University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Adam Gabriel Dobrakowski
Student no. 359226

Clustering of Medical Free-Text Records Based on Word Embeddings

Master's thesis
in MATHEMATICS

Supervisor:
dr hab. inż. Przemysław Biecek
Institute of Applied Mathematics and Mechanics
University of Warsaw

June 2019

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfils the requirements for the degree of Master of Computer Science.

Date Supervisor's signature

Author's statement

Hereby I declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Author's signature

Abstract

Is it true that patients with similar conditions get similar diagnoses? This thesis presents natural language processing methods applied to a corpus of medical records to validate the positive answer. The presented methods are designed to (1) generate and validate concept embeddings, i.e. vector representations of medical concepts, (2) find representation and clustering of medical visits based on concepts extracted from free-text descriptions recorded by doctors. The methods are validated on a real-world dataset with 100,000 medical visits. Based on the proposed methods we obtained stable and separated clusters of visits for several doctors' specialties. The clusters were positively validated against a final medical diagnosis. Finally, this thesis shows how the proposed algorithm may be used to aid doctors during their practice.

Keywords

Electronic Health Records, natural language processing, text clustering, hierarchical clustering, word embeddings

Thesis domain (Socrates-Erasmus subject area codes)

11.2 Statystyka
11.4 Sztuczna inteligencja

Subject classification

62 Statistics
62-07 Data analysis
62P10 Applications to biology and medical sciences

Tytuł pracy w języku polskim

Klastrowanie opisów wizyt medycznych na podstawie zanurzeń wyrazów

Contents

Introduction

This chapter presents the importance of the main issue of the work, namely the medical records clustering. Then we show some problems and the motivation for the work. Finally, we outline the content of this document.

Motivation

A growing amount of medical data collected in Electronic Health Records (EHR) opened a possibility to develop computer-supported medicine based on information extracted from the data. A clustering of the records plays an especially important role because knowledge about identified groups of visits/patients can be utilized in many ways in dealing with new visits. One of the most well-known examples of clustering are Diagnosis Related Groups (Fetter et al., 1980) which aim to divide patients into groups with similar costs of treatment.

A clustering of visits can be used when it comes to diagnosis of a new patient, because (1) we can follow recommendations that were applied to patients with similar visits in the past to create a list of possible diagnoses, (2) reveal that the initial diagnosis is unusual, (3) identify subsets of visits with the same diagnosis but different symptoms.

The EHRs collected from medical centers can contain a lot of structured contents like age and sex of the patient, place, history of diseases, ICD-10 code, etc. An example of patients' clustering based only on their history of diseases is Ruffini et al. (2017). However, a description of an interview with a patient, a medical examination and recommendations given by doctors is stored often in an unstructured way as free text, hard to process but rich in important information. Free-text descriptions are hard to process due to the heterogeneous, noisy and incomplete nature of such data. Some attempts of processing the medical notes exist for English, while for other languages the problem is still challenging (Orosz et al., 2013).

Lately, linguists from the Polish Academy of Sciences created a semi-automated methodology of extracting medical concepts from Polish free-text medical records and evaluated it on a corpus of about 3 million visits from health-care centers in Poland. Their work resulted in a clean, structured form of doctors' descriptions, more convenient to process by automatic algorithms.

This thesis is based on the created data set and aims to cluster visits based on the descriptions of interviews and examinations of patients. We propose a new methodology, which can be applied to new visits with known descriptions of the interview or examination. We applied this methodology to the real data set and show proprieties of groups of visits derived from the free-text descriptions.

Content of the work

The main aim of this thesis was to create a methodology for the clustering of medical records and show its usefulness in obtaining recommendations.

In the work we present the intermediate steps to achieve this goal, which are:

- Research in the field of existing methods of visits' clustering and application of these methods on our data set (Chapter 1).
- Analysis of the considered real data set of about 100,000 visits (Chapter 3).

The main result of the thesis – an algorithm of visits' clustering – is described in Chapter 2.

During experiments with the proposed methodology we observed valuable partial results which can be used and developed in other fields of medical data mining, not only in the clustering of visits. These results are:

- good-quality vector representations of medical concepts in Polish,
- analogies between pairs of concepts,
- ICD-10 codes representation.

These results are presented in Chapter 4.

Chapter 5, the most extensive, provides an application of the method on the considered corpus. We show that it allows us to obtain better results than the presented baseline.

In the Summary, we recapitulate the main achievements of the thesis and present conclusions, suggestions of a future work and propositions of application of the results.

Acknowledgments

I would like to thank my supervisor, Przemysław Biecek PhD, for his taking care of the whole project, proposing valuable ideas and discussions.

I am also thanking linguists Małgorzata Marciniak PhD, Agnieszka Mykowiecka PhD and Wojciech Jaworski PhD for their help.

The research was financially supported by the Polish Centre for Research and Development (Grant POIR.01.01.01-00-0328/17).

Chapter 1

State of the art

In this chapter we will find a general overview of existing approaches to (1) vector representations of texts and (2) medical concepts' representations and visits' clustering, in Section 1.1. Then, Section 1.2 presents an algorithm of clustering based only on the history of diseases of patients. We will see the results of applying this algorithm to our data set. Section 1.3 presents a simple idea of using of the extracted concepts and occurring problems. Finally, in Section 1.4, we will meet an idea of creating word embeddings and find a detailed description of GloVe algorithm.

1.1. Related works

In this section we can find a short overview of the current status of natural language processing methods of text representation. Then we will focus especially on the domain of medical texts: the problem of the representation of medical texts and the clustering of visits.

1.1.1. Vector representation of texts

In general, there are two main approaches to generate vector representations of a text. The first takes into account the occurrence and the frequency of words in the considered text. The simplest example is one-hot encoding or weighting by Term Frequency-Inverse Document Frequency (TF-IDF, Salton and Buckley (1988), see Subsection 1.3.1). The main disadvantage of these methods is that they do not take into account the semantic similarity between words. Thus very similar texts that do not have common words but do have some synonyms can be represented by totally different vectors. This is an especially serious problem in creating short text representations, like medical descriptions, where two random visits very often do not have any common word.

The second approach takes into account the similarity between words. Examples of such techniques are latent semantic analysis (LSA, Deerwester et al., 1990) or semantic hashing (Salakhutdinov and Hinton, 2009). During the last few years the most successful techniques became words/concepts embeddings, which are vector representations computed during training some kind of a neural network. For a more detailed description of embedding methods, see Section 1.4. Embeddings for texts usually are generated based on word/concept embeddings. Some authors use pretrained word embeddings (i.e. embeddings from publicly available embeddings bases computed on a large data set containing billions of words, like Wikipedia), especially when their data set is too small to train their own embeddings. Others try to modify existing embeddings and adjust to the specific set. But the biggest drawback of

these approaches is that the corpus for training embeddings can be not related to the specific task where embeddings are utilized. These disadvantages have been lastly overcome by the state-of-the-art contextual representations generated by ELMo tool (Embeddings from Language Models, Peters et al., 2018) and BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2018).

In this thesis we use a classic, non-contextual GloVe (Global Vectors) algorithm (Pennington et al., 2014) for generating embeddings. The results are very promising. However, in the future the methodology can be simply expanded by newer embedding algorithms.

If we have word/concept embeddings, the simplest way to generate text embeddings is to use some kind of an aggregation of term embeddings such as an average. This approach was tested for example by Banea et al. (2014) and Choi et al. (2016b). De Boom et al. (2016) compute a weighted mean of term embeddings by the construction of a loss function and training weights by the gradient descent method.

In this work we will see the results of the two approaches to obtaining vector representations: weighted one-hot encoding and embeddings. The new proposed method of obtaining vector representations of visits is based on the second approach. We will see that this new method resulted in better results.

1.1.2. Medical concepts' representation and visits' clustering

The problem of processing medical descriptions is well-studied for English. Medical concepts to be extracted from texts very often are taken from Unified Medical Language System (UMLS, Bodenreider, 2004), which is a commonly accepted base of biomedical terminology. Representations of medical concepts are computed based on various medical texts, like medical journals, books, etc. (Minarro-Giménez et al., 2014; De Vine et al., 2014; Newman-Griffis et al., 2017; Choi et al., 2016c; Chiu et al., 2016) or based directly on data from EHRs (Choi et al., 2016a,b,c).

However, the problem of creating such a base for Polish is still open. This work is based on state-of-the-art propositions of generating Polish medical terminology and a new corpus of medical descriptions (see Section 2.1).

An interesting algorithm for patient clustering is given by Choi et al. (2016a). A subset of medical concepts (e.g. diagnosis, medication, procedure codes) and computed embeddings is aggregated for all visits of a patient. This way we get patient embedding that summaries the medical history of a patient.

In this work we present a different approach. Our data contains medical records for different medical domains. This allows us to create a more comprehensive description of a patient. A second difference is that the our aim is a clustering of visits, not patients. This way a single patient may belong to several clusters.

1.2. Tensor decomposition algorithm

Ruffini et al. (2017) proposed an algorithm of patient clustering based only on the history of diseases of patients. This history is encoded as a vector of ICD-10 codes of diseases. Here we see how the algorithm works and then summarize the results obtained on our data set. We will see that the information encoded in ICD-10 codes is not sufficient and that it is important to extract the information directly from the descriptions of the interview and examination.

1.2.1. Methodology

The algorithm represents visits as high-dimensional binary vectors. Hence, it firstly builds a dummy matrix representation with zero-one elements, X of dimension $N \times d$, where N is the number of patients considered in the clustering and d – the number of diseases. In the matrix X one on the position (i, j) indicates that the patient i has diagnosed the disease j in his history. It is worth noticing that this encoding omits the order of the diseases and their repetitions.

We assume the number of desired clusters, k. Then the vector of diseases of the patient i, $X^{(i)} = (x_1^{(i)}, ..., x_d^{(i)})$, is a sample of the mixture k independent d-dimensional Bernoulli distributions. Let $Y \in \{1, ..., k\}$ be non-observable variable indicating the component of the mixture of $X^{(i)}$. Let $\omega_j := \mathbb{P}(Y = j)$ and $\mu_{ij} := \mathbb{P}(x_i|Y = j)$. We assume that μ_{ij} are independent for each i and j. If we know the parameters ω_j and μ_{ij} then the cluster of the patient i is simply $argmax_{j=1,...,k}\mathbb{P}(Y = j|X^{(i)})$, where, in this case:

$$\mathbb{P}(Y = j|X) \propto \omega_j \prod_{i=1}^{d} \mu_{ij}^{x_i}(1 - \mu_{ij})^{1-x_i}. \tag{1.1}$$

We estimate the parameters ω_j and μ_{ij} based on the first three moments of the distribution. Namely, it is enough to estimate from the data the following three operators:

$$M_1 := \sum_{i=1}^{k} \omega_i \mu_i,\ M_2 := \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i,\ M_3 := \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i, \tag{1.2}$$

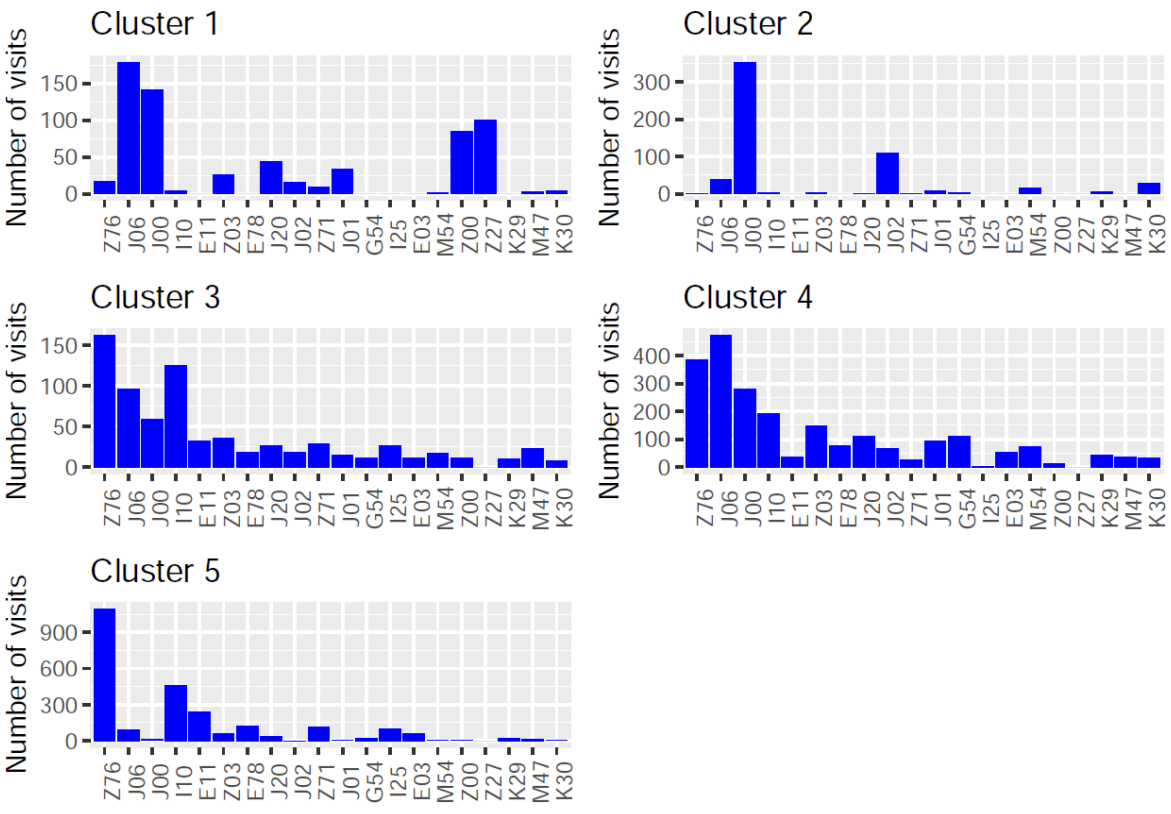
where $M_1 \in \mathbb{R}^d$, $M_2 \in \mathbb{R}^{d \times d}$, $M_3 \in \mathbb{R}^{d \times d \times d}$, $\mu_i := (\mu_{1i}, ..., \mu_{di})$ and $\otimes$ denotes the Kronecker product. Then, from the equations 1.2 we can compute μ_{ij} and ω_j by a tensor decomposition algorithm. Authors in Ruffini et al. (2017) showed an algorithm where the direct computation of M_3 is not necessary (for d equals to several hundreds it would require a huge memory costs). The estimators of μ_{ij} and ω_j are finally improved using Expectation Maximization (EM, Dempster et al., 1977).

1.2.2. Results

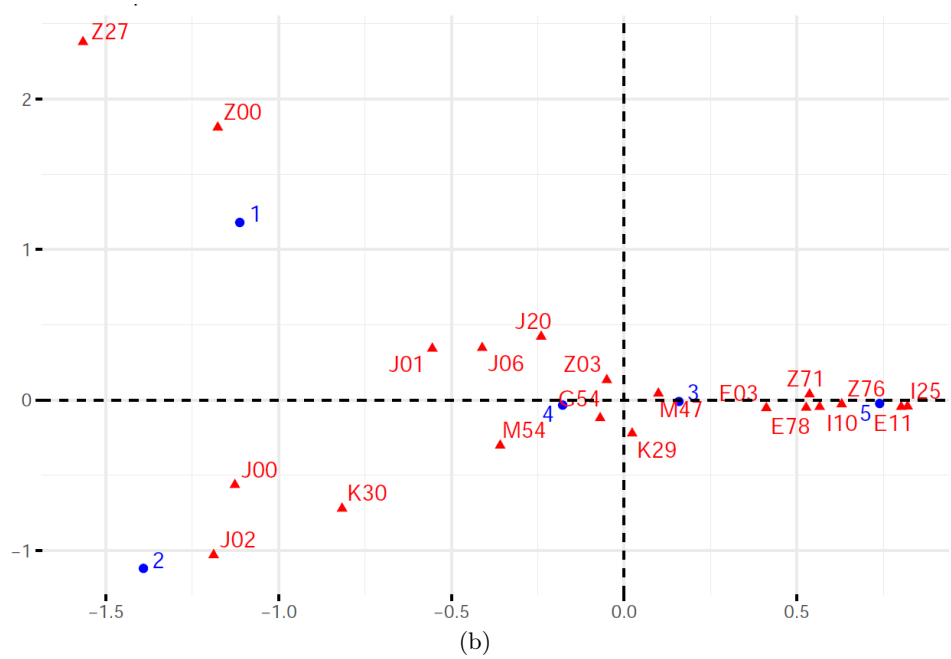
We perform the clustering for nine doctors' specialties separately (every clustering in this work is performed on the same data set, for the details see Chapter 3). In order to make a comparison, we choose the number of clusters as in Chapter 5. We choose only such visits that a patient has at least three diseases in his history (including the current diagnosed ICD-10 code) – the same requirement was set in Ruffini et al.). In contrast to Ruffini et al. we perform the clustering on visits, not patients, so one patient could belong to many clusters. In every case we apply 5 iterations of EM.

On Figure 1.1a we can see the distribution of the diagnosed ICD-10 codes in clusters for family medicine clustering. We selected the 20 most popular codes for family medicine and then count the number of visits assigned to these codes in every cluster.

Such distributions can be conveniently visualized by the correspondence analysis (CA, Hirschfeld, 1935). By this technique we can display clusters and ICD-10 codes in two-dimensional graphical form in such a way that every cluster and every ICD-10 code has an adequate point on the plane. The closer points, the more similar objects. More precisely, ICD-10 codes popular in one cluster are close to the point representing this cluster on the graph. In such a way we can easily see characteristic codes for clusters. We will often use the



(a)



(b)

Figure 1.1: The distribution of ICD-10 codes in clustering of family medicine (a), and the correspondence analysis between codes and clusters (b). We can see the characteristic codes for clusters.

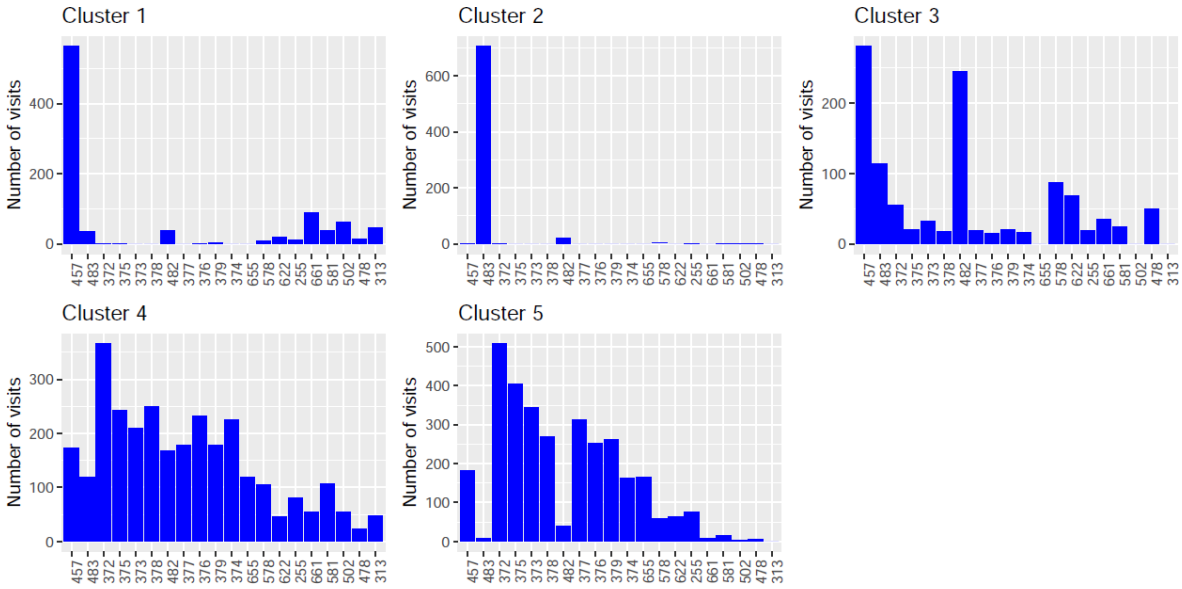


Figure 1.2: The distribution of doctors' IDs in clustering of family medicine. Based only on the history of ICD-10 codes we fitted two clusters to one doctors.

correspondence analysis in the subsequent parts of this work to show relationships between clusters and diagnoses or doctors.

Here, in Figure 1.1b, we can see for example that codes Z27 and Z00 are closely related to Cluster 1 (the point representing Cluster 1 is close to Z00 and Z27). From the histograms in Figure 1.1a we can read that indeed Z27 and Z00 codes occur almost only in Cluster 1. Similarly, codes J00 and J02 are more specific to Cluster 2 than to other clusters. Clusters 3 and 5 are closer to such diseases as I10, I25, E11, E78, Z71, Z76. So, we can see that distributions of ICD-10 in the obtained clusters are very different, as we expected.

However, there occurred some disadvantages of this algorithm. The first is that the number of obtained clusters sometimes is different than assumed (some clusters are empty). For example, in the clustering of family medicine, we assumed 6 clusters and obtained 5.

The second problem is that we can consider only patients that have at least three different ICD-10 codes in their history; it considerably reduced the number of visits in the data set (see Table 3.1) and, moreover, made this method intractable in clustering of new patients (without a history of ICD-10 codes).

Figure 1.2 illustrates the third problem: some clusters are perfectly fitted to one doctor. It can be caused by the fact that ICD-10 codes are simplified diagnoses and hard to express every specific situation. In such ambiguous cases doctors can have their preferred codes. Hence, to make a more reliable representation of a visit, we should make use of the descriptions of interviews and examinations. In such a way we will able to omit ICD-10 codes and then to validate the adequacy of ICD-10 assigning by doctors.

1.3. Description-based algorithms

Now we present the first approach to the clustering of visits based on the descriptions. We will check if indeed this approach allows us to overcome the disadvantages of the ICD-10-based method.

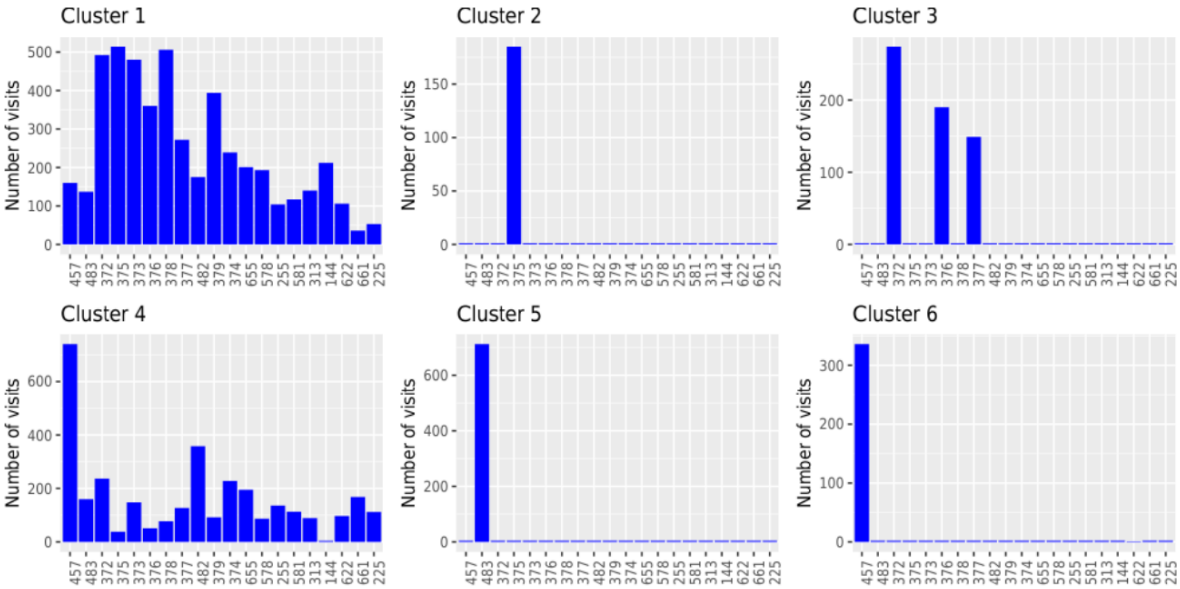


Figure 1.3: The distribution of doctors' IDs in clustering of family medicine. Three clusters are perfectly fitted to one doctor.

1.3.1. One-hot encoding approach

If we have concepts derived from descriptions, the simplest way of achieving a vector representation of the description is one-hot encoding, i.e. creating a dictionary of all terms and representing the description as a binary vector of length equal to the dictionary size.

A little bit more advanced approach is weighting entries of vectors by TF-IDF (term frequency–inverse document frequency, Salton and Buckley, 1988). By these coefficients we want to give higher weights to more rare terms.

For the term i and the document j:

$$(tf\text{-}idf)_{i,j} = tf_{i,j} \times idf_i, \tag{1.3}$$

where $tf_{i,j}$ (*term frequency*) is calculated by the formula:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{1.4}$$

where $n_{i,j}$ is the number of occurrences of the term i in the document j and the denominator is the sum of occurrences of all terms in the document j.

The coefficient idf_j (*inverse document frequency*) is given by the formula:

$$idf_i = log\frac{|D|}{|\{d, t_i \in d\}|}, \tag{1.5}$$

where $|D|$ is the number of all documents in the corpus and the denominator is the number of documents containing at least one occurrence of the term i.

Here, for weighting terms in our data set, we omit TF coefficient, because preprocessed text records do not have multiplied terms.

1.3.2. Results

In the application we firstly obtain one-hot representations of the descriptions weighted by IDF. The representation of the whole visit is the concatenation of the interview vector and the

examination vector. Then we cluster these vectors by the hierarchical clustering method with Ward's method for merging clusters (Ward Jr, 1963). We choose Ward's method (instead of single-linkage or complete-linkage) because here we obtain the most balanced clusters. We use the Euclidean metric for calculating the distance matrix.

Like before, we perform the clustering for each specialty of doctors separately with the same numbers of assumed clusters.

In this algorithm, like in the previous, a lot of the obtained clusters are perfectly fitted to one doctor. Figure 1.3 shows the distribution of doctors among the obtained clusters in family medicine clustering. Three out of six clusters contain visits described by one doctor. We can ask why we get a similar problem like in clustering by the history of ICD-10 codes. Here we do not look at ICD-10 codes. The answer is that a vocabulary and schemes of descriptions are also very often specific for doctors. If we do not take into account the similarity between terms, we build our clusters based on the characteristic expressions for doctors not necessarily semantics.

Another disadvantage of this method is a very high computational cost, because the vector representation of visits has a length equal to the sum of interview dictionary size and examination dictionary size, which is equal to above 7,000 (detailed statistics of the data set are presented in Chapter 3).

1.4. Word embeddings

The first method revealed that it is not enough to cluster visits based only on the history of ICD-10 codes assigned by doctors. The second revealed that a naive usage of concepts extracted from the descriptions is biased by terms or phrases preferred by one doctor. In both methods clusters very often cover visits from only one doctor.

In this section we meet an idea which is a base of the designed methodology, namely word embeddings. This approach enables us to catch the semantic similarity between terms and overcome the problem of doctor-biased terms. Then we will be able to get representations of visits that are more independent of the vocabulary used by the doctors.

Embedding algorithms aim to find vectors for the words in such a way that similar words have vectors close to each other. Three most common classic non-contextual approaches to obtain word embeddings are skip-gram, Continuous Bag of Words (two algorithms from Mikolov et al., 2013) and GloVe (Pennington et al., 2014, where higher accuracy than in previous algorithms was proved).

All these methods are unsupervised algorithms and are based on the observation that two words are similar if they occur in similar contexts. It enables us to find the vector for each word in such a way, that the closeness of the words (in the sense of a small angle between them or a scalar product near to 1) corresponds to close semantic of the words.

1.4.1. GloVe algorithm

GloVe algorithm (*Global Vectors*, Pennington et al., 2014) combines the advantages of two major model families for learning word vectors: 1) global matrix factorization methods, such as latent semantic analysis (LSA, Deerwester et al., 1990) and 2) local context window methods, such as the skip-gram model (Mikolov et al., 2013).

In the algorithm after creating the vocabulary, we compute the term co-occurrence matrix X of dimension $v \times v$, where v is the number of words in the dictionary, where each element X_{ij} indicates how many times the word i occurs in the context of the word j. We build the matrix X_{ij} based on setting the maximal distance between co-occurring words and then

computing the weighted sum (if the word i is in the distance d words from the word j then the contribution to X_{ij} is equal to $1/d$). We can choose other weights or the different left and right context size.

The goal is to obtain for each word two vectors w_i and $\tilde{w}_i$ from $\mathbb{R}^n$ (n is given in the beginning), to minimize the cost function:

$$J = \sum_{i,j=1}^{v} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij}))^2, \tag{1.6}$$

where b_i, $\tilde{b}_j \in \mathbb{R}$, and function f is given by the formula:

$$f(x) = \begin{cases} (x/x_{max})^\alpha & if\ x < x_{max}, \\ 1 & otherwise. \end{cases}$$

The function f is designed in such a way to reduce problems with the indeterminacy of the logarithm where $X_{ij} = 0$ and to reduce the impact of very often co-occurring words (more often than x_{max}). The parameter α was empirically chosen as 3/4 (see Pennington et al., 2014). The vectors are initialized randomly and trained by AdaGrad with an initial learning rate of 0.05.

The result of the algorithm for the word i is the vector $w_i + \tilde{w}_i$ (vectors w_i and $\tilde{w}_i$ are equal if the matrix X is symmetric; if it is not, the separate training of two different vectors allows for obtaining better results).

Chapter 2

Methodology

In this chapter we will find the description of the new algorithm for visits clustering. The clustering is derived in four steps.

(1) Medical concepts are extracted from free-text descriptions of an interview and examination (Section 2.1). (2) A new representation of identified concepts is derived with concepts embedding (Section 2.2). (3) Concept embeddings are transformed into visit embeddings (Section 2.3). (4) Clustering is performed on visit embeddings (Section 2.4).

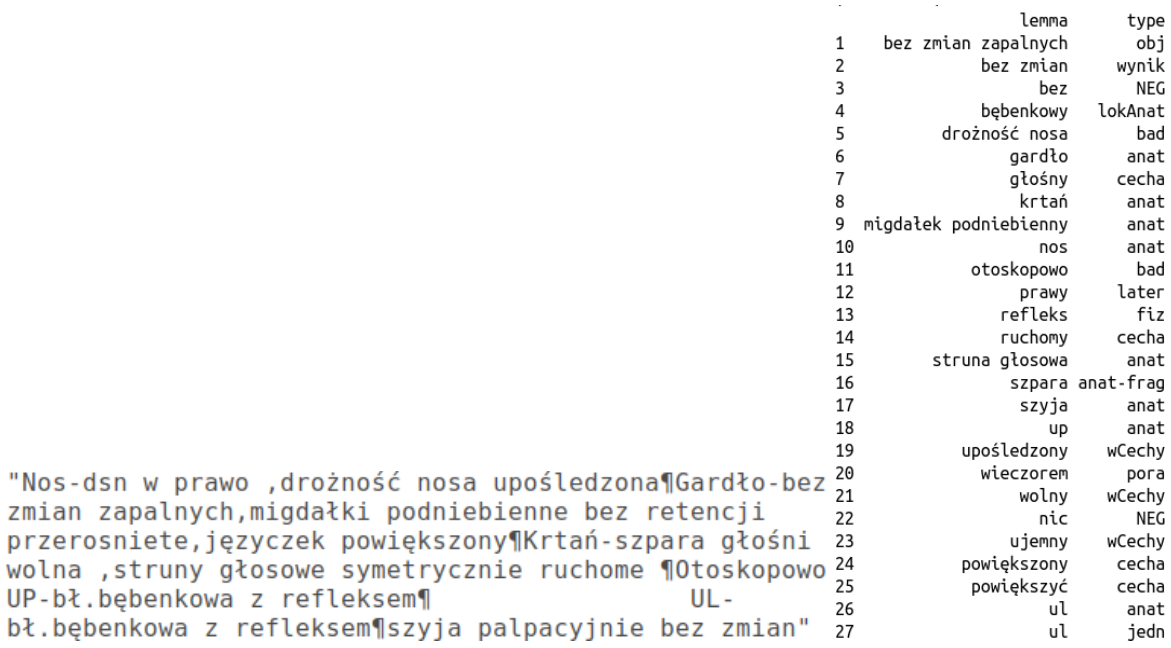
2.1. Extraction of medical concepts

This section presents in details the process of extracting medical concepts from raw free-text medical descriptions. This step was performed by linguists from the Institute of Computer Science, Polish Academy of Sciences. In brief, in the original text we find phrases related to the same semantic concepts. Then we link each concept with a proper semantic label. Figure 2.1 shows an example of original and preprocessed text.

Firstly, we anonymize the data by removing all structural parts with identification data and possible person names mentioned in the text. As there are no generally available terminological resources for Polish medical texts, the first step of data processing is aimed at automatic identification of the most frequently used words and phrases. These concepts will be base for computing embeddings. The doctors' notes are usually rather short and concise, so we assume that all frequently appearing phrases are domain related and important for text understanding. The notes are built mostly from noun phrases so we decide to extract simple noun phrases which consist of a noun optionally modified by a sequence of adjectives (in Polish they can occur both before and after a noun) or by another noun in genitive. We extract only sequences that can be interpreted as phrases in Polish, i.e. nouns and adjectives have to agree in case, number and gender.

Phrase extraction and ordering is performed by TermoPL (Marciniak et al., 2016). The program processes text which is tokenized, lemmatized and tagged with part of speech (POS) and morphological features values. It allows for defining a grammar describing extracted text fragments and order them according to the modified version of the C-value coefficient (Frantzi et al., 2000). Texts are preprocessed using Concraft tagger (Waszczuk, 2012) and we use the standard grammar for describing noun phrases included in TermoPL.

In order to get the most common phrases, we process 220,000 visits by TermoPL. The first 4,800 phrases (with C-value equal at least 20) from the obtained list are manually annotated with semantic labels. The list of labels covered most general concepts like *anatomy, feature, disease, test.* It contained 137 labels. Some labels are assigned to multi-word expressions



"Nos-dsn w prawo ,drożność nosa upośledzona¶Gardło-bez zmian zapalnych,migdałki podniebienne bez retencji przerosniete,języczek powiększony¶Krtań-szpara głośni wolna ,struny głosowe symetrycznie ruchome ¶Otoskopowo UP-bł.bębenkowa z refleksem¶ UL-bł.bębenkowa z refleksem¶szyja palpacyjnie bez zmian"

(a)

	lemma	type
1	bez zmian zapalnych	obj
2	bez zmian	wynik
3	bez	NEG
4	bębenkowy	lokAnat
5	drożność nosa	bad
6	gardło	anat
7	głośny	cecha
8	krtań	anat
9	migdałek podniebienny	anat
10	nos	anat
11	otoskopowo	bad
12	prawy	later
13	refleks	fiz
14	ruchomy	cecha
15	struna głosowa	anat
16	szpara	anat-frag
17	szyja	anat
18	up	anat
19	upośledzony	wCechy
20	wieczorem	pora
21	wolny	wCechy
22	nic	NEG
23	ujemny	wCechy
24	powiększony	cecha
25	powiększyć	cecha
26	ul	anat
27	ul	jedn

(b)

Figure 2.1: An example of an original and preprocessed free-text description of an examination of a patient. The preprocessed text abandons the original order, repetitions of concepts and not-identified tokens (in particular digits).

(MWEs), in some cases all or some of their elements are also labeled separately, e.g. *left hand* is labeled as *anatomy* while *hand* is also labeled as *anatomy* and *left* as *lateralization*. The additional source of information is the list of 9,993 names of medicines and dietary supplements.

The above list of terms together with their semantic labels is then converted to the format of lexical resources of Categorial Syntactic-Semantic Parser „ENIAM" (Jaworski and Kozakoszczak, 2016; Jaworski et al., 2018). The parser recognizes lexemes and MWEs in visits according to the provided list of terms, then the longest sequence of recognized terms is selected, and semantic representation is created. Semantic representation of a visit has a form of a set of pairs composed of recognized terms and their labels (not recognized tokens are omitted). The average coverage of semantic representation was 82.06% of tokens and 75.38% of symbols in section *Interview* and 87.43% of tokens and 79.28% of symbols in section *Examination*.

Texts of visits are heterogeneous as they consist of: very frequent domain phrases; domain important words which are too infrequent to be at the top of the term list prepared by TermoPL; some general words which do not carry relevant information; numerical information; and words which are misspelled.

The clustering task is performed on a subset of about 100,000 preprocessed visits. We neglect the original text and the experiments described in this work are solely performed on the set of extracted concepts from each interview and examination.

2.2. Embeddings for medical concepts

We compute embeddings of concepts by GloVe algorithm (see Subsection 1.4.1) for interview descriptions and for examination descriptions separately. During creating the term co-occurrence matrix (TCM) the whole description is treated as the neighborhood of the concept. All the terms from the description contribute to TCM with the same weights equal to 1.

We compute two separate embeddings due to catch the similarity between terms in their specific context, because terms similar in the interview may not be similar in the examination description (for example we computed that the nearest terms to *cough* in embeddings of interview are *runny nose*, *sore throat*, *fever*, *dry cough* but in embeddings of examination these are *rash*, *sunny*, *laryngeal*, *dry cough*).

2.3. Visit embeddings

To obtain a representation of a visit, we compute averages of embeddings of concepts from interview and examination descriptions, separately. Final embeddings for visits are obtained by concatenation of these two averages, see Figure 2.2. Such a method allows us to generate representations of visits where at least one of the descriptions of the interview and examination is not empty. For the empty part, the missing embedding is filled with zeros.

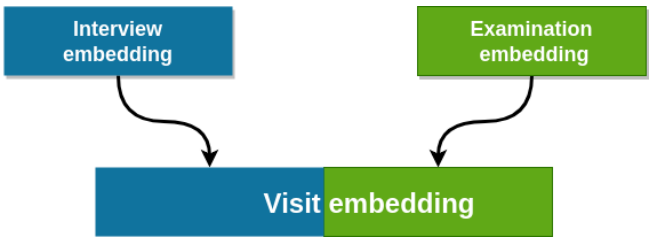


Figure 2.2: The visit embedding is the concatenated vector of the average of concepts embeddings from the interview and average of concepts embeddings from the examination.

Let us notice that the vector representation of a visit is derived only from the description of the interview and examination of a patient (precisely, from the extracted concepts) and does not contain any information about ICD-10 code of disease nor about recommendations given by the doctor. It is especially important, because in Chapter 4 we will see that obtained representations are strongly related to ICD-10 codes and to recommendations.

2.4. Visits clustering

Based on Euclidean distance between vector representations of visits we apply and compare two clustering algorithms: k-means and hierarchical clustering with Ward's method for merging clusters (Ward Jr, 1963). The similarity of these clusterings is measured by the adjusted Rand index (Rand, 1971). The adjusted Rand index measures the similarity of two divisions by a correction an expected value of random similarity.

For the final results we chose the hierarchical clustering algorithm due to greater stability.

Chapter 3

Data set

This chapter presents the structure of the considered data set of medical records. We will see basic statistics related to the number of visits, the number of medical concepts, ICD-10 distributions, and recommendation terms distributions.

3.1. Data structure

The data set consists of about 100,000 patients' visits from different primary health care centers and specialist clinics in Poland. The first part of the data contains descriptions of visits, which have a free-text form. They are written by doctors representing a wide range of medical professions, e.g. general practitioners, dermatologists, cardiologists or psychiatrists. Each description is divided into three parts: interview, examination, and recommendations. The process of extracting medical concepts from free-text is described in Section 2.1.

The second part of the data contains structured information about:

- doctor ID,
- specialty of the doctor,
- ICD-10 code assigned by the doctor, according to commonly used ICD-10 classification – International Statistical Classification of Diseases and Related Health Problems (Organization et al., 2004). The code contains a letter (a type of the disease) and two digits (e.g. E11, F32).
- history of ICD-10 codes from previous visits of the patient,
- sex of the patient,
- age of the patient.

3.2. Number of visits

We should note that a lot of records have some shortages, so the number of clustered visits in the algorithms is considerably lower.

In the experiments we consider 99,973 visits. 80,570 of them contain ICD-10 code – only these visits are considered in the clustering (but we use all descriptions to compute embeddings). 13,770 of them do not contain a specialty of a doctor. The other 66,800 belong to some of 45 specialties (a lot of doctors have more than one specialty, so many

specialty	# visits	# descr.	# ≥ 3 dis.
Family medicine	28,543	11,230	9,566
Internal medicine	18,515	6,419	4,330
Pediatrics	10,067	4,742	3,532
Gynecology	6,771	3,456	535
Orthopedics	2,794	1,869	116
Dermatology and venereology	2,275	1,204	47
Cardiology	1,899	1,201	327
Endocrinology	1,814	1,510	709
Psychiatry	1,693	1,012	71
SUM	74,371	32,643	19,233

Table 3.1: The number of visits per specialty. The second column presents the total number of visits. The third column shows the number of visits with non-empty descriptions. The fourth column presents the number of visits clustered by tensor decomposition algorithm (at least 3 different ICD-10 codes in the history).

visits are considered in several clusterings). In the experiments we cluster 9 the largest specialties, see Table 3.1. We obviously omit visits with empty descriptions of interview and examination (see Section 2.3). Furthermore, because we will validate clusters against prescribed recommendations, we remove visits with an empty recommendation description. The final number of clustered visits is presented in the third column of Table 3.1.

The tensor decomposition algorithm requires at least three ICD-10 codes in the history of the patient (see Section 1.2). The number of visits fulfilling this assumption is presented in the fourth column of Table 3.1.

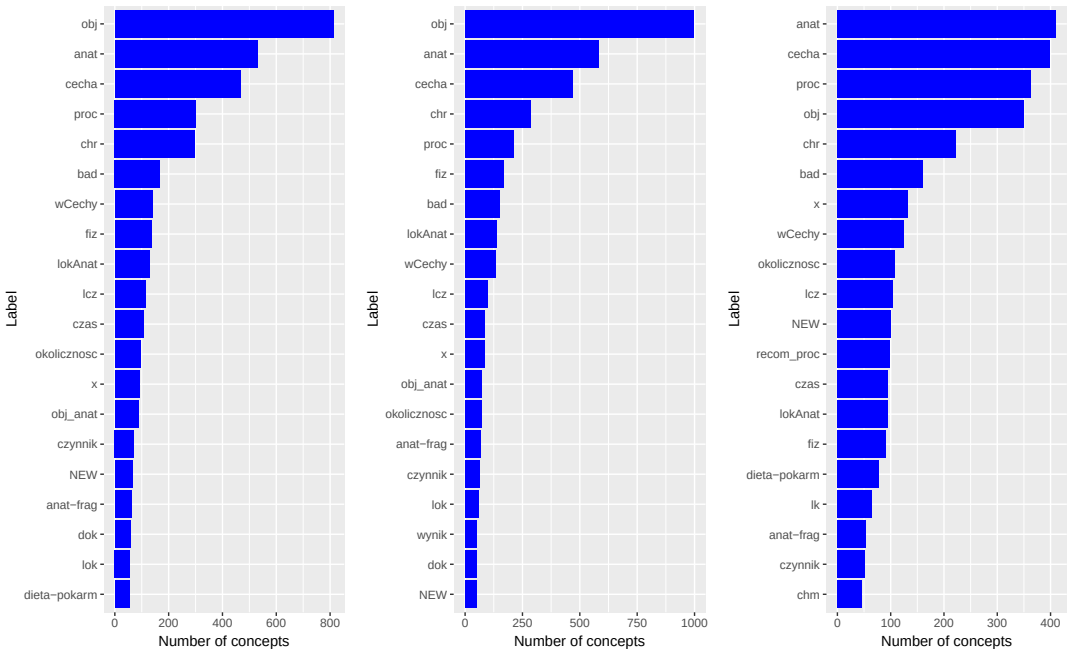
3.3. Medical concepts

The total number of extracted concepts from the descriptions is as follows:

part of description	# extracted con.	# embedded con.	mean # concepts
interview	4,603	3,816	11.2
examination	3,812	3,286	13.5
recommendation	3,282	2,879	7.6

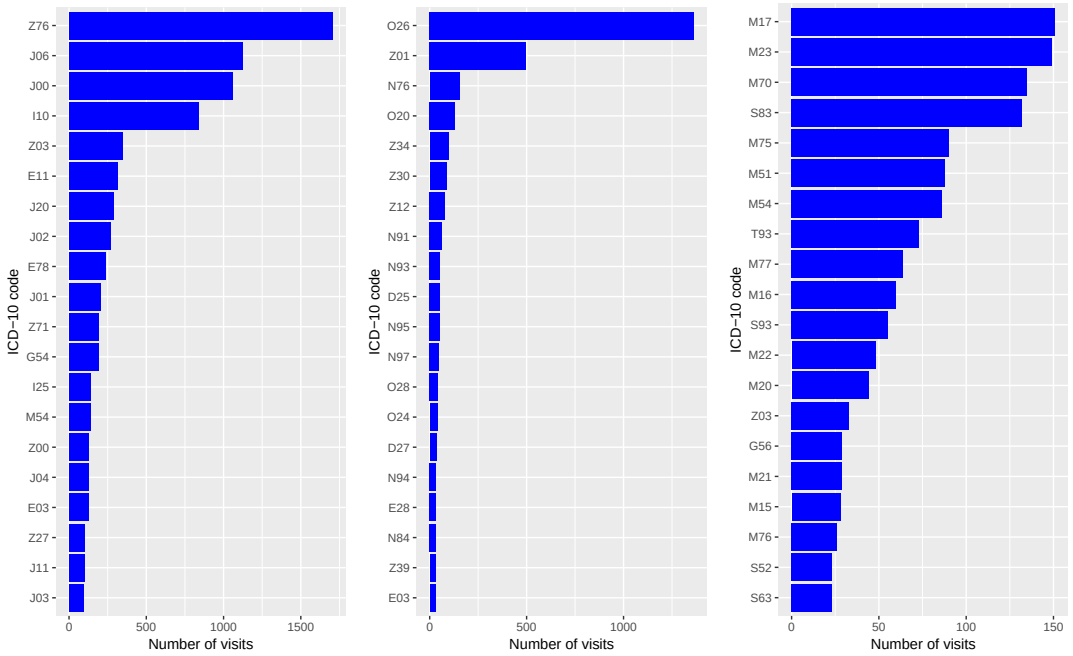
The second column shows the total number of extracted concepts from each description part. The third column shows the number of embedded concepts – the concepts that occurred at least 5 times in all descriptions. There were 2,556 common embedded terms for interview and examination (for these terms we have two different embeddings). The fourth column presents the mean number of embedded concepts per one description (regarding only non-empty descriptions).

As it was described in Subsection 2.2.1, for each concept we assigned a semantic label (for some concepts we assigned several labels). Figure 3.1 presents the 20 most popular labels for each description type and the number of derived concepts with these labels (not every concept has an embedding).



(a) Interview (b) Examination (c) Recommendation

Figure 3.1: The number of concepts with the most popular semantic labels. The most popular labels are: *symptom* (Polish abbreviation: obj), *anatomy* (anat), *feature* (cecha), *procedure* (proc), *disease* (chr).



(a) Family medicine (b) Gynecology (c) Orthopedics

Figure 3.2: Frequency of the 20 most popular ICD-10 codes per specialty.

specialty	most frequent recommendations
Family medicine	kontrola (13.5%), lek (9.1%), konieczna kontrola lekarska (8.5%), kontynuacja leczenia (6.7%), kontynuacja leków (5.8%), zalecenie (5.7%), systematyczne przyjmowanie leków (3.8%), badanie (3.5%), morfologia (3.2%), terapia (3.1%), leczyć (3%), rtg (3%), ssanie (3%), inhalacja (2.7%), kontrola lekarska (2.5%), toaleta noska (2.5%), odpoczynek (2.3%), probiotyk (2.2%), żel (2.2%), dieta (2.2%)
Gynecology	kontrola (40.6%), szpital (10%), morfologia (8.8%), badanie (6.7%), zalecenie (6.6%), tryb życia (6.2%), zus (4.9%), acidum (4.4%), badanie usg (4.1%), badanie ogólne moczu (2.6%), terapia (2.2%), witamina (2.2%), leczyć (2.1%), dieta (2%), cytologia (1.9%), lek (1.9%), potas (1.8%), badanie laboratoryjne (1.7%), najbliższy szpital (1.7%), odpowiednia dieta (1.7%)
Orthopedics	kontrola (25.9%), ćwiczenie (20.1%), rtg (16.2%), rehabilitacja (11.1%), zus (10%), rezonans magnetyczny (7.1%), orteza (6.7%), terapia (6.5%), operacja (5.6%), lek (5.3%), laser (5%), leczyć (4.9%), leczenie operacyjne (4.5%), oszczędzanie (4.5%), fizykoterapia (4.2%), terapia manualna (4.1%), ap (4%), krioterapia (3.6%), badanie usg (3.6%), wkładka (3.2%)
specialty	most frequent recommendations
Family medicine	control (13.5%), medicament (9.1%), necessary medical control (8.5%), continuation of treatment (6.7%), continuation of medicaments (5.8%), recommendation (5.7%), systematic taking of medicines (3.8%), examination (3.5%), morphology (3.2%), therapy (3.1%), to treat (3%), X-ray (3%), suction (3%), inhalation (2.7%), medical control (2.5%), toilet of the (small) nose (2.5%), repose (2.3%), probiotic (2.2%), gel (2.2%), diet (2.2%)
Gynecology	control (40.6%), hospital (10%), morphology (8.8%), examination (6.7%), recommendation (6.6%), style of life (6.2%), *zus* (Social Insurance Institution) (4.9%), acidum (4.4%), ultrasound examination (4.1%), urinalysis (2.6%), therapy (2.2%), vitamin (2.2%), to treat (2.1%), diet (2%), cytology (1.9%), medicament (1.9%), potassium (1.8%), laboratory examination (1.7%), nearest hospital (1.7%), proper diet (1.7%)
Orthopedics	control (25.9%), exercise (20.1%), X-ray (16.2%), rehabilitation (11.1%), *zus* (Social Insurance Institution) (10%), magnetic resonance (7.1%), orthosis (6.7%), therapy (6.5%), operation (5.6%), medicament (5.3%), laser (5%), to treat (4.9%), surgery (4.5%), thrift (4.5%), physiotherapy (4.2%), manual therapy (4.1%), *ap* (unknown abbreviation) (4%), cryotherapy (3.6%), ultrasound examination (3.6%), insertion (3.2%)

Table 3.2: The most frequent recommendations for three specialties in an original and translated form. In brackets we see a percentage of all visits from this specialty that contain a specified term.

3.4. ICD-10 codes

Figure 3.2 presents the most popular ICD-10 codes for three specialties. We should notice that for every specialty there are only a few very frequent codes.

3.5. Recommendation

The recommendation part of the descriptions, preprocessed in the same way as interviews and examinations, consists of the extracted concepts with some labels. However, not every concept is related to a specific procedure that a doctor proposed to a patient. To extract only concepts related to a specific recommendation, we limit us only to some categories of the terms. From all terms we chose five categories: *procedure* (proc), *examination* (bad), *treatment* (lcz), *diet* (dieta-pokarm) and *medicament* (lk). We assumed that terms from these categories will be related to complete recommendations.

Table 3.2 presents the 20 most popular recommendation terms from the chosen five categories. Some terms are still too general (e.g. *to treat*, *examination*, *nearest hospital*). However, the most of terms are related to a concrete action.

Chapter 4

Results of term embeddings

This chapter presents very interesting partial results of the proposed methodology, which can be used and developed in many fields of medical data mining. Section 4.1 shows examples of obtained embeddings of medical concepts. Section 4.2 presents a method of validation embedding quality. We will see astonishingly good results of a proposed new term analogy task. Section 4.3 presents a map of ICD-10 codes that we can treat as a kind of validation of ICD-10 classification.

4.1. Embeddings for medical concepts

Embeddings are computed for terms occurring at least 5 times in the data set (separately for interview and examination descriptions).

Embeddings of terms can be visualized with the T-distributed Stochastic Neighbor Embedding algorithm (t-SNE, Maaten and Hinton, 2008). This dimensionality reduction technique, unlike PCA, is not a linear projection. It uses the local relationships between points to create a low-dimensional mapping. This allows it to capture the non-linear structure. The t-SNE algorithm comprises two main stages. First, t-SNE constructs a probability distribution over pairs of high-dimensional objects (we define the probability that two objects are similar). Second, t-SNE defines a probability distribution (Student t-distribution) over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence between the two distributions (using gradient descent) with respect to the locations of the points in the map.

Here we generate two-dimensional mapping of 20-dimensional concept embeddings. Such visualization can tell us if the embeddings reflect the semantic similarity between terms. Figure 4.1 shows a fragment of visualization of *symptomps* derived from interview descriptions. Figure 4.2 shows a visualization of *anatomic* terms derived from examination descriptions. In the same way we can obtain embeddings of terms related to *medicaments*, *diseases*, *medical examinations*, etc.

Projections of embeddings suggest that embeddings are good and store some semantic information about medical concepts. Below section will verify this claim in a more systematic way.

4.2. Analogies in medical concepts

To evaluate the quality of term embeddings and to determine the optimal dimension of embedded vectors we employ a word analogy task introduced by Mikolov et al. (2013) and examined in details in a medical context by Newman-Griffis et al. (2017).

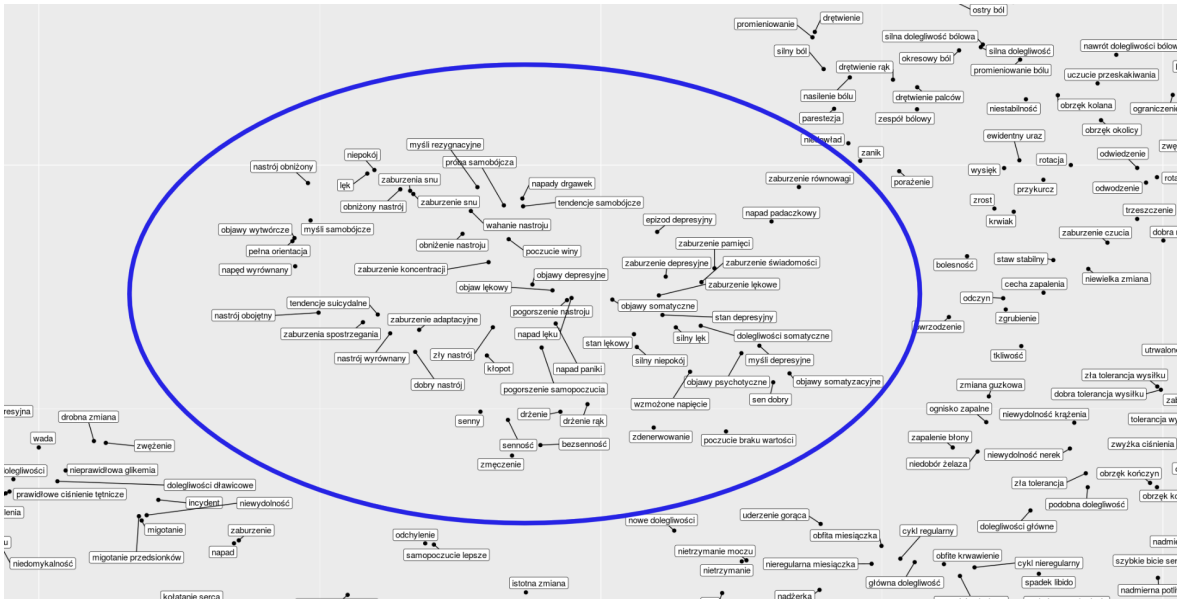


Figure 4.1: A fragment of a t-SNE visualization of embeddings of symptomps from interview descriptions. All terms from the circled group are related to mental disorders.

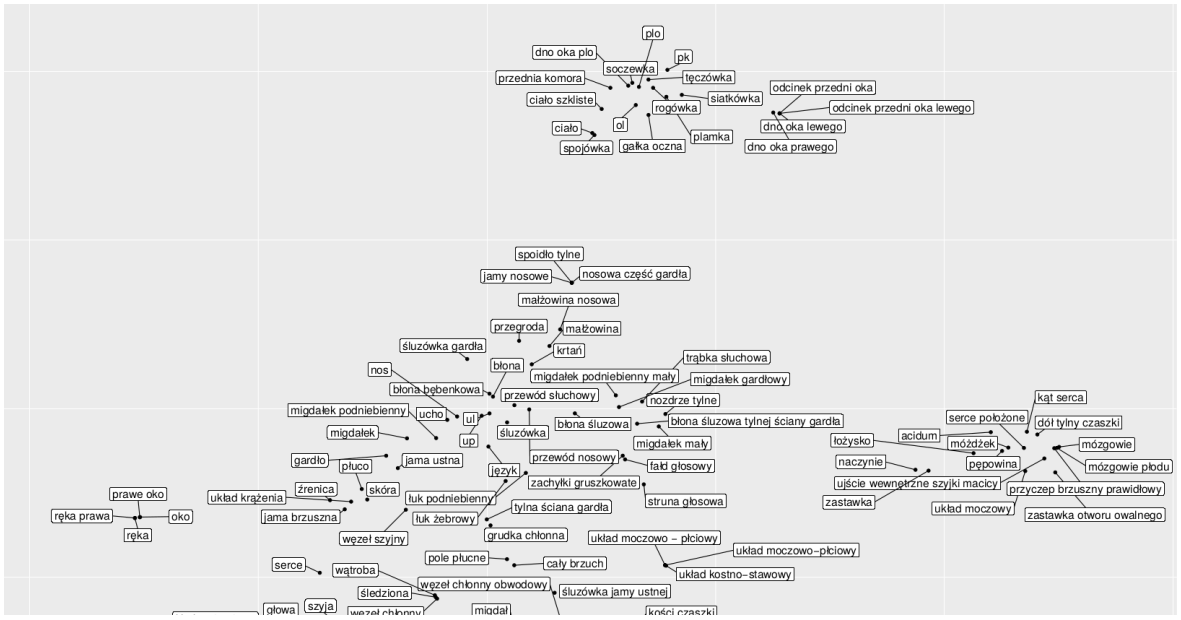


Figure 4.2: A fragment of a t-SNE visualization of embeddings of anatomic terms from examination descriptions. The terms in the upper group are related to parts of eye, like *cornea, conjunctiva* (in Polish *rogówka, spojówka*). On the right we can see the terms from a fetus examination.

Type of relationship	# Pairs	Term Pair 1		Term Pair 2	
Body part – Pain	22	eye	eye pain	foot	foot pain
specialty – Adjective	7	dermatologist	dermatological	neurologist	neurological
Body part – Right side	34	hand	right hand	knee	right knee
Body part – Left side	32	thumb	left thumb	heel	left heel
Spec. – Consultation	11	surgeon	surgical consult.	gynecologist	g. consult.
specialty - Body part	9	cardiologist	heart	oculist	eye
Man - Woman	9	patient (male)	patient (female)	brother	sister

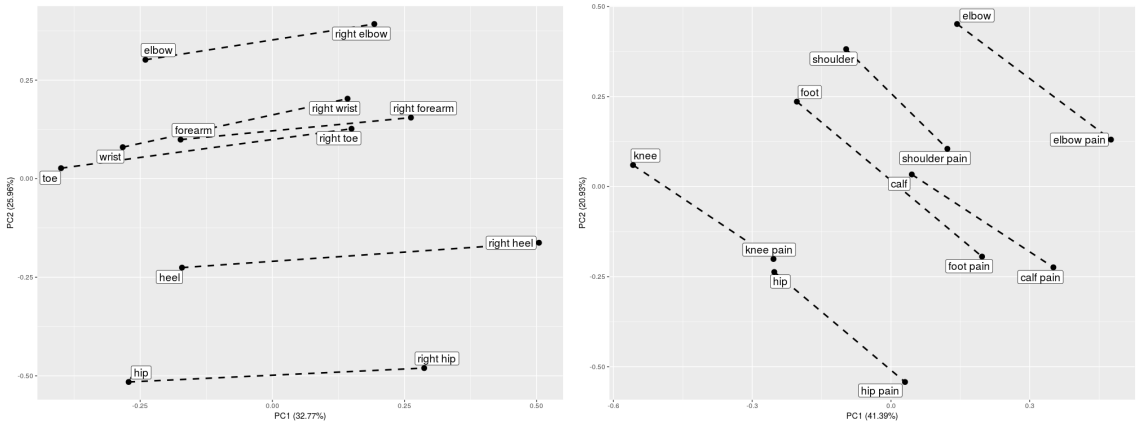
Table 4.1: The categories of questions in the term analogy task with total number of pairs in every category and two examples. We would like to obtain: $vector(eye) - vector(eye\ pain) + vector(foot\ pain) \approx vector(foot)$.

Dim. / Context	1	3	5
10	0.1293	0.2189	0.2827
15	0.1701	0.3081	0.4123
20	**0.1702**	0.3749	0.4662
25	0.1667	0.4120	0.5220
30	0.1674	0.4675	0.5755
40	0.1460	**0.5017**	0.6070
50	0.1518	0.4966	**0.6190**
100	0.0435	0.4231	0.5483
200	0.0261	0.3058	0.4410

Table 4.2: A mean accuracy of correct answers on the term analogy tasks. The rows present different embeddings sizes. Columns 2 - 4 show results by considering different neighborhoods of computed vectors.

The motivation of the word analogy task is that if we have two pairs of words related in the same way, this relation should be reflected in vector representations of these words. For example between words *man* and *woman* there is the same relation as between words *king* and *queen*. Hence, if we have embeddings of these words, we should be able to obtain: vector(*man*) - vector(*woman*) $\approx$ vector(*king*) - vector(*queen*). In their work Mikolov et al. (2013) defined five types of semantic and nine types of syntactic relationship, e.g. Man-Woman, City-in-state, Adjective to adverb, Opposite.

Here, for medical concepts it is necessary to find other relationships, more related to the context of medical descriptions. In designing new analogies we exploit the fact that in the corpus we have a lot of multiword concepts and very often the same words are included in different terms. Therefore there appears some relationship between terms in a natural way. For example, there exist such terms as *foot*, *foot pain*, *left foot*, *right foot*, but also *hand*, *hand pain*, *left hand*, *right hand*, etc. Such relationships allow us to design a new term analogy task, more appropriate for our corpus. We chose seven types of relationships.



(a) Body part – Right side

(b) Body part – Pain

(c) Specialty – Adjective

(d) Specialty – Body part

Figure 4.3: Visualization of analogies between terms. Pictures show term embeddings projected into 2d-plane using PCA. Each panel shows a different type of analogy.

A question in the term analogy task is computing a vector: $vector(left\ foot) - vector(foot) + vector(hand)$ and checking if the correct $vector(left\ hand)$ is in the neighborhood (in the metric of cosine of the angle between the vectors) of this resulting vector. We compute answers' accuracy in a similar way as in Mikolov et al. (2013): we create manually the list of similar term pairs and then form the list of questions by taking all two-element subsets of the pairs list. So if we created N term pairs in one category, we obtained $\binom{N}{2}$ questions. Table 4.1 shows the categories of questions, the number of created pairs and two examples of pairs of terms.

We also created one additional task, according to the observation that sometimes two different terms are related to the same object. This can be caused for example by the different order of words in the terms, e.g. *left wrist* and *wrist left* (in Polish both options are acceptable). In good embeddings such terms should be very close to each other. Hence, we define a list of synonym pairs and check if the embedding of one term lies in the neighborhood of the second.

We evaluate term embeddings for vectors of length from 10 to 200. For every embedding of interview terms we measure an accuracy of every eight tasks. Table 4.2 shows the mean of eight task results. The second column presents the results of the most restrictive rule: the question is assumed to be correctly answered only if the closest term (in the sense of the smallest angle between vectors) to the computed vector is the same as the desired answer. It should be noticed that the total number of terms in the considered data set (about 900,000 for interviews) was many times lower than sets examined in Mikolov et al. (2013). Furthermore, words in medical descriptions very often appear in a very specific context, which can do not give the relationship that was expected. Taking this into account, the accuracy of about 0.17 is very high and better than we expected.

Then we check the closest 3 and 5 words to the computed vector and assume a correct answer if in this neighborhood we can find the correct vector. In the biggest neighborhood the majority of embeddings returned the accuracy higher than 0.5. This indicates that the quality of obtained embeddings of medical concepts is very good and these embeddings are able to reflect the semantic similarity between concepts.

For computing visit embeddings we choose embeddings of dimensionality 20, since this resulted in the best accuracy of the most restrictive analogy task and it allows us to perform more efficient computations than higher dimensional representations.

Analogies between terms can be very clearly visualized by vector projections. Segments connecting pairs of terms should be almost parallel. Figure 4.3 shows that indeed this phenomenon appears in the created embeddings. Embeddings are projected using PCA method.

4.3. ICD-10 codes embeddings

If we have vector representations of visits, we are able to generate embeddings of ICD-10 codes. As we said, ICD-10 is a commonly used classification of diseases, based on expert knowledge (Organization et al., 2004). With exploiting the fact that every visit has a proper ICD-10 code assigned by the doctor, we assume that the representation of ICD-10 code can be computed simply as an average of embeddings of all visits with this code. Figure 4.4 shows t-SNE visualization of these representations. It turns out that indeed there appear very clear groups of codes from the same categories of diseases. Furthermore, we can see that codes from Z group do not form a separate cluster and they are spread on the whole map. It very well corresponds with their meaning, because codes from Z group are *Factors influencing health status and contact with health services.*

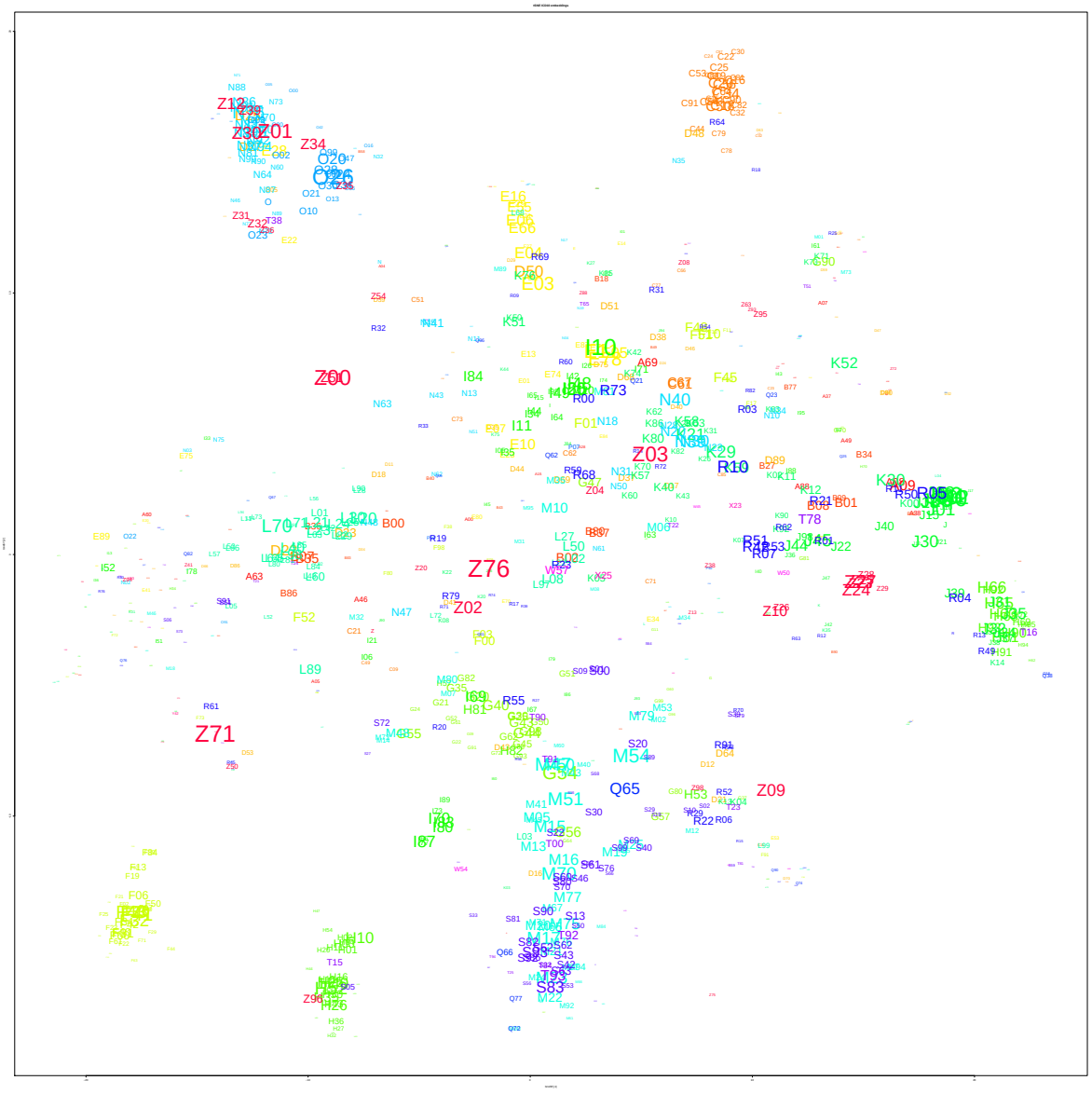


Figure 4.4: Map of ICD-10 codes in the space of embeddings for all visits, colored by the first character of the code. More popular codes have larger labels.

Chapter 5

Results of visits clustering

This chapter presents the details of the implementation of the proposed methodology and an analysis of achieved results. In Section 5.1 we will find a visualization of the clusters and understand their structure and stability. Section 5.2 will show how we measure the quality of the clusters and why the proposed methodology is better than clustering by the weighted one-hot visits' encoding. In Section 5.3 we will see that based on the clusters we can predict the doctor's diagnosis (ICD-10 code). Section 5.4 will develop these results and we will try to predict detailed recommendations that doctors prescribe to patients.

5.1. Visits clustering

specialty	# clusters	# visits	clusters' size	K-means - hclust
Cardiology	6	1201	428, 193, 134, 303, 27, 116	0.87
Dermatology and venereology	6	1204	455, 89, 176, 30, 391, 63	0.64
Endocrinology	5	1510	389, 412, 208, 183, 318	0.8
Family medicine	6	11230	3108, 2353, 601, 4518, 255, 395	0.69
Gynecology	4	3456	1311, 1318, 384, 443	0.8
Internal medicine	5	6419	1915, 1173, 1930, 1146, 255	0.76
Orthopedics	4	1869	360, 1257, 102, 150	0.19
Pediatrics	5	4742	1751, 658, 666, 715, 952	0.46
Psychiatry	5	1012	441, 184, 179, 133, 75	0.81

Table 5.1: The statistics of clusters for selected domains by hierarchical clustering algorithm. The last column shows adjusted Rand index between k-means and hierarchical clustering.

We perform the clustering separately for each specialty of doctors. For determining the optimal number of clusters, for each specialty we consider the number of clusters between 2 and 15. We choose the number of clusters so that adding another cluster does not give a relevant improvement of a sum of differences between elements and clusters' centers (according to the the so-called *Elbow method*, see Figure 5.1).

Table 5.1 gives basic statistics of clusters obtained by hierarchical clustering method. For every specialty we chose between 4 and 6 clusters. The last column contains the adjusted Rand index between hierarchical clustering and k-means. It can be interpreted as a measure

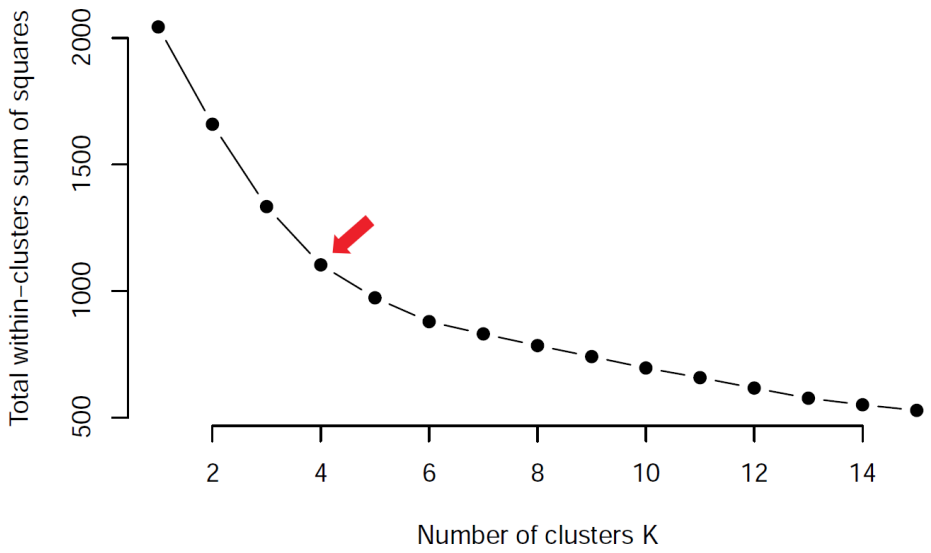


Figure 5.1: *Elbow method*: for each k plot the sum of squares of differences between elements in clusters and clusters' centers. Four clusters is the optimal number, because the higher does not give any relevant improvement.

of the stability of clustering. The higher similarity of two algorithms the higher stability of clustering. Except for orthopedics and pediatrics, the Rand index is very high.

Figure 5.2 illustrates two-dimensional projections (from 40-dimensional space) of visits' embeddings colored by clusters. The projections are created by t-SNE algorithm.

5.2. Comparison with one-hot representation clustering

This section presents three indicators that the proposed methodology is better than one-hot representation clustering: the clusters are more balanced, much more rarely fitted to one doctor and better in the distribution of ICD-10 codes.

5.2.1. Small clusters and doctors' distribution

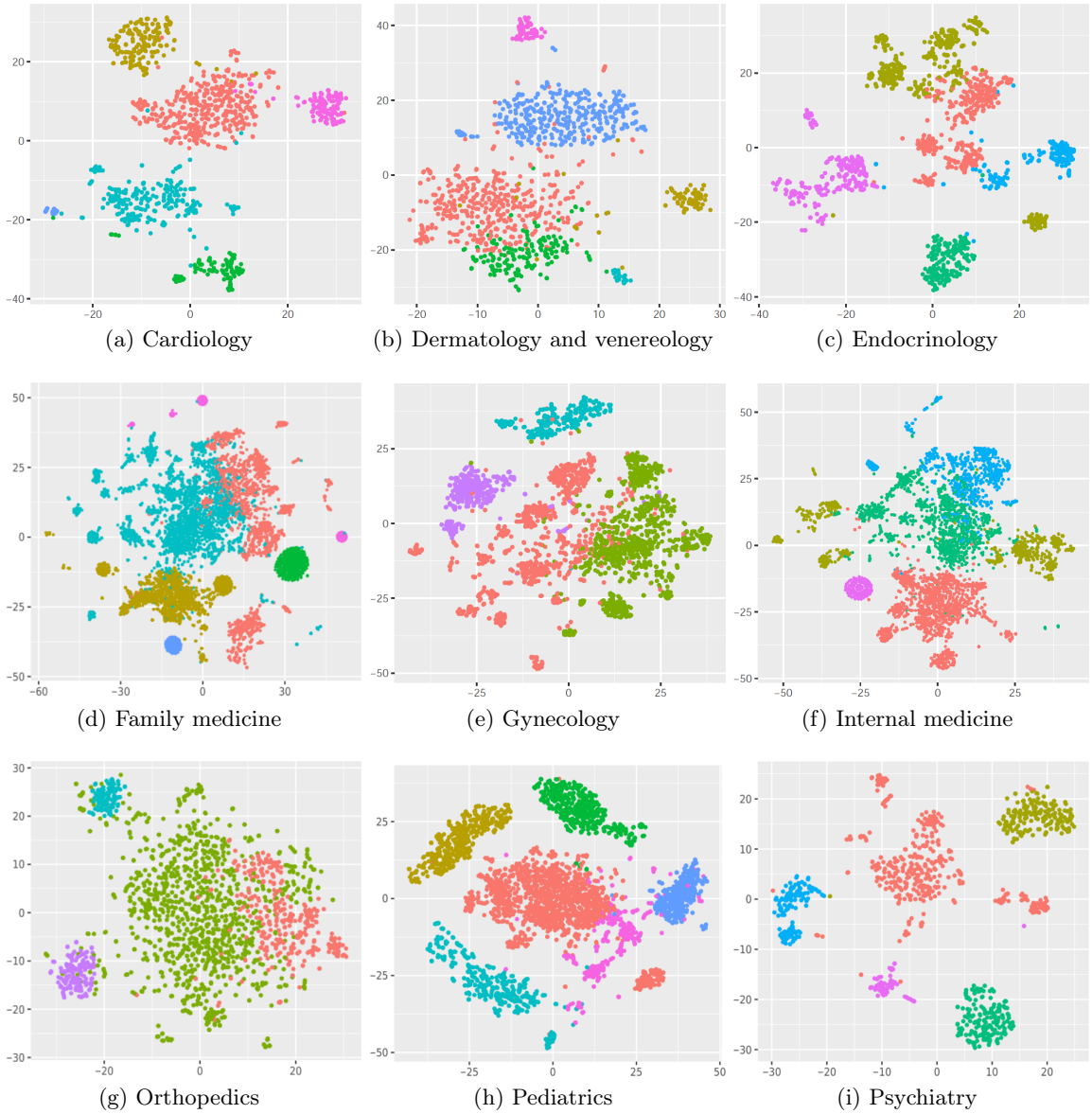
As we saw in Section 1.3, the main issue of clustering visits represented by one-hot encoded vectors was that clusters very often were perfectly fitted to one doctor. The second problem was the unbalanced clusters.

To compare the quality of these two clustering methods, we count the number of obtained small clusters (smaller than 5% of all visits) and the number of clusters fitted to one doctor (where one doctor covers more than 95% visits in this cluster). Table 5.2 shows that the proposed methodology returns much better clusters.

For every clustering of one-hot encoding visits at least two clusters were fitted to one doctor. In embedding representations, the most clusterings return 0 or 1 such clusters.

Figure 5.3 presents the distribution of doctors in family medicine clustering (cf. with Figure 1.3). There is only one cluster fitted to one doctor.

However, there remained three specialties (cardiology, pediatrics, psychiatry) where it failed to overcome doctor fitting. Figure 5.4 presents t-SNE visualizations of visits' embeddings colored by doctors and the correspondence analysis between doctors and clusters. Some doctors use very specific terms or their descriptions are based on characteristic schemes. Such



(a) Cardiology (b) Dermatology and venereology (c) Endocrinology

(d) Family medicine (e) Gynecology (f) Internal medicine

(g) Orthopedics (h) Pediatrics (i) Psychiatry

Figure 5.2: Clusters of visits for selected domains. Each dot corresponds to a single visit. Colors correspond to segments. Visualization created with t-SNE.

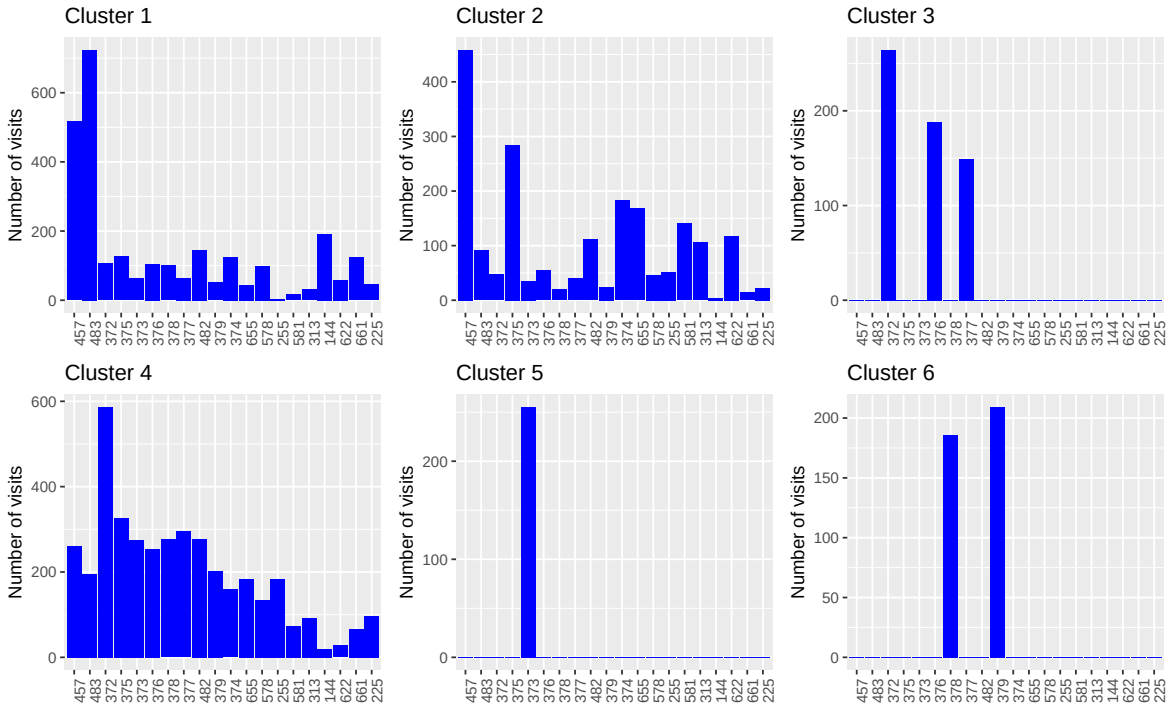


Figure 5.3: The distribution of doctors' IDs in clustering of family medicine by embedding representations.

descriptions have representations very similar to each other and separated from the others. So, our algorithm is not able to link these visits with others.

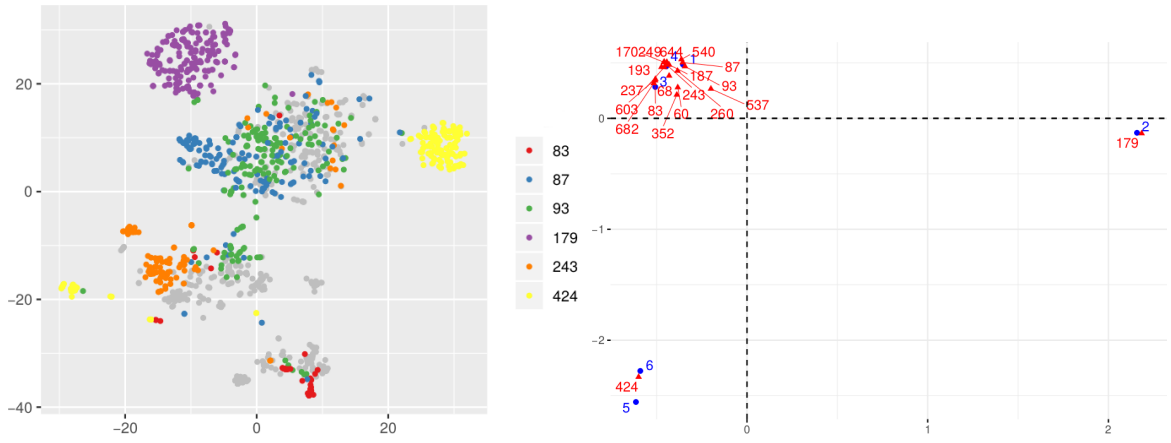
In Appendix A we can find results for the remaining six specialties.

5.2.2. ICD-10 distribution

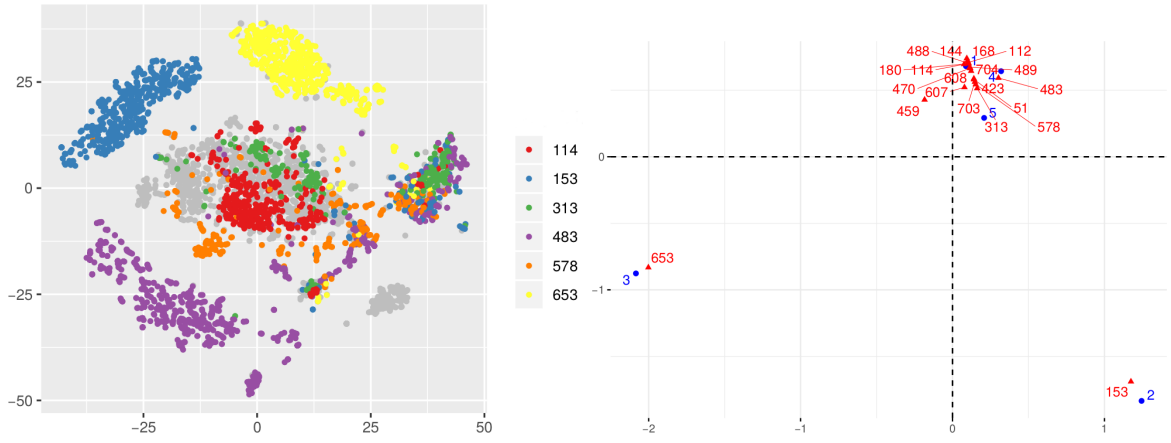
In order to validate the proposed methodology of clustering, we also evaluate how pure are derived segments when it comes to medical diagnoses (ICD-10 codes). As we said, no information about recommendations nor diagnosis is used in the phase of clustering to prevent data leakage. An optimal situation would be if the distributions of ICD-10 codes among clusters would be totally different. Hence, we perform chi-squared test (Pearson, 1900) to measure the difference between the distributions of ICD-10 codes among clusters. The higher result of chi-squared, the more different distributions and the better quality of clusters. Columns 7 and 8 in Table 5.2 show that for every specialty the embedding representation allows us to obtain better ICD-10 distributions than one-hot representations.

Figure 5.5 shows correspondence analysis between clusters and ICD-10 codes (the 20 or 50 most popular codes) for clusterings of three specialties (the rest can be found in Appendix A). In the clustering of family medicine there appeared two large groups of codes: the first related to diseases of the respiratory system (J) and the second related to other diseases, mainly endocrine, nutritional and metabolic diseases (E) and diseases of the circulatory system (I). The first group corresponds to Cluster 1 and the second to Cluster 4. Clusters 3, 5 and 6 (the smallest clusters in this clustering) covered Z76 ICD-10 code (encounter for issue of repeat prescription).

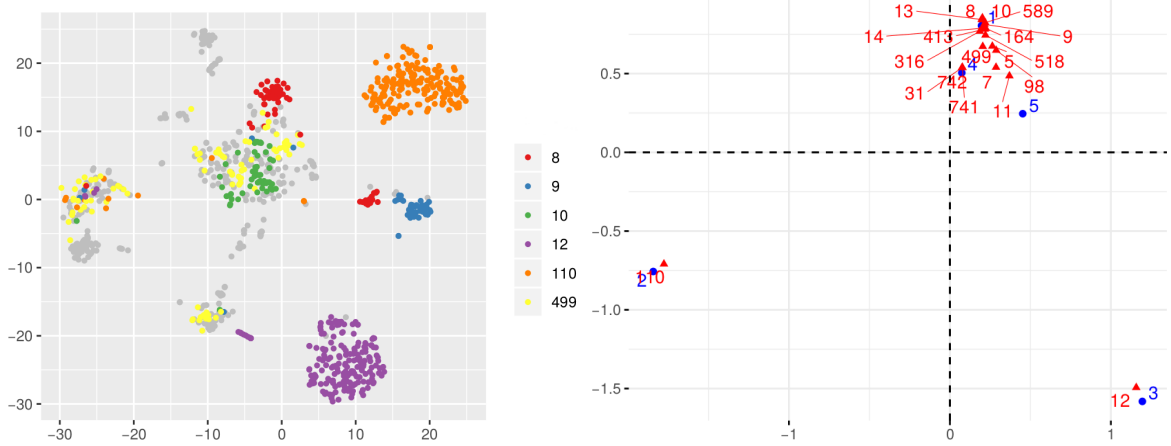
In clustering of gynecology we also have two groups: the diseases of genitourinary system (N) connected with Clusters 1 and 3; and pregnancy, childbirth and the puerperium (O),



(a) Cardiology

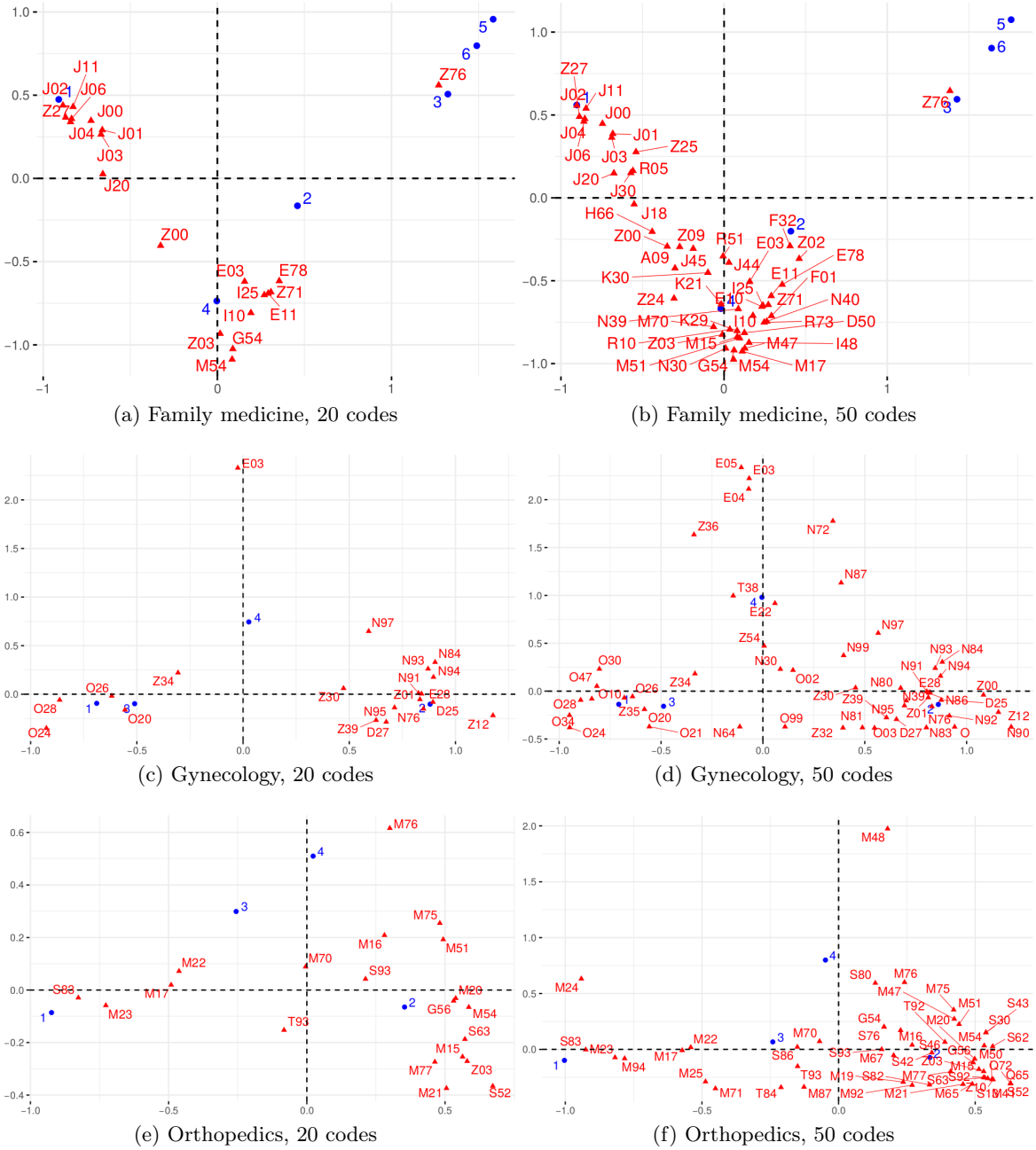


(b) Pediatrics



(c) Psychiatry

Figure 5.4: The distribution of doctors in clusters. In the left panels each dot corresponds to a single visit. Colors correspond to doctors. The right panels show the correspondence analysis between clusters and doctors.



(a) Family medicine, 20 codes

(b) Family medicine, 50 codes

(c) Gynecology, 20 codes

(d) Gynecology, 50 codes

(e) Orthopedics, 20 codes

(f) Orthopedics, 50 codes

Figure 5.5: Correspondence analysis between ICD-10 codes and clusters.

specialty	# clusters	# small clusters		# one-doctor cl.		chi-sq. ICD-10	
		one-hot	*embed.*	*one-hot*	*embed.*	*one-hot*	*embed.*
Cardiology	6	2	1	4	2	949.5	995.1
Dermatology and venereology	6	3	1	3	1	1688.5	1733.2
Endocrinology	5	1	0	2	1	623.1	1014.1
Family medicine	6	2	2	3	1	10282.5	12561.1
Gynecology	4	0	0	2	0	1511.7	2443.3
Internal medicine	5	2	1	2	1	7775.2	11858.7
Orthopedics	4	1	0	2	0	661.5	826.4
Pediatrics	5	1	0	4	3	3288.8	4154.1
Psychiatry	5	1	0	2	2	278.5	289.8

Table 5.2: A comparison of clustering based on weighted one-hot and embedding representations of visits. Columns 3 - 4: the number of clusters smaller than 5% of visits. Columns 5 - 6: the number of clusters where one doctor covers over 95% of visits. Columns 7 - 8: chi-squared test of the distribution of ICD-10.

connected with Cluster 2.

Thus, the presented methodology allowed us to obtain groups of visits with similar diagnoses expressed in ICD-10 codes.

5.3. Recommendations in clusters

The final step of evaluating clusters is an analysis of the recommendation part of the descriptions. Like in Subsection 5.2.2 we would like to obtain different diagnoses in different clusters.

We examine the recommendation terms from five categories, see Section 3.5. Then we find the most popular recommendation terms for every cluster.

Figure 5.6 shows the correspondence analysis between clusters and the 20 most popular recommendations for three specialties (for the rest specialties, see Appendix A). Some terms are more related to one cluster that to others but we have also a lot of terms spread on many clusters.

Hence, in the next step we filter the terms popular in many clusters to find only characteristic recommendations. More specifically, for every cluster we treat a term as popular if it belongs to one of the 15 most frequent terms in this cluster (from Table 5.3 we see that 15 is enough to exclude terms covering 2-3% of visits). Then we discard this term if it is popular in at least three clusters (the total number of clusters was between 4 and 6 so this filtering aims to remove terms popular in at least the half of the clusters).

Tables 5.4 - 5.6 present the 5 most popular recommendations after excluding non-characteristic terms. Unfortunately the number of excluded terms was high and the remained terms are mostly non-frequent. Moreover, characteristic concepts in the clusters are sometimes very semantically similar to the rest of the clusters, e.g. the most frequent recommendation in Cluster 3 in family medicine (Table 5.6), *necessary medical control* (konieczna kontrola lekarska) is very similar to *medical control* (kontrola lekarska) from Cluster 5 or *use of medicines* (stosowanie leków) from Cluster 6 is semantically similar to *systematic taking of medicines* (systematyczne przyjmowanie leków) from Cluster 2. Hence, the using of concepts from recommendation part should be improved.

cluster	size	most frequent recommendations
1	1311	kontrola (50.5%), szpital (21%), zalecenie (16.6%), morfologia (9.3%), zus (8.5%), tryb życia (6.9%), badanie ogólne moczu (5.6%), badanie (4.1%), dieta (4.1%), witamina (4%), profilaktyka stomatologiczna (3.4%), potas (3.1%), obserwacja ruchów płodu (3.1%), badanie laboratoryjne (2.8%), badanie usg (2.6%)
2	1318	kontrola (32.3%), badanie (7.2%), badanie usg (4.6%), terapia (4.1%), leczyć (4%), cytologia (3.8%), szpital (3.2%), usg piersi (3%), morfologia (2.5%), lek (2.2%), konsultacja (2%), zabieg (2%), operacja (1.9%), mammografia (1.8%), tryb życia (1.7%)
3	384	morfologia (35.4%), acidum (31.5%), kontrola (28.9%), tryb życia (25.3%), badanie (15.1%), najbliższy szpital (14.3%), odpowiednia dieta (14.3%), zachowanie prozdrowotne (14.3%), zus (11.2%), badanie usg (10.4%), kontrola położnicza (10.2%), usg ciąży (7.3%), szpital (5.7%), usg prenatalne (3.9%), liczenie ruchów płodu (2.9%)
4	443	kontrola (45.8%), badanie (5.4%), morfologia (2.7%), terapia (2%), leczyć (2%), szpital (1.8%), zus (1.8%), witamina (1.8%), badanie usg (1.6%), dieta (1.6%), lek (1.6%), tryb życia (1.4%), badanie ogólne moczu (1.4%), gin (1.4%), cytologia (1.1%)

Table 5.3: 15 most frequent recommendations for each cluster in clustering of gynecology. In brackets we see a percentage of visits in this cluster that contain a specified term.

cluster	size	most frequent recommendations
1	1311	zalecenie (16.6%), badanie ogólne moczu (5.6%), dieta (4.1%), witamina (4%), profilaktyka stomatologiczna (3.4%)
2	1318	terapia (4.1%), leczyć (4%), cytologia (3.8%), usg piersi (3%), lek (2.2%)
3	384	acidum (31.5%), najbliższy szpital (14.3%), odpowiednia dieta (14.3%), zachowanie prozdrowotne (14.3%), kontrola położnicza (10.2%)
4	443	leczyć (2%), terapia (2%), witamina (1.8%), dieta (1.6%), lek (1.6%)

Table 5.4: Characteristic recommendations for gynecology, skipped terms: *badanie, badanie usg, kontrola, morfologia, szpital, tryb życia, zus.*

cluster	size	most frequent recommendations
1	360	leczenie operacyjne (7.2%), terapia manualna (5.6%), basen (5.3%), zakres ruchomości (4.4%), fizykoterapia (3.9%)
2	1257	laser (5.5%), fizykoterapia (4.4%), terapia manualna (4.4%), ap (4.2%), leczenie operacyjne (4.1%)
3	102	badanie usg (8.8%), basen (3.9%), obciążanie (3.9%), rehabilitacja ruchowa (3.9%), oszczędzanie kończyny (2.9%)
4	150	laser (6.7%), ap (6%), fizykoterapia (6%), iniekcja (6%), powtórzenie (6%)

Table 5.5: Characteristic recommendations for orthopedics, skipped terms: *ćwiczenie, kontrola, leczyć, lek, operacja, orteza, oszczędzanie, rehabilitacja, rezonans magnetyczny, rtg, terapia, zus.*

(a) Family medicine

(b) Gynecology

(c) Orthopedics

Figure 5.6: Correspondence analysis between recommendations (from 5 categories) and clusters.

cluster	size	most frequent lemmas
1	3108	ssanie (9.3%), inhalacja (9.1%), toaleta noska (7.9%), żel (6.9%), probiotyk (5.6%)
2	2353	systematyczne przyjmowanie leków (14.6%), odpoczynek (4%), lekka dieta (3.7%), zwiększenie podaży płynów (3.7%), przepisany lek (3.2%)
3	601	konieczna kontrola lekarska (99.8%), kontynuacja leków (75%), stosowanie (1.2%), badanie ekg (0.2%), badanie laboratoryjne (0.2%)
4	4518	konieczna kontrola lekarska (5.6%), kontynuacja leków (4.4%), dieta (3.8%), regularne przyjmowanie leków (3.7%), pomiary domowe (3.4%)
5	255	kontrola lekarska (96.1%), następna wizyta (2.4%), wskazana kontrola (1.2%), pomiara glikemii (0.8%), gastroskopia (0.4%)
6	395	stosowanie leków (52.2%), stosowanie (0.5%), wizyta domowa (0.5%), działanie (0.3%), kontrolna wizyta (0.3%)

Table 5.6: Characteristic recommendations for family medicine, skipped terms: *badanie, konsultacja, kontrola, kontynuacja leczenia, leczyć, lek, morfologia, rtg, terapia, wizyta, zalecenie*.

Chapter 6

Summary

In the Summary we recapitulate the main achievements of the thesis, make some propositions of an application of the results and present conclusions and suggestions of a future work.

6.1. Achieved results

The most important results of this thesis are as follows:

- proposition of a new methodology of clustering of medical visits from Polish health centers, based on the concepts extracted from free-text descriptions of interviews and examinations written by doctors. The clustering is performed on visits' embeddings created from concepts' embeddings. The method was validated on a large corpus of real medical records and compared with a baseline algorithm. A visual and numerical examination of derived clusters showed an interesting structure among visits. Obtained clusters are linked with medical diagnoses even if neither the information about recommendations nor about diagnosis were used for the clustering. This additionally convinces that the identified structure is related to some subgroups of medical conditions.
- generating good-quality embeddings of Polish medical terminology with GloVe algorithm. The quality of the embeddings was measured by the specific analogy task designed specifically for this corpus. It turns out that analogies work well, what ensures that concept embeddings store some useful information.
- generating embeddings of ICD-10 codes that reflected the structure of ICD-10 classification.

Based on the results obtained in this thesis there was prepared a paper presenting the developed methodology (Dobrakowski et al., 2019). The paper was submitted to BioNLP 2019: 18th ACL Workshop on Biomedical Natural Language Processing.

6.2. Proposition of applications

The proposed methodology can be used to help doctors in their practice during patients‘ visits. It allows us to assign new visits to already derived clusters. Based on a description of an interview or a description of patient examination we can generate an embedding of the visit and find the closest cluster. Then we can identify similar visits and show corresponding ICD-10 codes and recommendations. The computer program could show the list of proposed

diagnoses to a doctor during filling up the form, which would shorten the duration of the visit. On the other hand, if the doctor wanted to prescribe an unusual recommendation, he could be warned that he probably made a mistake.

Some partial results of the presented methodology also can be used in many other applications. The obtained embeddings can be utilized in creating publicly available Polish medical data resource, like the Unified Medical Language System for English, with a medical ontology and the GloVe concept embeddings. The embeddings can be a base for other algorithms, like ICD-10 classification, etc.

6.3. Future work

The presented methodology is the first trial known to us to cluster visits based on descriptions written in the Polish language. Although the results are very interesting and promising, there are still a lot of directions that could improve this methodology.

Firstly, in generating concepts' embeddings it is worth to make experiments with the state-of-the-art contextual embedding techniques like BERT or ELMo. It could enable us to adapt the existing embeddings to a specific context.

Secondly, in the representations of the visits we skip the order of the terms and their repetitions. Maybe it is important to assign different weights to different concepts during generating visit's embeddings. Now we compute a simple arithmetic mean of terms' embeddings. Moreover, in the process of extracting the concepts, we omit the numbers, which could store a lot of information, for example about a result of a measurement. Including these elements could improve the quality of visits' embeddings.

Especially it is desired to overcome problems with doctors' schemes and specific vocabulary which resulted in doctor-fitted clusters. However, this problem can disappear when we will consider the bigger data set (now we have available only 100,000 visits out of about 3 million), where will be fewer terms used by only one doctor.

The representations of visits can be further developed in the context of ICD-10 codes. Our embeddings of ICD-10 codes reflect their meaning but still it is unclear why some codes are preferred by some doctors. It would be worth to check if there exist codes that are assigned interchangeably by one doctor. Also, we would like to detect the situations where two doctors assign two different codes in the same case.

The proposed methodology performs a clustering based only on the visit's description written by a doctor. In Section 1.2 we have seen an example of an algorithm that uses only a history of ICD-10 codes. The natural way of developing the methodology would be a trial for connecting these two algorithms. The interesting way could be using of the embeddings of ICD-10 codes (instead of ICD-10 codes) as a part of information encoded in visit embedding. In tensor decomposition algorithm we could replace one-hot representation of diseases history by the representation computed from embeddings of ICD-10 codes.

Additionally for clustering we could use other information from medical records, like age, sex of patients.

The process of extracting information about recommendations in clusters also can be improved. In this work we examined only five categories of recommendation terms. But not all concepts represented a complete recommendation. Also, we did not look at the similarity of recommendation terms. Maybe generating of recommendation terms' embeddings would help in better recommendation prediction.

Finally, the evaluation could include some feedback by human experts. We could gain more insights into the findings.

Appendix A

Results of visits clustering

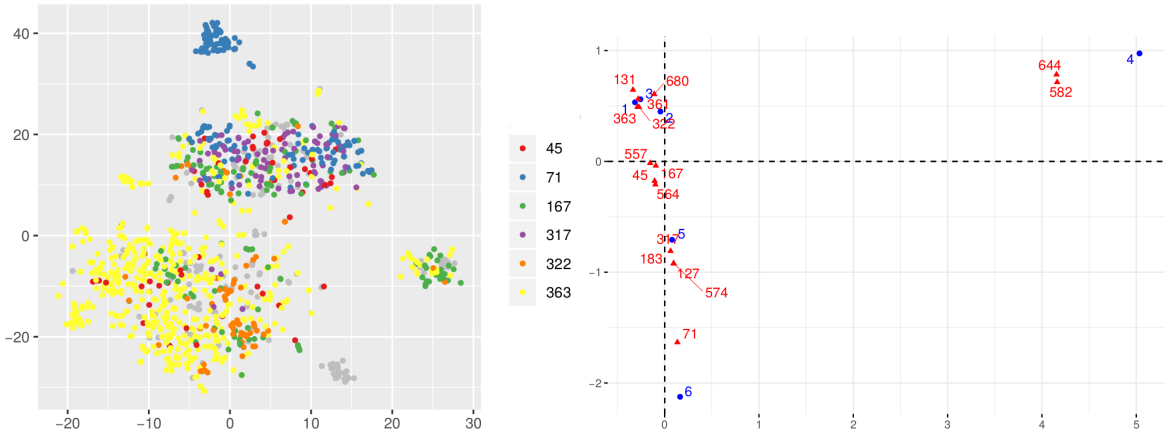
In Appendix we can find the rest of the analysis of the obtained clusters (for specialties not included in Chapter 5).

Figure A.1 continues the plots presented in Figure 5.4 and illustrates the distribution of visits by doctors and the relationships between the obtained clusters and the doctors.

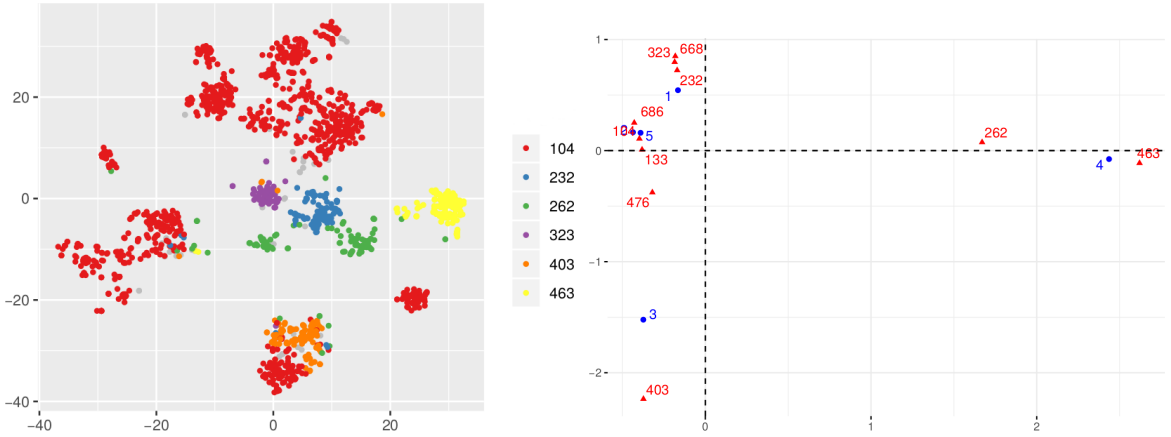
Figure A.2 continues the plots presented in Figure 5.5 and shows the correspondence analysis between clusters and ICD-10 codes (the 20 or 50 most popular codes).

Figure A.3 continues the plots presented in Figure 5.6 and shows the correspondence analysis between clusters and the 20 most popular recommendations.

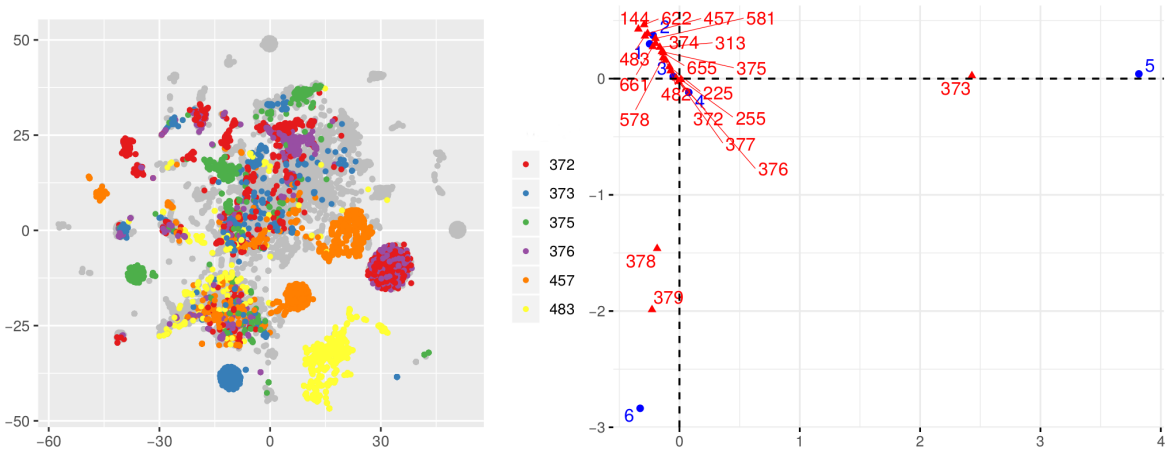
Tables A.1 - A.6 continue the results presented in Tables 5.6 - 5.5 and show the 5 most popular recommendations for every cluster after excluding non-characteristic terms.



(a) Dermatology and venereology

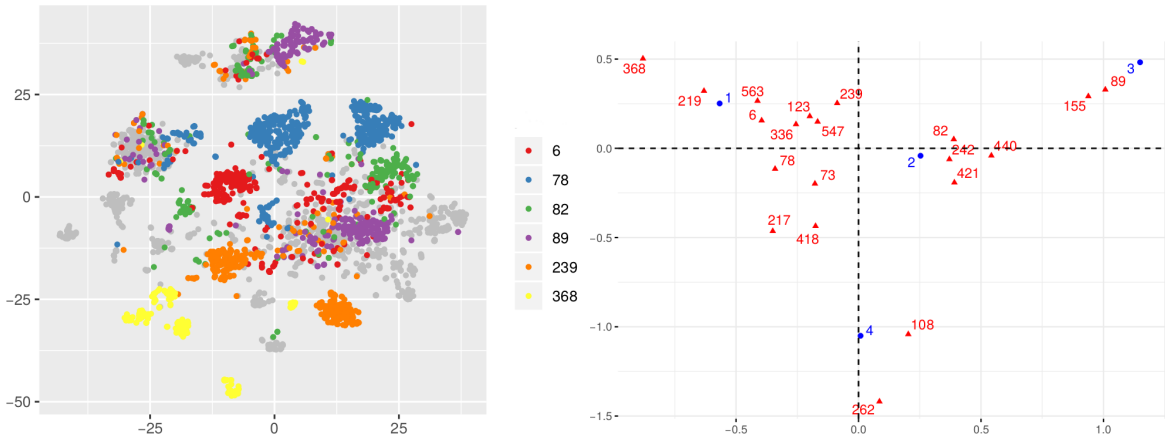


(b) Endocrinology

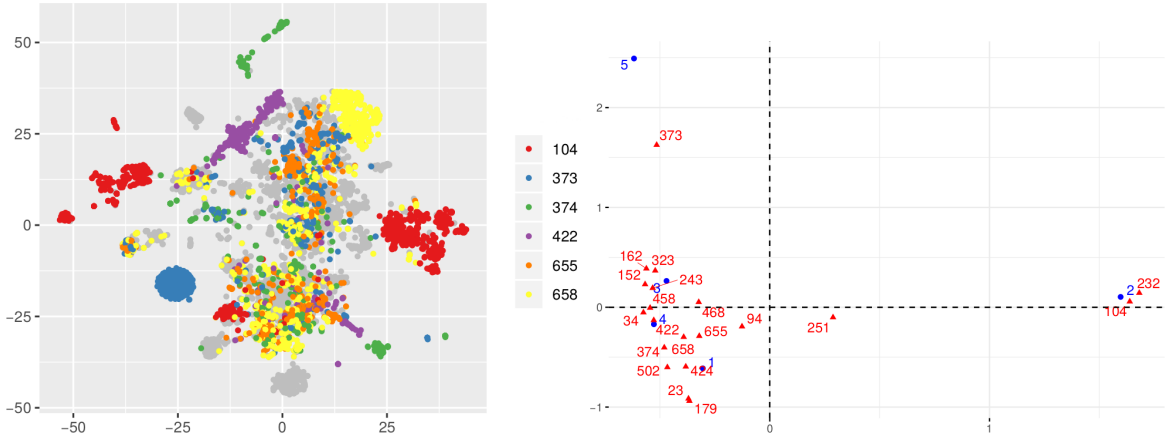


(c) Family medicine

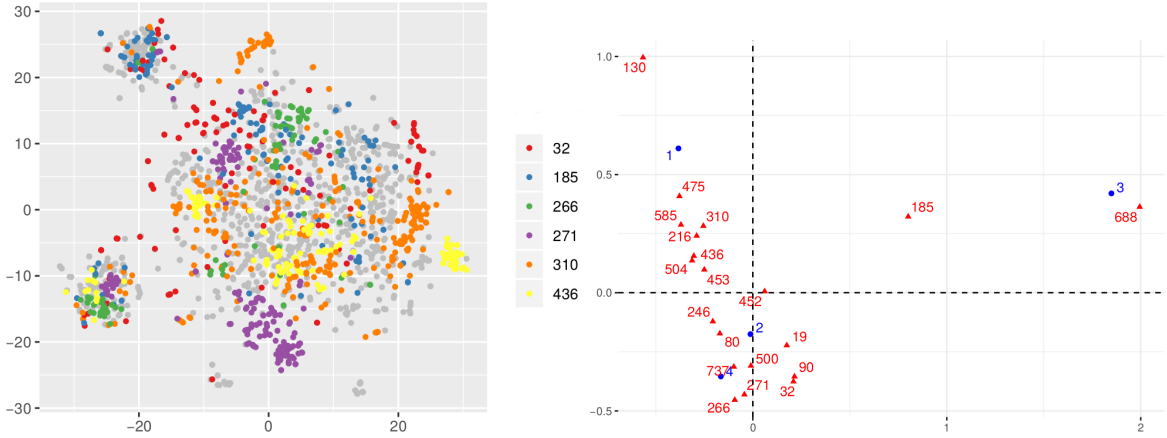
Figure A.1: The distribution of doctors in clusters. In the left panels there are projected embeddings of visits colored by doctors' IDs. The right panels show the correspondence analysis between clusters and doctors.



(d) Gynecology

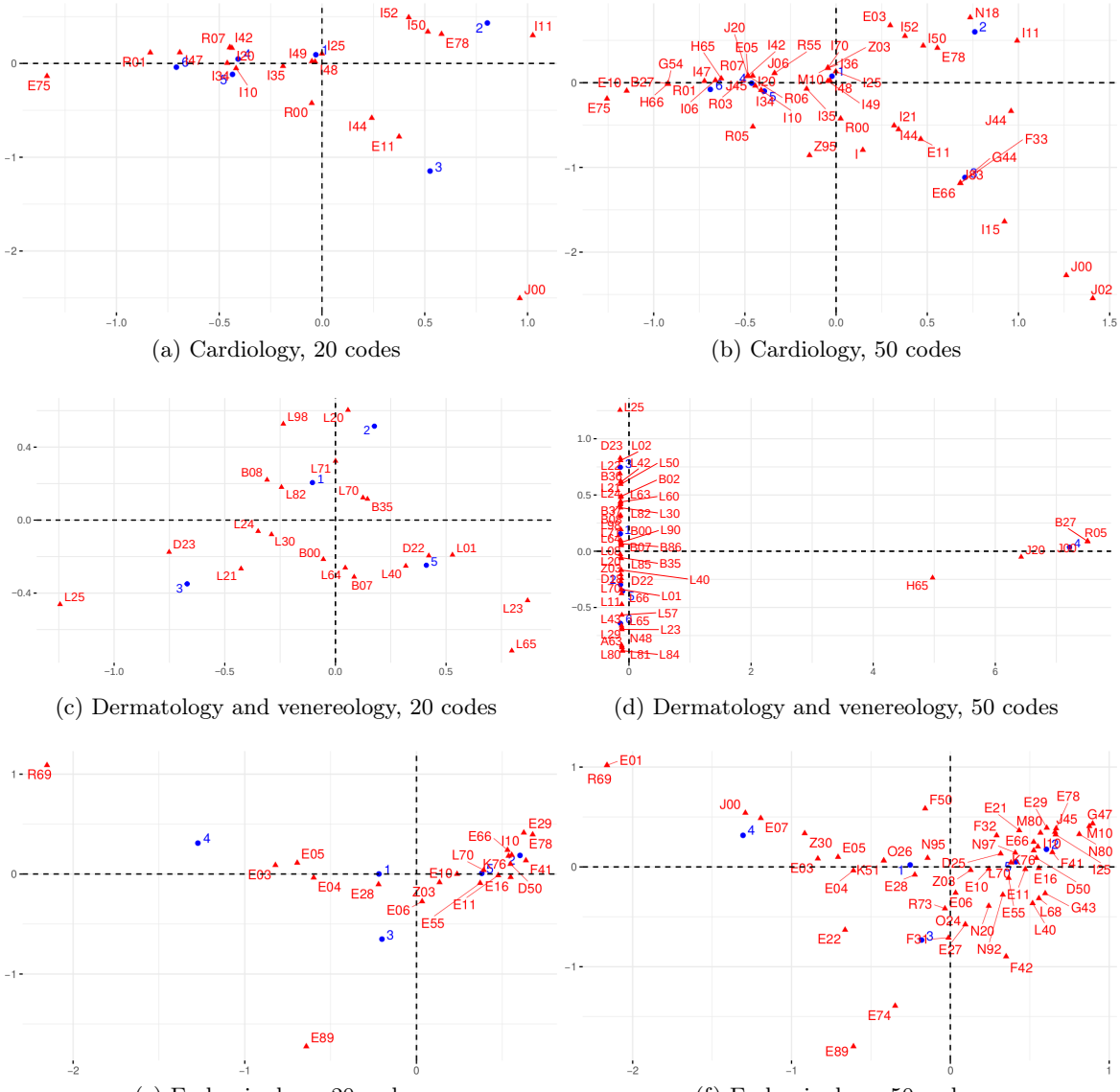


(e) Internal medicine



(f) Orthopedics

Figure A.1: The distribution of doctors in clusters (cont.).



(a) Cardiology, 20 codes

(b) Cardiology, 50 codes

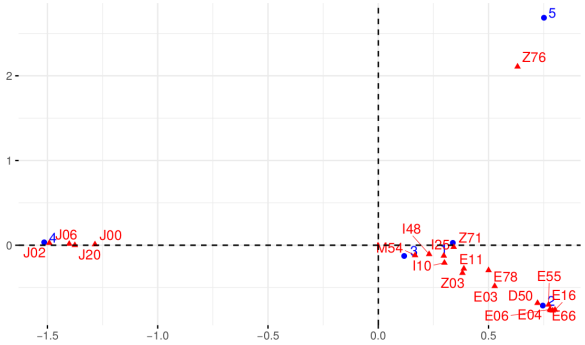
(c) Dermatology and venereology, 20 codes

(d) Dermatology and venereology, 50 codes

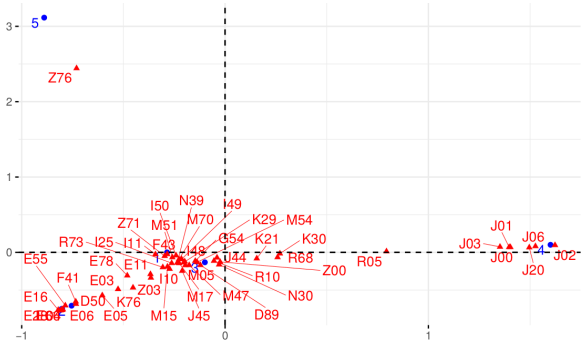
(e) Endocrinology, 20 codes

(f) Endocrinology, 50 codes

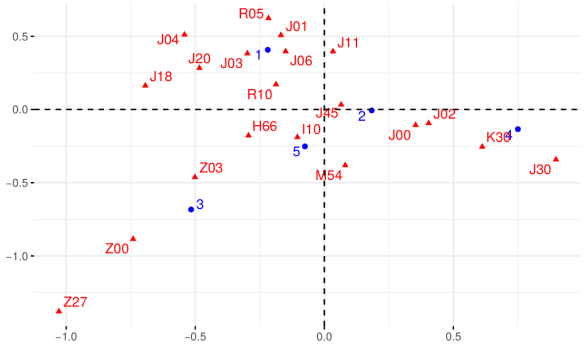
Figure A.2: Correspondence analysis between ICD-10 codes and clusters.



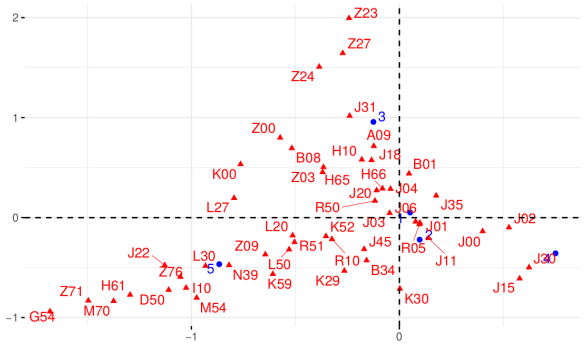
(g) Internal medicine, 20 codes



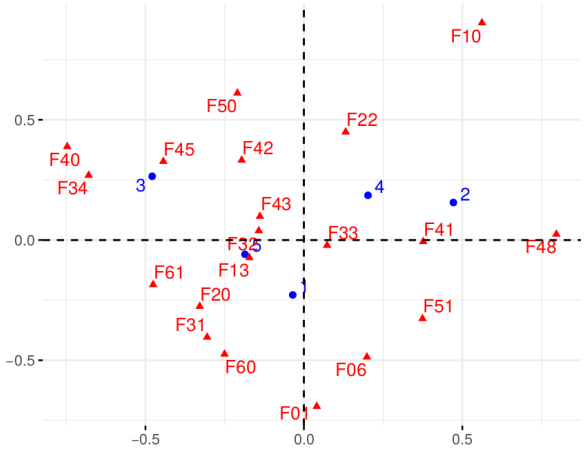
(h) Internal medicine, 50 codes



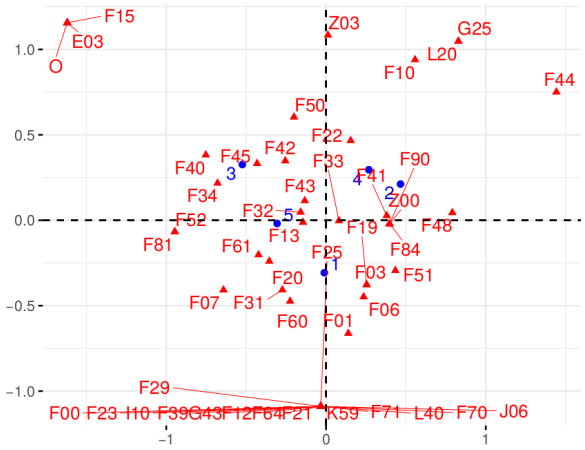
(i) Pediatrics, 20 codes



(j) Pediatrics, 50 codes

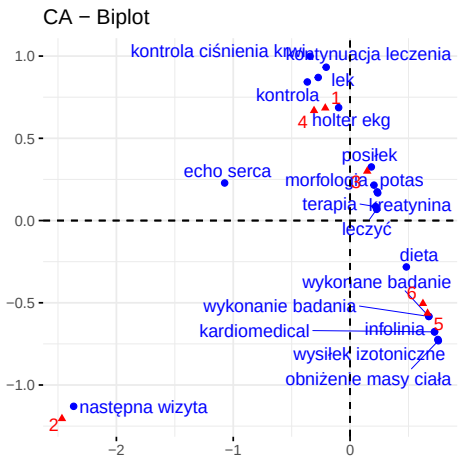


(k) Psychiatry, 20 codes

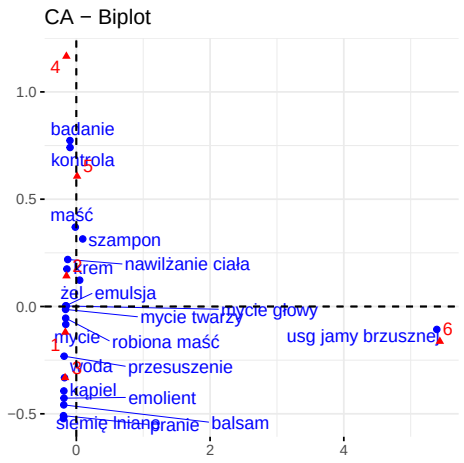


(l) Psychiatry, 50 codes

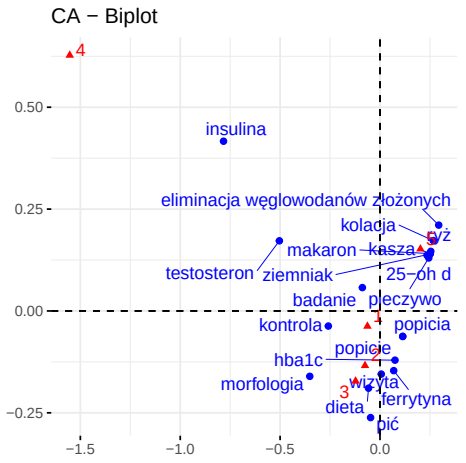
Figure A.2: Correspondence analysis between ICD-10 codes and clusters (cont).



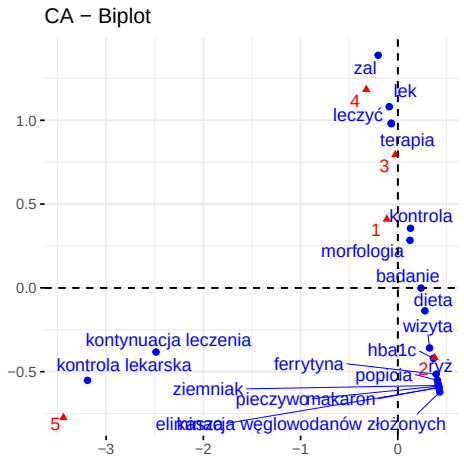
(a) Cardiology



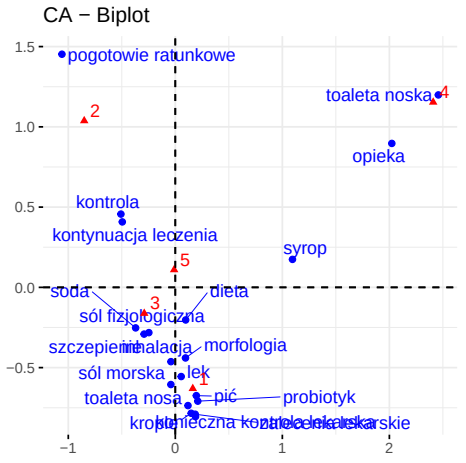
(b) Dermatology and venereology



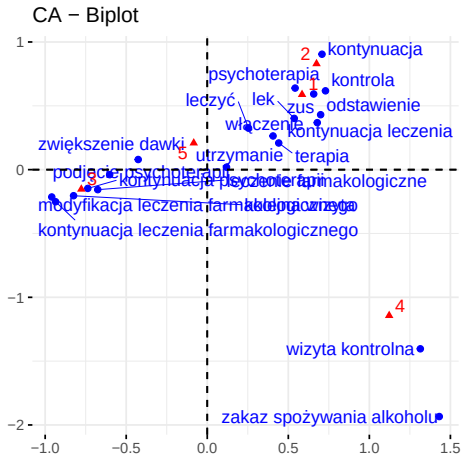
(c) Endocrinology



(d) Internal medicine



(e) Pediatrics



(f) Psychiatry

Figure A.3: Correspondence analysis between recommendations (from 5 categories) and clusters.

cluster	size	most frequent lemmas
1	428	holter ekg (7.5%), kontynuacja leczenia (7%), badanie ekg (5.6%), codzienny spacer (5.1%), przestrzeganie zaleconej diety (5.1%)
2	193	echo serca (8.8%), wizyta (3.1%), b12 (1%), badanie ekg (0.5%), codzienny spacer (0.5%)
3	134	posiłek (18.7%), dieta bezcukrowa (17.2%), aktywność fizyczna (14.2%), wysiłek programowany (14.2%), regularny pomiar ciśnień tętniczych krwi (12.7%)
4	303	kontrola ciśnienia krwi (12.9%), echo serca (11.2%), holter ekg (9.2%), ponowna kontrola (7.6%), pilna konsultacja (6.6%)
5	27	infolinia (100%), kardiomedical (100%), obniżenie masy ciała (66.7%), wykonane badanie (48.1%), wykonanie badania (48.1%)
6	116	infolinia (90.5%), kardiomedical (90.5%), obniżenie masy ciała (61.2%), wykonane badanie (39.7%), wykonanie badania (39.7%)

Table A.1: Characteristic recommendations for cardiology, skipped terms: *kontrola, leczyć, lek, morfologia, następna wizyta, terapia.*

cluster	size	most frequent recommendations
1	455	przesuszenie (20.7%), balsam (13.2%), pranie (12.5%), woda (10.5%), siemię lniane (8.8%)
2	89	kontrola (13.5%), przesuszenie (13.5%), probiotyk (9%), natłuszczanie (6.7%), terapia (6.7%)
3	176	balsam (18.8%), pranie (17.6%), woda (15.9%), siemię lniane (12.5%), szklanka ziaren siemienia lnianego (11.9%)
4	30	probiotyk (40%), inhalacja (36.7%), sól (36.7%), lek (26.7%), antybiotyk (13.3%)
5	391	kontrola (15.1%), morfologia (11.3%), acidum (9.7%), badanie (8.7%), rozmaz (7.2%)
6	63	usg jamy brzusznej (87.3%), insulina (85.7%), leczenie zaostrzenia (6.3%), leczenie zapobiegawcze (6.3%), acidum (3.2%)

Table A.2: Characteristic recommendations for dermatology and venereology, skipped terms: *emolient, emulsja,kąpiel, krem, maść, mycie, mycie twarzy, robiona maść, sód, szampon, żel.*

cluster	size	most frequent recommendations
1	389	eliminacja węglowodanów złożonych (12.6%), 25-oh d (12.1%), usg tarczycy (11.6%), testosteron (9.3%), witamina (7.5%)
2	412	testosteron (18.9%), eliminacja węglowodanów złożonych (17.7%), 25-oh d (14.6%), krople (6.1%), estradiol (4.6%)
3	208	krople (8.2%), eliminacja węglowodanów złożonych (6.2%), testosteron (5.8%), 25-oh d (4.8%), estradiol (2.9%)
4	183	plaster (51.4%), usg tarczycy (49.2%), obniżyć (42.6%), kontrola endokrynologa (41%), witamina (39.3%)
5	318	eliminacja węglowodanów złożonych (48.1%), 25-oh d (33.6%), testosteron (17.6%), estradiol (16%), krople (9.4%)

Table A.3: Characteristic recommendations for endocrinology, skipped terms *badanie, dieta, ferrytyna, hba1c, insulina, kasza, kontrola, makaron, pieczywo, popicia, popicie, ryż, wizyta, ziemniak.*

cluster	size	most frequent recommendations
1	1915	terapia (7.1%), leczyć (7%), następna wizyta (7%), kontynuacja leczenia (6.4%), infolinia (5.5%)
2	1173	ferrytyna (45.9%), hba1c (37.3%), pieczywo (26.5%), makaron (26%), kasza (25.8%)
3	1930	terapia (7.9%), leczyć (7.8%), rtg (4.5%), kreatynina (3.5%), badanie ekg (3.3%)
4	1146	żel (11.3%), ssanie (11.2%), zus (10.8%), nawilżanie gardła (10.7%), probiotyk (8.1%)
5	255	kontynuacja leczenia (100%), kontrola lekarska (96.1%), następna wizyta (2.4%), wskazana kontrola (1.2%), pomiara glikemii (0.8%)

Table A.4: Characteristic recommendations for internal medicine, skipped terms: *badanie, dieta, kontrola, lek, morfologia, wizyta, zal.*

cluster	size	most frequent recommendations
1	1751	toaleta nosa (8.2%), zalecenia lekarskie (7.8%), konieczna kontrola lekarska (7.5%), pić (6.8%), krople (5.9%)
2	658	pogotowie ratunkowe (58.4%), zus (3.6%), toaleta nosa (1.4%), dieta (1.2%), paracetamol (1.2%)
3	666	szczepienie (17.4%), soda (15%), sól morska (7.5%), leczyć (3.5%), terapia (3.5%)
4	715	toaleta noska (30.6%), syrop islandzki (11%), cebula (10.8%), żel (9.2%), opieka (8.7%)
5	952	dieta (4.4%), toaleta noska (3.2%), potas (2.9%), zalecenie (2.8%), badanie (2.7%)

Table A.5: Characteristic recommendations for pediatrics, skipped terms: *inhalacja, kontrola, kontynuacja leczenia, lek, morfologia, probiotyk, sól fizjologiczna, syrop.*

cluster	size	most frequent recommendations
1	441	psychoterapia (5.9%), kontrola (5%), następna wizyta (3.4%), kontynuacja (3.2%), próba (2.5%)
2	184	kontynuacja (12.5%), zwiększenie (8.7%), psychoterapia (2.7%), ponowne włączenie (1.6%), spacer (1.6%)
3	179	kontynuacja leczenia farmakologicznego (33%), modyfikacja leczenia farmakologicznego (31.8%), podjęcie psychoterapii (17.9%), leczenie farmakologiczne (10.1%), prowadzenie pojazdów (10.1%)
4	133	wizyta kontrolna (30.1%), zakaz spożywania alkoholu (30.1%), kontrola stanu psychicznego (6.8%), farmakoterapia (5.3%), umawianie wizyty (5.3%)
5	75	podjęcie psychoterapii (9.3%), leczenie farmakologiczne (6.7%), modyfikacja leczenia farmakologicznego (6.7%), prowadzenie pojazdów (6.7%), badanie (5.3%)

Table A.6: Characteristic recommendations for psychiatry, skipped terms: *kolejna wizyta, kontynuacja leczenia, kontynuacja psychoterapii, leczyć, lek, odstawienie, terapia, utrzymanie, włączenie, zus, zwiększenie dawki.*

Bibliography

Banea, C., D. Chen, R. Mihalcea, C. Cardie, and J. Wiebe
2014. Simcompass: Using deep learning word embeddings to assess cross-level similarity. In *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)*, Pp. 560–565.

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

Chiu, B., G. Crichton, A. Korhonen, and S. Pyysalo
2016. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th workshop on biomedical natural language processing*, Pp. 166–174.

Choi, E., M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun
2016a. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Pp. 1495–1504. ACM.

Choi, E., A. Schuetz, W. F. Stewart, and J. Sun
2016b. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*.

Choi, Y., C. Y.-I. Chiu, and D. Sontag
2016c. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41.

De Boom, C., S. Van Canneyt, T. Demeester, and B. Dhoedt
2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156.

De Vine, L., G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza
2014. 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. ACM.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman
1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Dempster, A. P., N. M. Laird, and D. B. Rubin
1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova
2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dobrakowski, A., A. Mykowiecka, M. Marciniak, W. Jaworski, and P. Biecek
2019. Patients' visits segmentation based on word embeddings. *arXiv preprint arXiv:1907.04152*.

Fetter, R., A. Novák, and G. Prószéky
1980. Case mix definition by diagnosis related groups. volume 18, Pp. 1—-53.

Frantzi, K., S. Ananiadou, and H. Mima
2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *Int. Journal on Digital Libraries*, 3:115–130.

Hirschfeld, H. O.
1935. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, Pp. 520–524. Cambridge University Press.

Jaworski, W. and J. Kozakoszczak
2016. Eniam: Categorial syntactic-semantic parser for polish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Pp. 243–247.

Jaworski, W., D. Oklesiński, J. Lupa, S. Rutkowski, J. Kozakoszczak, J. Przetacka, H. Teleżyńska, B. Antonowicz, A. Markiewicz, J. Kowalewski, M. Pieńkosz, and A. Morusiewicz
2018. Categorial parser. CLARIN-PL digital repository.

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

Marciniak, M., A. Mykowiecka, and P. Rychlik
2016. TermoPL — a flexible tool for terminology extraction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds., Pp. 2278–2284, Portorož, Slovenia. ELRA, European Language Resources Association (ELRA).

Mikolov, T., K. Chen, G. Corrado, and J. Dean
2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Minarro-Giménez, J. A., O. Marin-Alonso, and M. Samwald
2014. Exploring the application of deep learning techniques on medical text corpora. *Studies in health technology and informatics*, 205:584–588.

Newman-Griffis, D., A. M. Lai, and E. Fosler-Lussier
2017. Insights into analogy completion from the biomedical domain. *arXiv preprint arXiv:1706.02241*.

Organization, W. H. et al.
2004. Icd-10: international statistical classification of deseases and related health problems. In *ICD-10: international statistical classification of deseases and related health problems*.

Orosz, G., Y. Shin, J. Freeman, and R. Averill
2013. Hybrid text segmentation for hungarian clinical records. In *Advances in Artificial Intelligence and Its Applications. MICAI 2013*. Springer.

Pearson, K.
1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Pennington, J., R. Socher, and C. Manning
2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Pp. 1532–1543.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer
2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Rand, W. M.
1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Ruffini, M., R. Gavaldà, and E. Limón
2017. Clustering patients with tensor decomposition. *arXiv preprint arXiv:1708.08994*.

Salakhutdinov, R. and G. Hinton
2009. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978.

Salton, G. and C. Buckley
1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Ward Jr, J. H.
1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Waszczuk, J.
2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Pp. 2789–2804.