, only structured EMR data was used.

Interestingly, there is no single machine learning algorithm that performs best on clinical data. Ford et al. [18] compare 67 studies that use clinical data –all including clinical notes– for the purpose of detecting a specific (type of) disease. They found no significant differences among the accuracy of different types of algorithms. Although these case-detection studies slightly differ from disease prediction studies in their aim, the methodology of these studies is often similar to disease prediction studies.

Except for choosing the optimal machine learning model, it is also important which features/variables are included in the model. This is highly influenced by characteristics of the dataset, the amount of text preprocessing applied, the maximum size of the feature set and the feature selection technique used. This complicates the comparison of different studies.

Another important consideration in choosing a machine learning algorithm for medical applications is the interpretability of the model. Interpretable models, such as logistic regression models, are often preferred over black-box models (e.g. neural networks) [47].

Chapter 3

Datasets

We analyse an anonymized primary care dataset named the Extramural Leiden Academic Network (ELAN) primary care database. It contains GP data of 105619 individuals from GPs centered around Leiden. The number of stroke cases in this dataset was 1415 (1.3%). The dataset contains demographic patient data (e.g. age, gender), ICPC codes for diagnoses (often with a short description), medication prescriptions, laboratory results and referrals.

Typically an EMR dataset also contains consultation notes written by the general practitioner, but unfortunately these were removed from ELAN to protect patient privacy. Instead we use the 'diagnosis descriptions' as our source of unstructured information. Diagnosis descriptions are provided for various purposes, for example to add information about the certainty of the given diagnosis, to describe additional symptoms or to indicate the location of the problem more precisely.

Using the diagnosis descriptions also has some disadvantages. Many GPs use a system that partly automates the input of diagnosis descriptions. Consequently, the diagnosis descriptions are often (almost) equal to the official 'name' of the ICPC code to which they have been assigned. A quick examination of the descriptions in the ELAN dataset showed that 20% of these descriptions are exactly equal to the ICPC code. Furthermore, diagnosis descriptions have a strict length limit of 40 characters.

The ELAN dataset is a subset of the STIZON dataset [44], which is a nationwide dataset containing data of approximately 3 million patients and also contains the consultation notes. The pipeline we develop in this work will be applied to the STIZON dataset in future work to make a more accurate stroke prediction model and to find more reliable risk factors for stroke in women.

3.1 Patient privacy

Due to legal and institutional concerns about patient privacy it is difficult to gain access to medical data for text mining and this has been a major obstacle to progress in this field [45]. Clinical notes can potentially contain highly sensitive information and is therefore even harder to obtain than regular (structured) health data.

We also faced many obstacles in the process of obtaining the STIZON dataset, which was initially the dataset we planned to use for this research. As an alternative approach, we obtained the NEO dataset containing the complete EMRs from 6671 individuals, who gave their informed consent by participating in an obesity study. Unfortunately, the number of stroke cases in this small dataset was too small for stroke prediction. Therefore, we used the ELAN dataset that we described above.

We are also unable to share the medical data used in this research.

Chapter 4

Analysis and results

Our analysis consists of four main phases. In this first phase, *Preprocessing*, we convert the raw data to a format that the computer can more easily work with and includes text cleaning, tokenization, spelling correction and key-phrase detection. The second phase, *Data Selection & Analysis*, is where we select relevant data for use in the prediction phase and we split the patient group in target and non-target patients. Furthermore, to obtain insight into the basic distributional properties of the data, we analyse token frequencies in both groups and build a word embedding model. In the third phase, *Feature Selection*, we select features that will represent the data in the prediction models. In this phase we apply two techniques for the extraction of predictors from the data, namely bag-of-words and topic modeling. Finally, in the fourth phase, *Prediction models*, we use logistic regression to build a prediction model and find potential risk factors for stroke.

4.1 Preprocessing

Natural language typically requires a large amount of preprocessing and this is even more the case for medical notes as we discovered in our previous research project [30]. The data contain many spelling errors, for which we developed a spelling correction algorithm (See Section 4.1.2).

Further analysis of the data shows that many words are incomplete (e.g. *probl* instead of *probleem*) and that many medical terms are multi-word phrases (e.g. *diabetes mellitus*), which could lose their meaning when split into separate tokens. We extend our existing preprocessing pipeline with techniques that deal with these issues.

Other preprocessing steps used in the medical domain are synonym detection and negation identification [3, 52]. A single disease or symptom often has many synonyms, ranging from complex medical jargon to terms used in day-to-day

language. Unfortunately, resources for Dutch medical synonyms are limited. Initial experiments with synonym detection using SNOMED-CT [34], the Dutch version of UMLS [7], led to retrieval of overly specific medical concepts given the limited information present in each diagnosis description. For example *'vitamine deficientie'* (vitamin deficiency) was matched to 'Neuropathy associated with hypervitaminosis B6 (disorder)' and *'hoofdpijn'* (headache) was matched to 'Medication overuse headache (finding)'. In some cases the matches were not only too specific, but also unrelated to the original text (e.g *'immunisatie preventieve medicatie'* (immunisation) was matched to 'Transfusion reaction due to leukagglutinins'). Further experiments using bi-word search queries, reduced recall to an insufficient amount of synonym coverage (e.g. concept 95655001 'Ophthalmic migraine' was not retrieved for search query *"migraine ophthalmic"*).

Negation words indicate that something does *not* apply to the patient. Ignoring or removing negation words results in an interpretation of the phrase that is opposite to the intended interpretation (e.g. 'no headache' could be interpreted as 'the patient has headache'). Negation detection algorithms often use pattern matching (e.g. NegEx) or part-of-speech (PoS) tagging (e.g. NegExpander) to obtain the (sentence) structure [24], which is needed to identify to which words in the sentence the negation applies.

Data analysis shows that negation words or phrases –as listed in Appendix A– occur in only 1.5% of the records and that the data typically does not contain any (sentence) structure. Consequently, we decided not to include negation detection in the preprocessing pipeline.

To summarize, the preprocessing pipeline consists of text cleaning, tokenization, spelling correction, word completion and key-phrase detection.

4.1.1 Cleaning and tokenization

We lowercase the diagnosis descriptions and replace all punctuation symbols –except hyphens– by white space. In Dutch hyphens are sometimes used to create compound words. Therefore, we keep all hyphens that appear within a word and remove only the hyphens occurring at word boundaries. Furthermore, we remove the stop words (commonly used words) listed in Appendix B [17]. We also remove words without any alphabetical characters (e.g. dates and telephone numbers). Finally, we remove words that contain fewer than three characters. We tokenize the strings of text into a list of words by splitting on white space.

The top 5 most frequent remaining words are *klachten, pijn, symptomen, eczeem* and *acute*. The frequency distribution is highly skewed (Fig. 4.1). At this point of preprocessing we have a total count of 4200400 tokens of which 94334 are unique tokens.

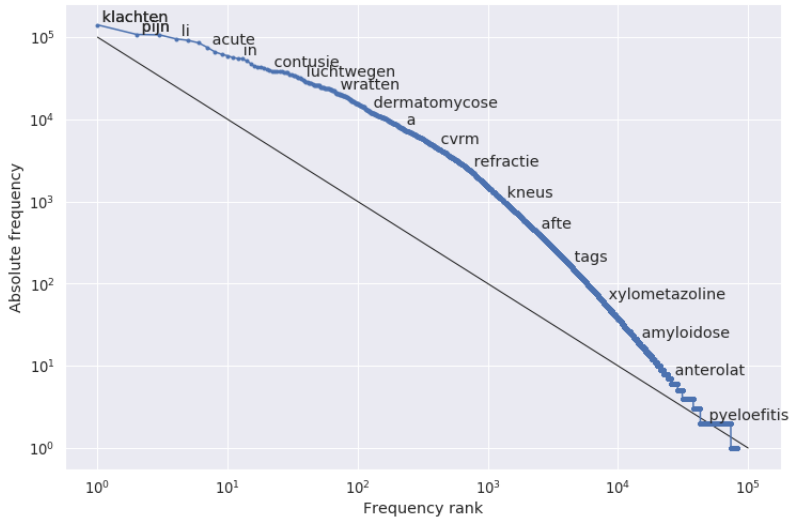

Figure 4.1: Zipfs plot showing the relation between the absolute frequency and frequency of the tokens. Both axes have a logarithmic scale.

4.1.2 Spelling correction

Texts in EMRs are generally written under high time pressure and often intended for personal use only, resulting in a larger amount of uncorrected misspellings or even intentional misspellings, since typing less characters is more efficient and even more beneficial when notes are limited to a certain maximum length.

Humans automatically relate misspelled words to correct words they already know or –when they know multiple correct words similar to the misspelled word– they understand it based on the context of the word. However, for computers each unique string of characters represents a unique token. Consequently a computer model that 'knows' the meaning of the word *migraine* might have no clue about the meaning of *migrain*.

In some cases a misspelling is also a correct word (e.g. mispelling *cold* as *old*), which we can not identify without using a context sensitive spelling correction algorithm [22]. Since these algorithms generally need more context than is available in the short EMR notes, we focus only on non-word misspellings.

Domain specific dictionary

A dictionary of Dutch words that includes a sufficient amount of medical terms is needed in order to use a dictionary-based spelling correction algorithm. We construct this dictionary by combining all unique words from the official ICPC

descriptions [35] with the Clinspell lexicon [17] and the CoNLL collection [16]. The Clinspell lexicon contains 474206 words from both general and medical sources and the CoNLL 2017 Shared Task Dutch is a large word embedding corpus containing 2.6 million words retrieved from Wikipedia and Common Crawl. The combined dictionary contains almost 2.8 million words.

Spelling correction algorithm

We identify non-word misspellings by comparing words from the dataset with words from the dictionary. When a word is an out-of-vocabulary (OOV) term (i.e. not found in the dictionary), it is either an actual misspelled word or a correct word that is not in the dictionary, which can occur when words are very domain-specific or when the dictionary is too small. By using a large dictionary and including additional medical terms we reduce the chance of these false positives. Before spelling correction 6% of the words in dataset were OOV words and the set of all *unique* words consisted for as much as 55.6% of OOV words.

For replacing the OOV terms by dictionary words, we use the Damerau- Levenshtein distance (DLD) [4], which is defined as the minimum number of insertions, deletions, replacements or transpositions of adjacent characters needed to change one word into another word. We set the maximum DLD distance to 1 character and the minimum word length to 6 characters to find words in the dictionary that are highly similar to the OOV terms. When the algorithm finds multiple correction candidates, we choose the candidate with the highest frequency in the dataset. Short words are more likely to have multiple correction candidates and are more difficult to correct. Interestingly, despite the strict spelling correction parameters many (severely) misspelled words obtained incorrect correction candidates. (e.g. *neurit* $\rightarrow$ *neuriet* instead of *neurit* $\rightarrow$ *neuritis*, *angstkl* $\rightarrow$ *angstel* instead of *angstkl* $\rightarrow$ *angstklachten*)

Improving precision

Precision of spelling correction in the medical domain is essential. Small incorrect word edits can drastically change meaning of not only the word itself (e.g. changing *oog-* (eye-) to *oor-* (ear-)), but also the meaning of the whole note and possibly even the behaviour of the prediction models when trained on this data. We reduce the amount of incorrect corrections by considering only correct words that were already present in the dataset as correction candidates instead of using all words from the dictionary as correction candidates. Manual evaluation of 250 randomly selected corrections showed that 93% of these were correct.

Algorithm efficiency

Computation of the DLD of a single word pair has a time complexity of $O(n*m)$ where n and m are the word lengths. A naive approach to find all correction

candidates for all OOV terms requires millions of DLD computations. While this is still possible within reasonable time for our dataset it is not scalable to larger datasets and/or larger dictionaries.

Given that the maximum DLD threshold is low, a more efficient approach is spelling correction using Symmetric Delete [19]. This algorithm deletes one or more characters (up to the maximum DLD) from both the dictionary words and the OOV terms. These words-with-deletions are stored in a hash table or similar data structure. After this preprocessing step, simple equality comparisons are sufficient to find correction candidates in the dictionary without any DLD computation while the results are identical.

The spelling correction algorithm was able to correct 20339 words, which reduced the percentage of OOV words in the dataset to 4%. Table 4.1 lists some misspellings and their corrections. Manual evaluation, by a medical expert and the author, of a random sample of 250 spelling corrections showed that only 2% of these were regarded as incorrect replacements. (e.g. *jindy→cindy, paralu→paraplu*)

OOV term	**Correction**
maagklacgten	maagklachten
schurtje	scheurtje
mammacarcinooml	mammacarcinoom
epydidymitis	epididymitis
oxazepma	oxazepam
ingegreoide	ingegroeide
apigastrio	epigastrio

Table 4.1: Examples of spelling corrections provided by the spelling correction algorithm

4.1.3 Word completion

Due to the strict limit imposed on the length of the diagnoses descriptions, many words are incomplete. The spelling correction procedure described in the previous section is not suited for the correction of incomplete words. Therefore, we look for each of the remaining OOV terms with a length of at least 3 characters, whether it occurs as prefix in the vocabulary. (e.g. for *transplantat* we look if any of the correct words in the dataset start with *transplantat-*). When we find multiple candidates we choose the candidate that is most frequent in the dataset. Figure 4.2 shows some examples of incomplete words in dataset and the complete word found by the algorithm. We found completions for 11124 words, which is 21.6% of all OOV terms.

word	candidates	completion
eetpro	[eetprobleem, eetproblemen, eetproblematiek]	eetprobleem
heupprobl	[heupprobleem, heupproblemen, heupproblematiek]	heupprobleem
psychisc	[psychische, psychisch]	psychische
marcouma	[marcoumar]	marcoumar
subcla	[subclavia]	subclavia
periostit	[periostitis]	periostitis
transplantat	[transplantatie]	transplantatie

Figure 4.2: Word completion examples

4.1.4 Key-phrase detection

In the first step of the pipeline we split all notes into single-word tokens. However, many medical terms are multi-word phrases (e.g. diabetes mellitus). When each word in a phrase is analysed as an individual concept, these phrases often lose their meaning. Phrases can be identified by computing the *phraseness*; the degree to which a given word sequence is considered to be a phrase [46]. We define the phraseness for phrase P with unigrams u and their relative frequencies in the dataset $RF(P)$ and $RF(u)$ as shown in Eq 4.1 (Pointwise Kullback-Leibler divergence).

$$Phraseness = RF(P) \log \left(\frac{RF(P)}{\prod_{u \in P} RF(u)} \right) \tag{4.1}$$

We computed the phraseness of $7.5M$ phrases consisting of two, three, four or five words. Figure 4.3 shows the phrases with the highest phraseness. Based on manual inspection of the top phrases, we selected the top 0.01% of phrases (94 phrases) with the highest phraseness as key-phrases and replaced the individual tokens of these phrases in the dataset with the key-phrase in which we replaced spaces with underscores.

The list of key-phrases in Fig 4.3 contains some key-phrases which are not multi-word phrases in common natural language, for example *immunisatie preventieve medicatie* (Preventive Immunisations/Medications), which is exactly the description of ICPC code `-44` after our preprocessing steps and is therefore a frequent 'phrase' in the cleaned diagnosis descriptions. This is a consequence of the semi-structured nature of the diagnosis descriptions and will occur less frequently in fully unstructured text.

term	count	n	phraseness
acute infectie bovenste luchtwegen	14366	4	0.028206
symptomen klachten	43170	2	0.019766
infectie bovenste luchtwegen	14480	3	0.019355
immunisatie preventieve medicatie	12435	3	0.017975
acute infectie bovenste	14385	3	0.017490
lage rugpijn zond uitstr	6537	4	0.014591
rugpijn zond uitstr	6539	3	0.009738

Figure 4.3: Multi-word phrases with highest *phraseness*

4.2 Data selection & Analysis

In this phase the divide the data into two groups; a target group (i.e. patients who have experienced one or more strokes) and a non-target group. Furthermore, we compare token frequencies among the groups and build a word embedding model, which we will use for data visualisation and word similarity computation, and we consulted a medical expert for interpretation of the output of these analyses.

4.2.1 Patient and time period selection

Earlier recognition of patients at risk of stroke increases the probability that preventive measures, such as medication and lifestyle changes, are effective. Therefore, a successful disease prediction model has to be able to predict the disease based on a limited set of records that were collected some time before the disease occurs. In this section we describe how we selected relevant data for training and testing the model.

Similar to survival analysis studies we define a study period (the roll-in period) and a follow-up period for each patient. [37] The roll-in period is a period with a fixed length from which we select data to use for the model and the follow-up period is a fixed length period directly after the roll-in period in which the event occurs at some moment if the patient is a target patient. For non-target patients we chose a random roll-in period.

Most data was collected between 2007 and 2017, which is a relatively short period and not long enough to predict a stroke 10 years before the actual stroke event. When a patient had a stroke in 2015 we would need to train the model with health information from **before** 2005, which is not available in the dataset. Therefore, we train our model to predict strokes 5 years or less before the first stroke (i.e. the follow-up period is 5 years). The roll-in period is 2 years. Sur-

vival analysis of the ELAN dataset shows that a 1-year roll-in period is already sufficient for accurate prediction [39].

The resulting dataset contains 48190 patients of which 517 are stroke patients. Each patient has on average 5 records, with a standard deviation of 3.85 records.

4.2.2 Word frequency analysis

Next, we compared target patients (stroke patients) with non-target patients by comparing token frequencies using the Kullback-Leibler divergence (Eq. 4.2), which is a measure of how one probability distribution is different from a second probability distribution. In Equation 4.2, $P(x)$ is the relative frequency of token x in the target-patient subset and $Q(x)$ is the relative frequency of token x in the non-target subset. Due to the asymmetry of this measure the tokens that are relatively more frequent in the target subset will obtain an higher score and thus we will obtain a list of potential risk factors for stroke using this measure.

$$D_{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{4.2}$$

A list of tokens with the highest KL divergence (Table 4.2) shows that a wide range of medical conditions and symptoms occur more frequently in the target group. Noteworthy, many of these terms appear to be age-related (e.g. presbyacusis: age-related hearing loss). Since stroke incidence is highest among elderly, the mean age of the target group is twice as high as the mean age of the non-target group. Consequentially, age-related symptoms and diseases are more frequent in this group.

To obtain some more insight into the confounding effect of age on the KL divergence we also plotted the KL divergence against the 'mean age' of each token in Fig. 4.4. The mean token age is computed by listing all occurrences of the token (in the full dataset) together with the patients' age at the time of diagnosis followed by calculating the mean of these ages. Using this method we obtained –for example– a mean age of 57 for *erysipelas* and a mean age of 47 for *neusbloeding* (nosebleed), which implies that *erysipelas* occurs on average more in elderly persons than nosebleeds.

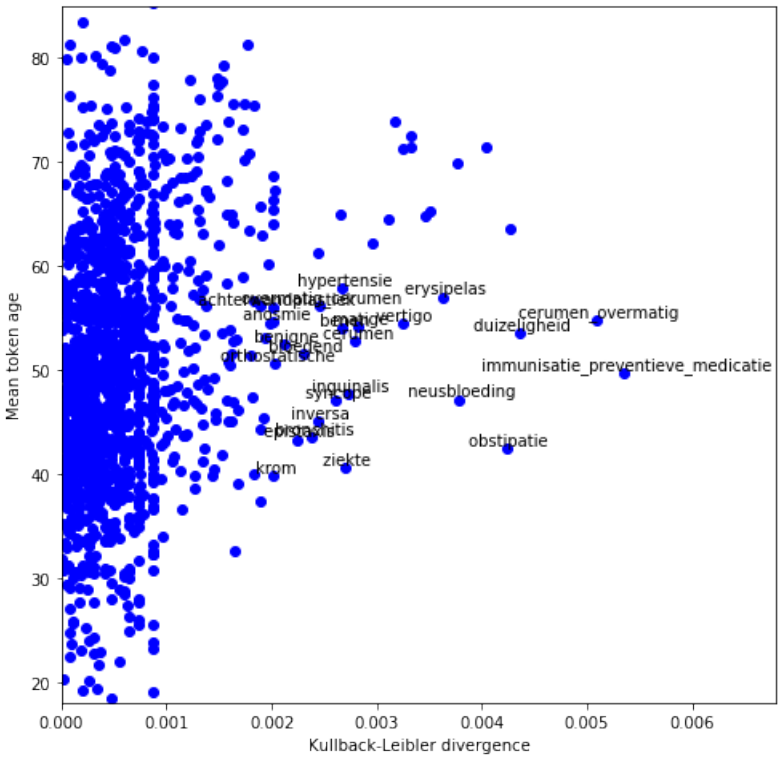

Figure 4.4: KL divergence and mean token age. We annotated all tokens with a mean token age smaller than 60 and a KL divergence bigger than 0.0020.

4.2.3 Word embedding

A word embedding model is another method for representing text data. In a word embedding model each word is represented as a vector in a vector space. These vectors are obtained using a neural network that is trained on word contexts in the input data. Similar words will have similar word vectors and therefore a word embedding model can be used to compute similarities between words. We use the Word2Vec (W2V) [33] implementation, which is frequently used for modeling semantic word relationships [12]. We build the word embedding model based on the preprocessed ELAN dataset, which we grouped by ICPC code to obtain a 'document' for each ICPC code. We used the common bag-of-words (CBOW) method and each word vector in the model is 300-dimensional.

For visualisation purposes the we selected the 2000 most frequent tokens and reduced the number of dimensions of their word vectors to 2 dimensions using t-Distributed Stochastic Neighbor Embedding (t-SNE) [32]. Figure 4.5 shows a scatter plot of these reduced word vectors in which we annotated the 150

Token	KL divergence
immunisatie_preventieve_medicatie (immunization)	0.005351
cerumen_overmatig (excessive earwax)	0.005084
duizeligheid (dizziness)	0.004352
copd	0.004261
obstipatie (constipation)	0.004231
presbyacusis	0.004032
neusbloeding (nosebleed)	0.003778
arteriitis	0.003762
erysipelas	0.003635
prostaathypertrofie (prostatitis)	0.003504

Table 4.2: Tokens with highest KL divergence. A high KL divergence indicates a relatively high frequency in the target group (stroke patients) in comparison to the frequency in the non-target group.

most frequent tokens. Furthermore a medical expert annotated several medical systems using colored ellipses. This shows that the word embedding model adequately represents and clusters medical information.

In a later phase of the pipeline we will also use the word embedding model to determine the optimal number of topics in the topic models by measuring the word similarities of the words in the generated topics.

4.3 Feature selection

In this section we describe two feature selection techniques: bag-of-words and topic modeling. Bag-of-words is a basic technique that transforms token counts to features, while topic modeling is a more complex technique that creates interpretable 'topics' based on token co-occurrence, which has been used in various studies in the medical domain [11, 20, 29]. We use both techniques to create two separate feature sets, which will be used as input for the prediction model in the next phase.

4.3.1 Bag-of-words

The bag-of-words model is a way of representing text and is traditionally commonly used in the natural language processing (NLP) domain. In this model a sentence or document is represented as a multiset of words. A multiset is a set that can contain a multiple instances of the same word. Due to this representation, grammar and word order are disregarded while multiplicity is kept. The number of features created by this model is consequently equal to the number of unique words in the whole dataset. Applying this model to our relatively large dataset will yield a immense feature set containing $19,000+$ features. Therefore, we restrict the feature set to words that occur at least 10 times in the dataset.

This reduces the feature set to 3409 features while simultaneously filtering rare words from the data.

4.3.2 Topic Modeling

A topic model is a type of statistical model for discovering abstract 'topics' –sets of related words– that occur in a collection of documents based on word co-occurrence. We use this technique both to obtain topical information about the data and to obtain a feature set (the topic probabilities) for later prediction models. Multiple approaches for obtaining topics exist, including the probabilistic Latent Dirichlet Allocation (LDA) [6] and the deterministic Non-Negative Matrix Factorization (NMF) [43].

Figure 4.5: Word embedding model after dimensionality reduction. The 150 most frequent words are annotated and some medical domains are indicated using ellipses based on expert knowledge. Red: psychosocial problems, yellow: ear-related problems, blue: upper respiratory tract problems, black: musculoskeletal problems, pink: dermatological problems, green: eye-related problems

Tokens	Topic 0	Topic 1	Topic 2
[epistaxis, neusbloeding, frequente_mictie_aandrang, ome]	0.0	0.000000	0.000000
[presbyacusis, eczeem]	0.0	0.022143	0.000000
[veronderstelde_gastro-intestinale_infect, uitslag, uitslag, obstipatie, down, depressief, gevoel, cystitis, herpes_zoster, griepvaccinatie, aortastenose]	0.0	0.000000	0.005414

Table 4.3: Three random patients with multiple medical problems, and their average topic probabilities of the first three topics obtained using the topic model described in Section 4.3.3. These topics are listed in Appendix D.

Topic modeling approaches are parameterized. One of the most important parameters is the number of topics. Specifying the desired number of topics is required since topic models are unable to automatically determine this based on the data. Other parameters include the number tokens to consider. Including all tokens would highly increase the time-complexity, while extremely rare tokens also do not provide any additional value to the topic model. We set the number of tokens to 1500. For LDA we choose the tokens with the highest term frequency (tf) and for NMF we choose those tokens with the highest tf-idf (term frequency-inverse document frequency) [41]. Furthermore, we only consider the first ten (highest-scoring) tokens in each topic.

The topic models are based on the dataset we obtained after the preprocessing phase. Single records typically describe one single problem and therefore co-occurrence of words in these records will yield more coherent topics opposed to using the data that was grouped per patient. Moreover, the full preprocessed dataset contains significantly more data and more patients than the filtered and grouped dataset. After obtaining the topic probabilities for each record we compute the average topic probability for each patient, which are the probabilities that we will use as features in the logistic regression model.

Finally, to obtain a single set of topic features for each patient, we compute the average of the topic probabilities of all his records. Table 4.3 lists three random patients with their topic features for the first three topics to demonstrate these topic features.

Topic coherence

The exact number of topics present in textual data is typically unknown. Even the ICPC diagnosis coding system changes regularly; an exact number of diseases or symptoms does not exist. Therefore, we create multiple topic models with topic sets ranging from 18 to 40 topics after which we compute which of

these topic models contains the most 'coherent' topics. Topic coherence is a measure of the degree of semantic similarity between the high scoring words in a topic, which we define as the euclidean distance between the word vectors of these words (TC-W2V) [38] in the word embedding model that we created in the previous phase. Similarly, we define the 'topic model coherence' as the average of the topic coherences of all topics in the topic model.

Fig. 4.6 shows the coherence of all NMF topics models. According to TC-W2V the topic model with 26 topics is the optimal topic model. However, inspection of the topics in this model reveals that the 26 topics lacked some common medical problems (e.g. *hypertensie* (hypertension) and diabetes mellitus), which might also be important for stroke prediction. Since the 32-topic model has only a slightly worse TC-W2V and its topics contain a broader range of medical problems, we use this model instead. Interestingly, the topic model coherence for the 32-topic LDA topic model is 0.300, which is substantially lower than the topic model coherence of the NMF topic model.

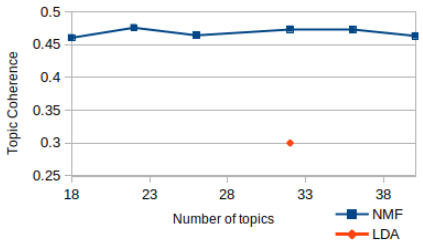

Figure 4.6: Topic model coherences

4.3.3 Topic modeling results

The topic model obtained using NMF[1] with 32 topics is listed in Appendix D. Most topics describe a clear set of symptoms and it would be relatively easy for a medical professional to assign a name to each topic that describes (almost) all top tokens in the topic. The topics are mainly about fairly common medical issues, including respiratory problems, eczema and common injuries, which corresponds with the general statistics about care provided by general practitioners in the Netherlands [36]. The results of the LDA[2] topic model with 32 topics are listed in Appendix C. The topics in this model are clearly less interpretable than the topics from the NMF model, which corresponds with the topic coherences observed earlier, and thus we choose NMF as our method for topic modeling.

Interestingly, even with this relatively small amount of topics there seem to be multiple topics for a single type of problem (e.g. ear infections) in the NMF topic model. We further investigated this observation using a correlation matrix to discover inter-topic correlations (see Fig. 4.7). For each topic pair we used the ranked Spearman coefficient to determine whether the topic probabilities are correlated (either positively or negatively). The correlation plot shows mainly weakly positively correlated topics and only one pair of topics (topic 13 and topic 29) which is strongly positively correlated with a correlation coefficient of 0.69. Both topics in this pair are about ear infection. Two other pairs, with correlation coefficients 0.60 and 0.59 respectively, also show some positive correlation. The first pair, topic 3 and topic 11, are both broad topics about pain and other complaints in specific joints and/or other parts of the body. The other pair, topic 13 and topic 26, share some tokens (e.g. *links* (left) and *rechts* (right)) which explains the positive correlation observed.

4.4 Prediction models

The final step in the stroke prediction pipeline is the development of the prediction models. Logistic regression models are a commonly used approach in the medical domain, due to their high interpretability. Unlike complex machine learning models (e.g. neural networks) these models return exactly how each input feature contributed to the prediction by means of regression coefficients.

4.4.1 Model development

We develop two logistic regression models, one with the bag-of-words features as predictors (BOW-LR) and one with the topics as predictors (NMF-LR). The effectiveness of the logistic regression models is measured using the precision (the fraction of relevant instances among the retrieved instances), recall (the fraction of all relevant instances that were retrieved) and the area under the

[1] Implementation details are listed in Appendix D.
[2] Implementation details are listed in Appendix C.

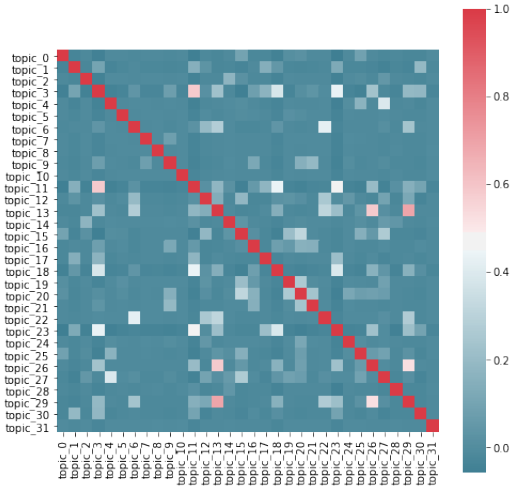

Figure 4.7: Topic correlation (Spearman)

precision-recall curve (AUCPR).

Additionally, we develop two similar models that predict the age group of the patient. We define two age groups similar to the grouping used in related research to distinguish 'young' stroke patients from 'old' stroke patients, with ages $18 - 50$ and $50+$ respectively. These models are used to investigate the confounding effect of age on the predictors of the stroke prediction model. A reduction of the confounding effect of age can also be obtained by including the patient's age as a predictor in the stroke classification model, but consequently the prediction model is no longer strictly based on features from text, which will also influence the evaluation results (e.g. AUCPR score).

4.4.2 Imbalanced groups

A naive prediction model that predicts all patients as non-target patient will already be 98.7% correct when applied to the ELAN dataset, since only 1.3% of the patients are stroke patients. We avoid this behaviour using Synthetic Minority Oversampling Technique (SMOTE) [10], which synthesises new minority instances between existing minority instances up to the desired ratio.

Unfortunately, when SMOTE is applied to high-dimensional data such as bag-of-words features, it only slightly attenuates the bias towards the majority class [31]. Although the accuracy on the minority (stroke) class might remain low despite using SMOTE, the model will be more sensible than a non-oversampled model that predicts each instance as a majority (non-stroke) instance. We set the oversampling ratio to 0.75.

4.4.3 Classification results

Each logistic regression model was ran with a maximum of 200 iterations. The classification report of both logistic regression models is shown in Table 4.4. As expected both the BOW-LR model and the NMF-LR model mainly predict instances as non-stroke cases, with recall values 0.29 and 0.15 for the stroke cases respectively. Interestingly, the recall of stroke cases in the BOW-LR model is twice as high as the recall in the NMF-LR model. Due to the class imbalance problem, both models have extremely low AUCPR scores.

		BOW-LR			**NMF-LR**		
Stroke	#	Precision	Recall	AUCPR	Precision	Recall	AUCPR
False	9377	0.99	0.89		0.99	0.93	
True	142	0.04	0.29		0.03	0.15	
Overall:				**0.027**			0.023

Table 4.4: Logistic regression classification report for BOW-LR and NMF-LR for stroke classification

Naturally the groups in the age prediction model are almost equally balanced and thus did not require oversampling. The AUCPR scores for age prediction using the BOW-LR and NMF-LR models, are 0.805 and 0.572 respectively. Interestingly, for age prediction the BOW-LR model performed substantially better than the NMF-LR model. This might indicate that BOW-LR is a better choice when the data is balanced, while NMF-LR is more suited when the data is (strongly) imbalanced.

In the next two sections we will investigate the BOW and NMF features that were the most important according to these models. Furthermore, we will visualise how the predictive value of the features in the stroke prediction models correspond to the predictive value of the same features in the age prediction models.

4.4.4 Bag-of-word predictors

Table 4.5 shows the top 7 predictors with the highest logistic regression coefficients, both for stroke classification (left) and age classification (right). The first predictor for stroke classification is *immunisatie_preventieve_medicatie* (immunization/preventive medication). This is in most cases the description for *griepprik* (flu vaccine). In the Netherlands everyone aged 60+ is called for a yearly flu vaccine by their general practitioner. Other predictors in the top 7 are also highly age-related (e.g. cataract (*staar*)). Some of the top predictors for age classification are similar to the predictors in the stroke prediction model, again indicating that these predictors might be influenced by the average age of the stroke-patients rather than being actual risk factors for stroke.

BOW-LR for stroke		BOW-LR for age	
Top predictors	**Coeff.**	**Top predictors**	**Coeff.**
immunisatie_prevent...	17.455890	immunisatie_prevent...	7.440464
aortastenose	5.179181	cataract	5.182603
goede	4.127771	dementie	3.616307
cataractoperatie	3.939823	presbyacusis	3.358967
statine	3.833086	gonartrose	3.147137
inguinalis	3.746110	nierfunctiestoornis	2.929755
erisypelas	3.733573	euthanasie	2.904145

Table 4.5: Bag-of-words predictors for stroke (left) and for age (right)

We further investigate the relation between age predictors and stroke predictors by visualising the relationship between the regression coefficients in both models in Figure 4.8. We normalized all regression coefficients to values ranging from -1 to 1 (See Equation 4.3). We also annotated all tokens with at least one positive regression coefficient and an euclidean distance of at least 0.45 from the mean. This threshold reduces the number of overlapping annotations and prevents the annotation of (less-informative) neutral predictors. The tokens in the upper left quadrant are annotated in red (because these are the tokens that are more positively correlated with stroke and negatively correlated with age), while all other annotations are blue. The point size and annotation size is relative to the frequency of the tokens to direct the focus to the more frequent and consequently more reliable tokens.

$$x_N = \frac{x - min(X)}{max(X) - min(X)} * 2 - 1 \tag{4.3}$$

As observed before, positive predictors for age include a wide range of age-related diseases, while negative predictors include pregnancy, STD and menstruation related problems. The annotations also contain some stroke-related diseases, but surprisingly not all of these are positive predictors for stroke in this model. Diabetes mellitus (*diabetes_mellitus_type*) is a positive predictor, but heart disease (*hvz*) and transient ischemic attack (*tia*) are not. Hypertension and high cholesterol, two known stroke-related comorbidities, are not among the annotated tokens.

Interestingly, the set of tokens with (strong) positive coefficients for stroke and neutral coefficients for age (upper middle) contain many possibly neurological symptoms, including tension headache (*spanningshoofdpijn)*, vertigo (dizziness) and syncope (fainting). Similar symptoms including headache, migraine and tiredness are also situated in the upper left quadrant. The upper-central area also contains several symptoms that are currently not associated with stroke risk, including excessive earwax (*cerumen_overmatig*), urinary problems (*mictieklachten*) and edema (*oedeem*). Although there might be no direct relation to

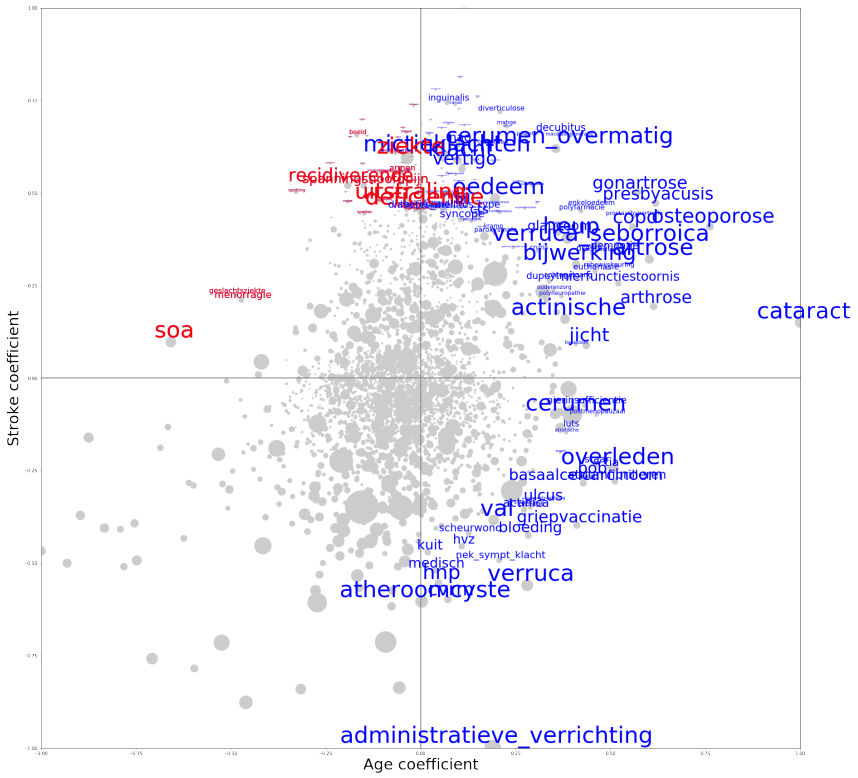

Figure 4.8: Normalized BOW-LR coefficients for both regression models.

stroke, these symptoms might be indicators for other (stroke-related) diseases or side effects of medications for these diseases.

4.4.5 Topic predictors

Similarly, we create a scatterplot for the stroke and age predictors of the NMF-LR model (Fig. 4.9). Again we normalized the regression coefficients to the range $[-1, 1]$ and we annotated all topics with the topic number and the highest scoring word in the topic. The top 10 words in each topic are listed in Appendix D.

The plot shows that the majority of topics is a positive predictor for age, which is sensible. More surprisingly, the majority of topics is also a positive predictor of stroke and none of the topics is both a positive predictor for age and a negative predictor for stroke. Some relatively similar topics pairs (e.g. 2-14 and 16-21) appear also close together in the plot.

Further inspection of the plot reveals that topic 28, which includes hypertension (*hypertensie*), cardiovascular riskmanagement (*cvrm*), hypercholesterolemia (*hypercholesterolemie*) and diabetes mellitus, is a good predictor for both age and stroke, whereas the BOW scatterplot was more ambiguous about these stroke comorbidities. Even more noticeable than in the BOW scatterplot, we observed that predictive value of topics 16 and 21, which together include headache (*hoofdpijn*), migraine, tension headache (*spanningshoofpijn*), tiredness (*moe, moeheid*), dizziness (*duizeligheid*) and burnout (*surmenage*) , is relatively high for stroke while being negative predictors for age. These observations are consistent with what is known about these risk factors from scientific literature.

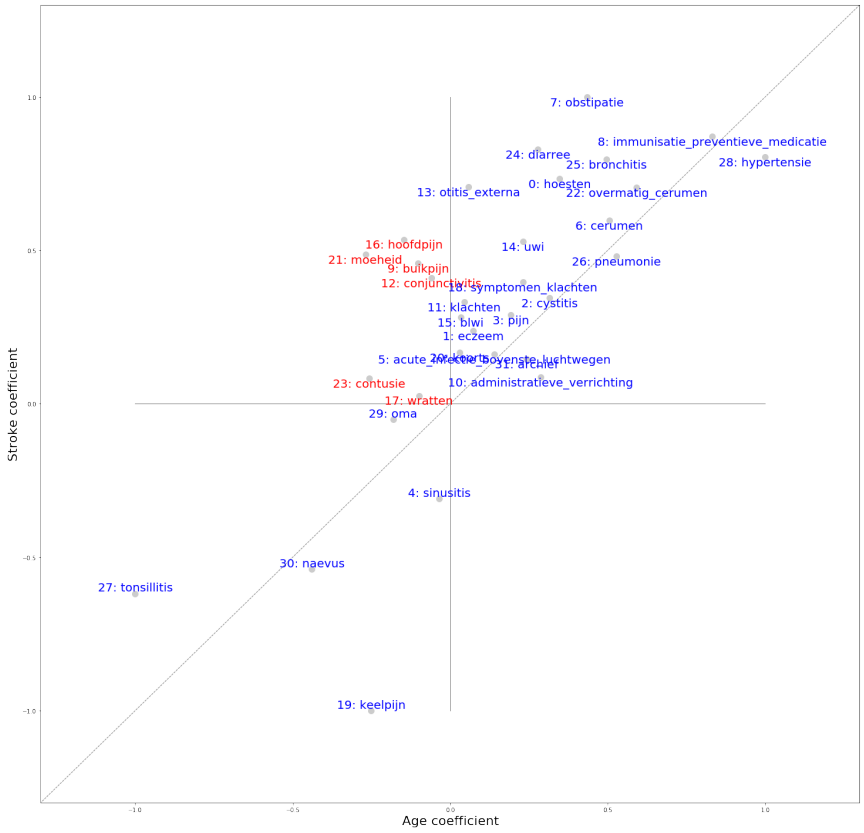

Figure 4.9: Normalized topic regression coefficients: stroke regression coefficients are on the y-axis and age regression coefficients are on the x-axis

Chapter 5

Discussion

The results of this research show that all techniques described in the previous chapter –logistic regression with bag-of-words and topic features, feature extraction and word frequency analysis– are able to extract sensible information from Dutch diagnosis descriptions in EMRs from general practitioners. Further analysis also revealed information that is potentially relevant for stroke prediction. When we compare the results of KL divergence and BOW-LR we see similar tokens being associated with stroke patients. Topic modeling using non-negative matrix factorization combined with topic coherence optimization led a set of 32 interpretable topics about common medical problems.

Stroke prediction using bag-of-words (BOW-LR) outperformed stroke prediction using the these topics features (NMF-LR) in terms of recall of the stroke cases (Table 4.4), which is consistent with the results of a similar study on colorectal cancer prediction [27]. However, the results of NMF-LR are particularity valuable for discovering the relationship between common medical problems and stroke. Furthermore, we build an word embedding model that showed a clear clustering of medical problems (Fig. 4.5).

Additionally, we observed an strong confounding effect of age consistent with the incidence pattern of stroke [1] and other diseases. We build a similar prediction model for age prediction and by comparing the results of this model with the results of stroke prediction we obtained clear visualisations of this effect (Fig. 4.8, Fig. 4.9).

Limitations

All techniques were applied to the diagnosis descriptions from the ELAN dataset. With a mean length of only 19 characters and a significant amount of descriptions being equal to the ICPC names, there is a limited amount of relevant information that can be potentially extracted. Factors that we hypothesised to play a role in stroke development (e.g. psychosocial problems) were simply

almost not present in this type of data. Furthermore, these short notes lack sentence structure, rendering part-of-speech tagging (PoS) for concept extraction, negation detection and the extraction of other contextual information infeasible. The relatively small amount of stroke patients in this dataset (517 individuals) limited our possibilities for balancing the patient groups. For similar reasons, techniques for reducing the confounding effect of age (e.g. age sampling) were also not feasible with this data. Furthermore, due to the imbalanced nature we were forced to use an oversampling, which can potentially introduce bias in the prediction model.

In spite of these limitations we were able to obtain sensible results from each technique. Furthermore, this was a great opportunity to show the value of diagnosis descriptions in disease prediction, which might also be a good alternative for others who face obstacles in the process of obtaining consultation notes. In addition to being regarded as less privacy sensitive, diagnosis descriptions are also less noisy, require less computational effort and less preprocessing.

Another important limitation to note is the current lack of tools for processing Dutch medical texts (e.g. for UMLS concept extraction) in comparison to the number of tools available for medical texts written in English. Lastly, we were only able to test a limited set of techniques and parameter settings, due to the narrow scope and the finite amount of time available for this research.

Strengths

Most disease prediction studies only use the structured data in EMRs and mainly use complex machine learning models. The approach we used in this research to create an explainable model while also obtaining possible risk factors from textual EMR data has a number of advantages. We not only extracted information that is possibly not present in the structured fields but also combined the information extraction with text cleaning methods, spelling correction, word completion and key-phrase detection such that the noise typically associated with natural language is reduced.

Additionally, we improved the interpretability even further by grouping the data into topics. By using a simple logistic regression model we also ensured that not only the features were explainable but also the final predictions. For medical applications, understanding and being able to interpret models is key, since unintended and undetected biases in the model could have disastrous consequences [8, 21].

We also demonstrated a possibly novel approach for filtering confounding effects, which does not require any form of statistical knowledge but is still a reliable way to obtain insights into these effects. Furthermore, we contributed to the field of medical NLP by developing a natural language preprocessing pipeline for Dutch medical notes.

At several stages of this research we consulted a medical expert (Hine van Os) for feedback and interpretation of the results, which led to new insights and directions to focus on.

Future directions

In future work we plan to combine the bag-of-words and/or topic features with the features extracted from the structured fields into a single model or into an ensemble model to investigate the added value of using the texts from EMRs for the prediction of stroke. We also plan to compare this model with the basic risk factor algorithm that is currently used by general practitioners.

Furthermore, once we have access to the STIZON dataset we plan to apply the techniques also to this significantly larger dataset. After some small modifications to the preprocessing phase we could also apply the pipeline to the consultation notes to extract even more information for stroke prediction.

Other interesting directions for future work include developing a tool for Dutch medical concept extraction, experimenting with other approaches for dealing with strongly imbalanced high-dimensional data or including diseases similar to stroke (e.g. cardiovascular diseases) in the prediction model. Finally and importantly, it would be valuable to build female-specific prediction models or age-specific prediction models, since evidence that different mechanisms play a role in these groups is mounting.

Chapter 6

Conclusion

In this research we proposed a novel approach for stroke prediction by utilizing unstructured information from EMRs. Stroke is one of the leading causes of death in the Netherlands and early recognition of persons at risk could potentially save lives. For real-life applications in diagnostication, the medical experts who will be using the model should be able to trust and understand the model completely.

Stroke occurs mainly in elderly persons and consequently prediction models for stroke tend to indicate the numerous amount of age-related problems as possible indicators for stroke, thereby obfuscating the actual predictors for stroke. This research aimed to answer to following questions:

1. How can we develop an explainable text mining pipeline for the prediction of stroke?

2. How can we discern the confounding effect of age from this information?

We developed a pipeline for the prediction of stroke based on unstructured information from Dutch medical notes, more specifically diagnosis descriptions in EMRs from general practitioners.

We have compared two feature selection techniques, namely bag-of-words and topic modeling, and we build an explainable logistic regression model for stroke prediction. Furthermore, we analysed the confounding effect of age, again using methods that are interpretable by medical experts.

Logistic regression models using either bag-of-words and topic features provide interpretable information about stroke risk. We found that logistic regression with topic features allow for the easiest interpretation about common medical problems and their relation to stroke. Additionally, we demonstrated that KL divergence applied to the relative frequencies of the words in each patient group,

is an interpretable method for showing which words are more frequent in the target group.

We developed clear visualisations of annotated logistic regression coefficients for both age prediction and stroke prediction, that discern predictors (tokens or topics) that are good predictors for both stroke and age from predictors that are positive predictors stroke but negative predictors for age. Additionally, we computed the *mean token age* of each token and found a broad distribution of mean token ages, which we plotted against the KL divergence (Fig 4.4). These visualisations discern to some extend the confounding effect of age from the extracted information, while also being easy to interpret by medical experts or others without advanced statistical knowledge.

Overall, we have shown multiple techniques for extracting relevant information from texts, which proved to be not only interpretable but also are promising approaches for stroke prediction in real-life applications. It can be considered as a first step in shifting the focus from structured health data towards the inclusion of free text data and although we were unable to demonstrate this due to the limitations of our dataset, we expect that these techniques have the potential to aid the discovery of novel risk factors from texts and potentially also improve the accuracy of stroke prediction models.

With this research we explored the utilization of unstructured clinical data in disease prediction models, we advanced the understanding of the steps involved in (pre)processing Dutch medical texts and we hoped to advance the knowledge about risk factors for stroke.

Bibliography

[1] Beroerte - cijfers & context - huidige situatie. https://www.volksgezondheidenzorg.info/onderwerp/beroerte/cijfers-context/huidige-situatie#node-prevalentie-beroerte-naar-leeftijd-en-geslacht

[2] Cijfers hart- en vaatziekten. https://www.hartstichting.nl/hart-en-vaatziekten/feiten-en-cijfers-hart-en-vaatziekten

[3] Afzal, Z., Pons, E., Kang, N., Sturkenboom, M.C., Schuemie, M.J., Kors, J.A.: Contextd: an algorithm to identify contextual properties of medical terms in a dutch clinical corpus. BMC bioinformatics **15**(1), 373 (2014)

[4] Bard, G.V.: Spelling-error tolerant, order-independent pass-phrases via the damerau-levenshtein string-edit distance metric. In: Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68. pp. 117–124. Citeseer (2007)

[5] Beeksma, M., Verberne, S., van den Bosch, A., Das, E., Hendrickx, I., Groenewoud, S.: Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. BMC medical informatics and decision making **19**(1), 36 (2019)

[6] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)

[7] Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)

[8] Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., Tsaneva-Atanasova, K.: Artificial intelligence, bias and clinical safety. BMJ Qual Saf **28**(3), 231–237 (2019)

[9] Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics **34**(5), 301–310 (2001)

[10] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)

[11] Cohen, R., Aviram, I., Elhadad, M., Elhadad, N.: Redundancy-aware topic modeling for patient record notes. PloS one **9**(2) (2014)

[12] De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., Bruza, P.: Medical semantic similarity with a neural language model. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management. pp. 1819–1822 (2014)

[13] DeLisle, S., South, B., Anthony, J.A., Kalp, E., Gundlapallli, A., Curriero, F.C., Glass, G.E., Samore, M., Perl, T.M.: Combining free text and structured electronic medical record entries to detect acute respiratory infections. PloS one **5**(10), e13377 (2010)

[14] Dirkson, A., Verberne, S., Kraaij, W.: Narrative detection in online patient communities. In: Text2Story@ ECIR. pp. 21–28 (2019)

[15] Divita, G., Zeng, Q.T., Gundlapalli, A.V., Duvall, S., Nebeker, J., Samore, M.H.: Sophia: a expedient umls concept extraction annotator. In: AMIA Annual Symposium Proceedings. vol. 2014, p. 467. American Medical Informatics Association (2014)

[16] Fares, M., Kutuzov, A., Oepen, S., Velldal, E.: Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden. pp. 271–276. No. 131, Linköping University Electronic Press (2017)

[17] Fivez, P., Šuster, S., Daelemans, W.: Unsupervised context-sensitive spelling correction of English and Dutch clinical free-text with word and character n-gram embeddings. arXiv preprint arXiv:1710.07045 (2017)

[18] Ford, E., Carroll, J.A., Smith, H.E., Scott, D., Cassell, J.A.: Extracting information from the text of electronic medical records to improve case detection: a systematic review. Journal of the American Medical Informatics Association **23**(5), 1007–1015 (2016)

[19] Garbe, W.: Improved edit-distance based spelling correction. `https://github.com/wolfgarbe/SymSpell` (2012)

[20] Ghassemi, M., Naumann, T., Joshi, R., Rumshisky, A.: Topic models for mortality modeling in intensive care units. In: ICML machine learning for clinical data analysis workshop. pp. 1–4 (2012)

[21] Gianfrancesco, M.A., Tamang, S., Yazdany, J., Schmajuk, G.: Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine **178**(11), 1544–1547 (2018)

[22] Golding, A.R., Roth, D.: A winnow-based approach to context-sensitive spelling correction. Machine learning **34**(1-3), 107–130 (1999)

[23] Goldstein, B.A., Navar, A.M., Pencina, M.J., Ioannidis, J.: Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association **24**(1), 198–208 (2017)

[24] Goryachev, S., Sordo, M., Zeng, Q.T., Ngo, L.: Implementation and evaluation of four different methods of negation detection. Boston, MA: DSG (2006)

[25] Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W.: Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of biomedical informatics **42**(5), 839–851 (2009)

[26] Henderson, K.M., Clark, C.J., Lewis, T.T., Aggarwal, N.T., Beck, T., Guo, H., Lunos, S., Brearley, A., Mendes de Leon, C.F., Evans, D.A., et al.: Psychosocial distress and stroke risk in older adults. Stroke **44**(2), 367–372 (2013)

[27] Hoogendoorn, M., Szolovits, P., Moons, L.M., Numans, M.E.: Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artificial intelligence in medicine **69**, 53–61 (2016)

[28] Kop, R., Hoogendoorn, M., Ten Teije, A., Büchner, F.L., Slottje, P., Moons, L.M., Numans, M.E.: Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Computers in biology and medicine **76**, 30–38 (2016)

[29] Lehman, L.w., Saeed, M., Long, W., Lee, J., Mark, R.: Risk stratification of icu patients using topic models inferred from unstructured progress notes. In: AMIA annual symposium proceedings. vol. 2012, p. 505. American Medical Informatics Association (2012)

[30] Louwe, A.: Exploring textual data in electronic medical records. Research project at LIACS/LUMC, Leiden University (2019)

[31] Lusa, L., et al.: Smote for high-dimensional class-imbalanced data. BMC bioinformatics **14**(1), 106 (2013)

[32] Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)

[33] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)

[34] NICTIZ: SNOMED CT. https://www.nictiz.nl/standaardisatie/terminologiecentrum/snomed-ct/

[35] NICTIZ: Waardelijst ICPC-1. https://decor.nictiz.nl/ketenzorg/kz-html-20141013T173536/voc-2.16.840.1.113883.2.4.3.11.60.103.11.12-2011-10-12T000000.html

[36] NIVEL: Zorg op de huisartsenpost. https://www.nivel.nl/sites/default/files/bestanden/jaarrapport_huisartsenpost_2017.pdf

[37] Ohno-Machado, L.: Modeling medical prognosis: survival analysis techniques. Journal of biomedical informatics **34**(6), 428–439 (2001)

[38] Ocallaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications **42**(13), 5645–5657 (2015)

[39] Pantazis, I.: Dealing with the temporal structure of routine primary care data (2019)

[40] Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., Goertzel, B., Watts, B., Flashman, L., McAllister, T.: Predicting the risk of suicide by analyzing the text of clinical notes. PloS one **9**(1), e85733 (2014)

[41] Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 133–142. Piscataway, NJ (2003)

[42] Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association **17**(5), 507–513 (2010)

[43] Shahnaz, F., Berry, M.W., Pauca, V.P., Plemmons, R.J.: Document clustering using nonnegative matrix factorization. Information Processing & Management **42**(2), 373–386 (2006)

[44] STIZON: Stichting informatievoorziening voor zorg en onderzoek. https://www.stizon.nl/

[45] Šuster, S., Tulkens, S., Daelemans, W.: A short review of ethical challenges in clinical natural language processing. arXiv preprint arXiv:1703.10090 (2017)

[46] Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction. In: Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment. pp. 33–40 (2003)

[47] Tu, J.V.: Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology **49**(11), 1225–1231 (1996)

[48] Tzourio, C., Benslamia, L., Guillon, B., Aidi, S., Bertrand, M., Berthet, K., Bousser, M.: Migraine and the risk of cervical artery dissection: a case-control study. Neurology **59**(3), 435–437 (2002)

[49] Tzourio, C., Tehindrazanarivelo, A., Iglesias, S., Alperovitch, A., Chedru, F., d'Anglejan Chatillon, J., Bousser, M.G.: Case-control study of migraine and risk of ischaemic stroke in young women. Bmj **310**(6983), 830–833 (1995)

[50] Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M., Qureshi, N.: Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS one **12**(4), e0174944 (2017)

[51] Yiyu, Y.: Medical entity recognition from patient forum data. Masters thesis (2017)

[52] Zeng, Q.T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S.N., Lazarus, R.: Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC medical informatics and decision making **6**(1), 30 (2006)

Appendices

A Negation trigger words

A list of Dutch negation trigger words based on [3].

‘afwezigheid van ’, ‘evenmin’, ‘gedaald’, ‘ geen ’,‘ geen aanwijzingen voor ’,‘ geen klachten van ’,‘geen oorzaak van ’,‘ geen teken van ’,‘ geen tekenen van ’,‘ heeft geen ’,‘ kan niet ’,‘ leek niet ’,‘ niet ’,‘ niet als ’,‘onbekend ’,‘ uitsluiten’,‘ verdwenen’,‘ vertonen geen ’,‘ vertoonde geen ’,‘vrij van ’,‘weg ’,‘ zonder’,‘ zonder tekenen van’

B Stopwords

A list of Dutch stopwords based on [17] excluding negation (stop)words.

aan, af, al, alles, als, altijd, andere, ben, bij, daar, dan, dat, de, der, deze, die, dit, doch, doen, door, dus, een, eens, en, er, ge, geweest, haar, had, heb, hebben, heeft, hem, het, hier, hij, hoe, hun, iemand, iets, ik, in, is, ja, je, kan, kon, kunnen, maar, me, meer, men, met, mij, mijn, moet, na, naar, niets, nog, nu, of, om, omdat, onder, ons, ook, op, over, reeds, te, tegen, toch, toen, tot, u, uit, uw, van, veel, voor, want, waren, was, wat, we, wel, werd, wezen, wie, wij, wil, worden, wordt, zal, ze, zei, zelf, zich, zij, zijn, zo, zou

C LDA Topic model

Implemented using `sklearn.decompisition.LatentDirichletAllocation`

Parameters:
Number of topics: 32
Maximum number of iterations: 10
Learning method: Online
Learning offset: 10
Learning decay: 0.7
Batch size: 128
Random state: 0

Coherence (W2V-TC): 0.3001

Topics:

0. contusie, overmatig_cerumen, rechter, verruca, zwangerschap, seborrhoica, onderbuik, nieuwe, gewricht, cyste
1. ziekte, geen, dermatomycose, hand, onychomycose, folliculitis, wondje, medische_gegevens, probleem, ontstoken
2. pneumonie, mictieklachten, trauma, hyperreactiviteit, handen, lichte, gewone_verkoudheid, contacteczeem, arthrose, hand_vinger_symptomen_klachten
3. rechts, vinger, urticaria, wond, menstruatie, dossier, ontsteking, gastroenteritis, zwanger, medisch
4. klachten, buikpijn, tonsillitis, lumbago, mond, cataract, acute, buikkrampen, tong, abces
5. huid, moe, viraal, droge, surmenage, bloedverlies, gelaat, hypercholesterolemie, artrose, epistaxis
6. linker, pijnlijke, benen, varices, ongeval, letsel, bijwerking, reflux, hielspoor, ander
7. smear, infectie, partner, dyspnoe, incontinentie, ziektegevoel, moeheid, urine, huid_subcutis, relatieproblemen
8. diarree, allergie, vrouw, kneuzing, keratose, actinische, malaise, borsten, chronisch, preventie
9. symptomen_klachten, schouder, pijn, borst, arm, migraine, verhoogde, oac, mycose, bloeddruk
10. gevoel, luchtweginfectie, anemie, schaafwond, bovenste, afwijkend, chron, tekenbeet, nekpijn, gehoor

11. links, fractuur, dig, been, oog, oor, lsp, voet, hand, ulcus

12. knie, rugpijn, acne, otitis_media, tgv, teen, acute, pharyngitis, spruw, mictie

13. sinusitis, moeheid, acute, chronische, obv, lwi, zwakte, unguis_incarnatus, hematoom, familie

14. jeuk, vaginale, astma, verruca_seborroica, onderzoek, zorgen, vit, klacht, osteoporose, b12

15. wratten, soa, zoon, dochter, partus, bevallen, anticonceptie_orale_anticonceptie, waterpokken, angst, normale

16. eczeem, zwelling, syndroom, duizeligheid, lies, constitutioneel, seborroisch, misselijkheid, vertigo, oogleden

17. hypertensie, bdz, rugklachten, reactie, mamma, allergische, lage-rugpijn_zonder_uitstraling, lage, zonder, orgaanbeschadiging

18. obstipatie, uwi, oorpijn, ogen, bvo, onderbeen, erysipelas, naevi, droge, visus

19. conjunctivitis, cerumen_overmatig, nvzb, herpes_zoster, knie_symptomen_klachten, lipoom, bacteriele, verdenking, wang, irritatie

20. blwi, virale, borstkas_symptomen_klachten, hordeolum, maagpijn, ome, vitamine, slaapstoornis, slapeloosheid, verkoudheid

21. administratieve_verrichting, oma, hoofd, snijwond, depressie, lage_rugpijn, verkouden, corpus_alienum, licht, influenza

22. acute_infectie_bovenste_luchtwegen, hoofdpijn, enkel, distorsie, anticonceptie, hooikoorts, braken, iud, collaps, verstuiking

23. problemen, bursitis, nek, elleboog, maagklachten, recidiverende, hyperventilatie, myalgie, operatie, diabetes_mellitus_type

24. otitis_externa, val, scheurwond_snijwond, duim, nekklachten, voeten, hallux, voet_teen_symptomen_klachten, hemorroiden, beet

25. naevus, candida, acute_bronchitis_bronchiolitis, schouderklachten, knieklachten, moedervlek, rug_symptomen_klachten, oedeem, cvrm, medicatie

26. hoesten, impetigo, pijnklachten, brandwond, spanningshoofdpijn, impetiginisatie, mollusca, contagiosa, slaapproblemen, atheroom

27. cystitis, neus, smear_bvo, candidiasis, overleden, epicondylitis_lateralis, urineweginfectie, nek_sympt_klacht, spierpijn, ganglion

28. koorts, keelpijn, bronchitis, atheroomcyste, hals, oud_dossier, roken, vasectomie, angina, bloeding

29. angst, archief, lage_rugpijn_zond_uitstr, eci, frequente_mictie_aandrang, tinea_pedis, hernia, ivm, man, oksel

30. cerumen, immunisatie_preventieve_medicatie, griep, overgewicht, jicht, mogelijk, balanitis, recidief, spanningsklachten, hematurie

31. pijn, voet, rug, tendinitis, niet, wrat, pols, dermatofibroom, thorax, thoracale

D NMF Topic model

Implemented using `sklearn.decompisition.NMF`

Parameters:
Number of features: 3000
Number of topics: 32
Initialization: Nonnegative Double Singular Value Decomposition (NNDSVD)
Alpha: 0.1
L1 ratio: 0.5
Random state: 42

Coherence (W2V-TC): 0.4734

Topics:

0. hoesten, hyperreactiviteit, obv, verkouden, kinkhoest, viraal, griep, astma, verkoudheid, dyspnoe
1. eczeem, constitutioneel, seborroisch, seborrhoisch, dyshidrotisch, roos, handen, mycose, seb, oogleden
2. cystitis, urineweginfectie, urineweginfecties, nao, recidiverende, rec, recidiverend, geen, prostatitis, recidief
3. pijn, knie, borst, onderbuik, schouder, been, thorax, heup, voet, rug
4. sinusitis, acute, chronische, chron, recidiverende, rhinitis, verkoudheid, verkouden, recidief, beginnende
5. acute_infectie_bovenste_luchtwegen, viraal, recidiverende, boven, griep, wsch, icm, hoest, vaak, rec
6. cerumen, bdz, ads, oor, verwijderd, ome, oren, cerumen_overmatig, beide, oorpijn
7. obstipatie, ibs, tgv, obv, buikklachten, zwangerschap, chronische, diverticulitis, colon, gebruik
8. immunisatie_preventieve_medicatie, griepvaccinatie, tetanus, griep, bcg, hep, reis, zweten, goed, gold
9. buikpijn, gelokaliseerde, buikkrampen, gegeneraliseerde, gegen, eci, onderbuik, ibs, diverticulitis, appendicitis
10. administratieve_verrichting, ion, kennismaking, aangemeld, niet, verhuizing, algemeen, nieuwe, verhuisd, inschr
11. klachten, mictie, knie, nek, gehoor, thoracale, schouder, psychische, vaginale, depressieve

12. conjunctivitis, bacteriele, allergische, virale, infectieuze, bacterile, reactie, ods, bact, bdz
13. otitis_externa, bdz, links, rechts, ads, lichte, ome, recidiverende, otitis, oor
14. uwi, geen, recidiverende, rec, hematurie, prostatitis, bewezen, niet, jaar, mictieklachten
15. blwi, virale, viraal, luchtweginfectie, infectie, hyperreactiviteit, lwi, infect, bovenste, obv
16. hoofdpijn, migraine, eci, nek, spanningshoofdpijn, moe, nekklachten, val, duizeligheid, tgv
17. wratten, voet, hand, handen, voeten, vingers, seb, voetzool, vinger, beide
18. symptomen_klachten, schouder, nek, keel, pols, heup, enkel, arm, bursitis, elleboog
19. keelpijn, viraal, virale, oorpijn, verkouden, infect, pharyngitis, griep, pfeiffer, heesheid
20. koorts, eci, viraal, infect, griep, obv, lwi, braken, vaccinatie, zonder
21. moeheid, ziektegevoel, zwakte, eci, malaise, surmenage, wrs, duizeligheid, tgv, anemie
22. overmatig_cerumen, bdz, rechts, links, oor, otitis, ads, ome, cerumen_overmatig, oorsuizen
23. contusie, voet, knie, hand, fractuur, pols, dig, enkel, links, distorsie
24. diarree, braken, infectieuze, overgeven, chronische, spugen, bijwerking, misselijk, bloed, obv
25. bronchitis, chronische, astma, lichte, hyperreactiviteit, chron, copd, acute_bronchitis_bronchiolitis, griep, recidiverende
26. pneumonie, links, rechts, influenza, griep, beginnende, verdenking, atypische, cave, opname
27. tonsillitis, acute, otitis_media, pharyngitis, lumbago, chronische, peritonsillair, abces, myringitis, acuta
28. hypertensie, orgaanbeschadiging, zonder, essentiele, essentile, cvrm, hypercholesterolemie, diabetes_mellitus_type, diabetes_mellitus, secundaire
29. oma, bdz, links, rechts, ome, ads, loopoor, oorpijn, beginnende, knieklachten
30. naevus, moedervlek, rug, dermale, naevocellularis, wang, buik, atypische, hals, benigne
31. archief, brieven, a62, kennismaking, oud, vorige, gegevens, inschrijfformulier, medisch, inschr