Journal of Intelligent Information Systems (2021) 57:447–465
https://doi.org/10.1007/s10844-021-00659-4

Interpretable segmentation of medical free-text records based on word embeddings

Adam Gabriel Dobrakowski[1] · Agnieszka Mykowiecka[2] · Małgorzata Marciniak[2] · Wojciech Jaworski[1] · Przemysław Biecek[1,3]

Received: 10 December 2020 / Revised: 2 July 2021 / Accepted: 12 July 2021 /
Published online: 28 October 2021

Abstract

Medical free-text records store a lot of useful information that can be exploited in developing computer-supported medicine. However, extracting the knowledge from the unstructured text is difficult and depends on the language. In the paper, we apply Natural Language Processing methods to process raw medical texts in Polish and propose a new methodology for clustering of patients' visits. We (1) extract medical terminology from a corpus of free-text clinical records, (2) annotate data with medical concepts, (3) compute vector representations of medical concepts and validate them on the proposed term analogy tasks, (4) compute visit representations as vectors, (5) introduce a new method for clustering of patients' visits and (6) apply the method to a corpus of 100,000 visits. We use several approaches to visual exploration that facilitate interpretation of segments. With our method, we obtain stable and separated segments of visits which are positively validated against final medical diagnoses. In this paper we show how algorithm for segmentation of medical free-text records may be used to aid medical doctors. In addition to this, we share implementation of described methods with examples as open-source `R` package `memr`.

Keywords Electronic health records · Natural language processing · Text clustering · Word embeddings

1 Introduction

Information extraction from free-text clinical records plays an important role in computer-supported medicine (Apostolova et al., 2009; Ganesan & Subotin, 2014). It is because a detailed description of symptoms, physical examination results and medical interviews are frequently stored in an unstructured way as free-text. Such free text is rich in important information but it is also hard to process for classical machine learning algorithms. Although there have been a number of attempts to automatize the processing of medical notes for

✉ Przemysław Biecek
P.Biecek@mimuw.edu.pl

Extended author information available on the last page of the article.

English, Dalianis (2018), and some for other languages, e.g. for Swedish (Névéol et al., 2018), in general, the problem is still a challenge (Orosz et al., 2013). The description of visits can be used for many purposes, such as: to automate diagnostics, to classify patients, to extract specific patient characteristics, to search in the historical data of patients that is similar to the examined cases. In this work, we are mainly interested in the problem of grouping visits, i.e. dividing visits into segments of visits with similar descriptions of the interview, the examination, and the therapeutic recommendations.

Segmentation of visits can fulfill many potential goals. If we are able to group visits into clusters based on patient interviews and medical examination results, we can aggregate recommendations that were suggested to patients with a similar history to create a list of possible consensus diagnoses; to reveal that the current diagnosis is atypical; and to identify subsets of visits with the same diagnosis but different symptoms. The goal of the segmentation of patients is usually formulated in general terms like dividing patients into groups with similar behavior or similar features. In the case of the segmentation of hospitalized patients, one of the most well-known examples is Diagnosis Related Groups (Fetter et al., 1980) which aim is to divide patients into groups defined based on the costs of treatment. This is an important issue for cost estimation and budgeting related to medical services. In this work, however, we will generate segments for a different purpose, as medical decision support for the physician's diagnosis. In order for the doctor to make decisions based on these analyses, it is crucial to ensure that the doctor has trust in them. We will focus on the issue of explainability and interpretability of the segmentation.

Segmentation (cluster analysis) is a well-studied domain for structured data such as age, sex, place, history of diseases, ICD-10 code etc. For an example of segmentation of patients based only on their history of diseases, see Ruffini et al. (2017). Many algorithms such as k-means or hierarchical clustering are typically used for numerical data, where it is easy to determine the distances between observations. On the other hand, segmentation is far from being solved for unstructured free-texts, which requires making many decisions on the contextual meaning of the text. The text itself may be of different length or have a different degree of details in the description. Medical concepts to be extracted from texts are very often taken from the Unified Medical Language System (UMLS, see Bodenreider (2004)), which is a commonly accepted base of biomedical terminology. Representations of medical concepts are computed based on various medical texts, such as medical journals, books, etc. Minarro-Giménez et al. (2014), De Vine et al. (2014), Newman-Griffis et al. (2017), Choi et al. (2016c), and Chiu et al. (2016) or based directly on data from the Electronic Health Records (Choi et al., 2016a; Choi et al., 2016b; Choi et al., 2016c). Another approach for patient segmentation is given in Choi et al. (2016a). A subset of medical concepts (e.g. diagnosis, medication, procedures) and embeddings is aggregated for all visits of a patient. This way we get a patient embedding that summarizes a patient's medical history.

In this work, we group individual visits not patients. For each visit, our data includes a description of the interview, the examination, and recommendations for the treatment, as well as the diagnosis and additional information about a patient and a physician. Because we group visits, a single patient can therefore belong to several clusters. However, this is not a problem, because if we want to support the work of the doctor during one visit, this visit belongs to only one cluster.

Moreover, our segmentation is based on a dictionary of medical concepts created from data, as The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is not translated into Polish. The UMLS resources for Polish are limited to Medical Subject Headings (MeSH) which is a controlled biomedical vocabulary that was created to index medical literature and make it easier to search. It contains 30% to 40% of terminology entries extracted from hospital documents according to Masarie and Miller (1987) (English) and Marciniak (2015) (Polish).

Some examples of visual exploration of supervised models for structured medical data are given in Gordon et al. (2019), Kobylińska et al. (2019), and Biecek (2018). Our paper addresses the problem of explainable machine learning for unsupervised models. Results of clustering obtained within this work can be shown by the several ways of visual exploring that facilitate interpretation of segments. All the presented methods are implemented in the R package memr.

This article is an extended version of the work presented at ISMIS 2020 (Dobrakowski et al., 2020). We added a detailed description of the extraction of medical concepts that was not included in the conference article, and validation of obtained segmentation, and a description of the open source package memr that implements all presented methods.

2 The Polish corpus of free-text clinical records

All results presented in this paper for segmentation are developed and validated on a dataset of free-text clinical records of about 100,000 visits. Our data set consists of descriptions of patients' visits from different primary health care centers and specialist clinics in Poland. They have a free-text form and are written by doctors representing a wide range of medical professions, e.g. general practitioners, dermatologists, cardiologists and psychiatrists. Each description is divided into three parts: interview, examination, and recommendations.

The interview with the patient includes a description of the symptoms with which the patient came and answers to the questions asked by the doctor about his or her health. It may also include the results of provided laboratory tests. The examination section consists of a description of the results of the examination performed during the visit. The most common are physical examination and gynecological examination. The recommendations mainly consist of dosing descriptions of prescribed drugs. There is also information about referrals for examinations and issued exemptions.

A characteristic feature of the texts is their considerable repetition. Individual doctors have their own predefined texts, which they paste into the appropriate fields and then edit them to fit a specific patient. This is due to the fact that a doctor performs a series of tests on a patient during a visit, and even for sick people most of results are normal, and only one or two tests are disturbed (e.g. the throat and lungs of a patient complaining of cough). There is also a phenomenon of copying recommendations, e.g. a set of recommendations for drugs and diet in the case of diabetes. Only the dosage of individual drugs in such recommendations is modified.

3 Methodology

In this section, we introduce the algorithm for the interpretable clustering of medical visits. The algorithms is based on the following four steps: (1) Medical concepts are extracted from free-text descriptions of an interview and examination. (2) A new representation of identified concepts is derived using concept embedding. (3) Concept embeddings are transformed into visit embeddings. (4) Clustering is performed on visit embeddings. The whole process is supplemented with visualizations to facilitate the evaluation of the obtained segmentation.

3.1 Extraction of medical concepts

We conducted our analysis on free-text descriptions in Polish. 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. The doctors' notes are usually 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 which consist of a noun optionally modified by a sequence of adjectives or by another noun in the genitive. We have only identified sequences that can be interpreted as phrases in Polish.

To get the most common phrases, we processed 100,000 visit descriptions. First, we preprocessed texts using the Concraft tagger (Waszczuk, 2012) which assigned lemmas, parts of speech, and morphological feature values. It also guessed descriptions (apart from lemmas) for words which were not present in its vocabulary. Phrase extraction and ordering was performed by TermoPL (Marciniak et al., 2016). As Polish is the inflectional language, the program collected all forms of phrases identified in text, e.g. the phrase *lewa ręka* 'left hand' was represented in data by the six following strings: *lewa ręka, lewej ręki, lewej ręce, lewą rękę, lewą ręką, lewym ręku*. The program allowed a grammar describing extracted text fragments to be defined, but we used the built-in grammar of noun phrases. The phrases were ordered according to a version of the C-value coefficient (Frantzi et al., 2000) which ranked all candidates according to their frequency, length and the contexts in which they appeared.

The first 4,800 phrases (all with a C-value equal to 20 at least) from the obtained list were manually annotated with semantic labels. Among the phrases, 330 synonymous pairs were identified. For example, the acronym *azs* was joined with the full form *atopowe zapalenie skóry* 'atopic dermatitis'; and *ból* 'acke' was connected with the common spelling error *bol*. The list of 132 labels covered most general concepts such as *anatomy, feature, disease*, and *test*. Most of the concept representations consist of only one word (11,083), while there are 4,144 two word phrases and 1,747 longer phrases (the longest consist of seven words), Table 1 gives a dozen of the top multiword phrases. Table 2 shows the number of different subtypes of the most important concepts, and examples of phrases that have these labels, while in Table 3, the number of all occurrences of phrases belonging to these concepts recognized within the entire data set together with the number of occurrences of their most frequent subtypes is depicted. It should be noted that our data confirm very frequent occurrence of negation within examination descriptions.

Many labels were assigned to multi-word expressions (MWEs). In some cases, elements of phrases were also labeled separately, e.g. 'left hand' had the *anatomy* label; 'hand' had the *anatomy* label too, while 'left' the *lateralization* one. The additional source of information was the list of 9,993 names of medicines and dietary supplements, but they were not common in the processed parts of visits (interviews and medical examinations).

The list of phrases together with their semantic labels was then converted to the format of lexical resources of Categorial Syntactic-Semantic Parser "ENIAM" (Jaworski & Kozakoszczak, 2016; Jaworski et al., 2018). The parser recognized lexemes and MWEs in texts according to the provided list of phrases. The longest sequence of recognized tokens was then selected, and semantic representation was created. Semantic representation of a visit had the form of a set of pairs composed of recognized terms and their labels (not recognized tokens were omitted). The same semantic representation was assigned to all forms of a phrase and its synonyms. As the vocabulary of texts was rather limited, the average

Table 1 Top phrases extracted from the free-text recordings

No	Phrase (Polish version and translation to English)	semantic label
1.	*stan ogólny dobry* 'good general condition'	condition
2.	*brzuch miękki* 'abdomen soft'	symptom
3.	*wywiad aktualny* 'current interview'	documentation
4.	*szmer pęcherzykowy* 'alveolar murmor'	symptom
5.	*drogi oddechowe* 'respiratory tract'	anatomy
6.	*objawy ogólne* 'general symptoms'	symptom
7.	*szmer pęcherzykowy prawidłowa* 'normal alveolar murmur'	symptom
8.	*tony czyste* 'pure tones'	symptom
9.	*czynność serca* 'activity of the heart'	physiology
10.	*porada lekarska ambulatoryjna* 'outpatient medical advice'	documentation
11.	*szmer oddechowy pęcherzykowy prawidłowa* 'normal alveolar respiratory murmur'	symptom
12.	*jama ustna* 'oral cavity'	anatomy
13.	*objaw otrzewnowy ujemny* 'negative peritoneal symptoms'	symptom

coverage of semantic representation was quite high: 82.06% of tokens and 75.38% of symbols in the *Interview* section and 87.43% of tokens and 79.28% of symbols in the *Examination* section. These statistics cover both lexemes included in the dictionary and recognized numeral tokens.

3.2 Embeddings for medical concepts

Operating on a relatively large number of very specific texts, we decided not to use any generic model for the Polish language. Given the amount of data available, in our experiments, we reduced the description of visits to extracted concepts and trained our own

Table 2 The most representative phrase labels; the number of phrases with the label and examples

Sem. label	# phr.	Examples of phrases
Symptom	1303	*węzeł chłonny powiększony* 'enlarged lymph node'; *ból gardła* 'sore throat'
Anatomy	701	*lewa noga* 'left leg'; *cewka moczowa* 'urethra'
Disease	342	*zapalenie płuc* 'pneumonia'; *zawał serca* 'myocardial infarction'
Procedure	198	*kontynuacja leczenia* 'continuation of treatment'; *dieta cukrzycowa* 'diabetic diet'
Physiology	178	*tętno obwodowe* 'peripheral pulse'; *wydolność fizyczna* 'physical performance'
Test	178	*posiew moczu* 'urine culture'; *TK głowy* 'head CT'
Invalid	172	*duży problem* 'big problem'; *norma Piersi* 'norm Breasts' – the words come from two unrelated phrases
Feature	144	*patologiczny* 'patological'; *prawidłowy echogeniczność* 'normal echogenicity'

Table 3 The most frequent labels within the data for the most numerous concept groups

Sem. label	# occ.	Most frequent types
Symptom	477,481	*symptom* 'symptom' (37,315); *ból* 'pain' (28,748); *zmiana* 'change' (26,115); *szmer* 'murmur' (18,052)
Anatomy	420,933	*gardło* 'throat' (28,313); *brzuch* 'abdomen' (20,177); *serce* 'heart' (19,913); *płuco* 'lung' (18,504); *skóra* 'skin' (12,706)
Disease	85,247	*disease* 'choroba' (7,422); *zapalenie* 'inflammation' (3,295); *uraz* 'injury' (3,279); *zmiana patologiczna* 'lesion' (3,233)
Procedure	91,909	*wywiad* 'interview' (17,884); *wizyta* 'visit' (9,189); *kontrola* 'control' (8,718); *stosowanie* 'use' (4,155); *powtórzenie* 'repetition' (3,520)
Physiology	142,487	*ton* 'tone' (12,566); *krążenie* 'circulation' (12,350); *ruch* 'movement' (8,315); *wydzielina* 'secretion' (6,647)
Test	94,244	*RR* 'RR' (29,078); *waga* 'weight' (5,214); *HR* 'HR' (3,542); *temperatura* 'temperature' (3,000); *glukoza* 'glocose' (2,426)
Feature	208,779	*prawidłowy* 'normal' (37,007); *ujemny* 'negative' (23,846); *niebolesny* 'painless' (16,218); *miękki* 'soft' (15,805); *miarowy* 'regular' (11,695)
Negation	146,893	*bez* 'without' (79,726); *nie* 'no/not' (49,363); *neguje* 'negates' (6,144)

domain embeddings on them. An additional advantage of this approach is that the original data cannot be in any way reproduced from the embeddings, which is extremely important in the case of personal medical data. During creation of the term co-occurrence matrix, the description of the whole visit was treated as the neighborhood of the concept. Furthermore, we chose only unique concepts and abandoned their original order in the description (we did this for simplicity).

We computed embeddings of concepts using GloVe (Pennington et al., 2014) for interview descriptions and for examination descriptions separately. Computing two separate embeddings, we aimed to catch the similarity between terms in their specific context. For example, the nearest words to *cough* in the interview descriptions are *runny nose*, *sore throat*, and *fever*, but in the examination description it is *rash*, *sunny*, *laryngeal*.

3.3 Visit embeddings

The simplest way to generate text embeddings based on term embeddings is to use some kind of aggregation of term embeddings, such as an average. This approach was tested, for example, in Banea et al. (2014) and Choi et al. (2016b). In De Boom et al. (2016), the authors computed a weighted mean of term embeddings using the construction of a loss function and training weights by the gradient descent method. We thus firstly computed embeddings of the descriptions (for interview and examination separately) as a simple average of concept embeddings. The final embeddings for visits were then obtained by concatenation of two description embeddings (see Fig. 1).

3.4 Visit clustering

Among many known clustering algorithms (like DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999), BIRCH (Zhang et al., 1996), CLUBS (Masciari et al., 2013)), we decided to use two of the most common: k-means and hierarchical clustering with Ward's method for merging clusters (Ward Jr. 1963). These algorithms cover two different clustering approaches (DBSCAN-based algorithms could be hard to use in this case due to

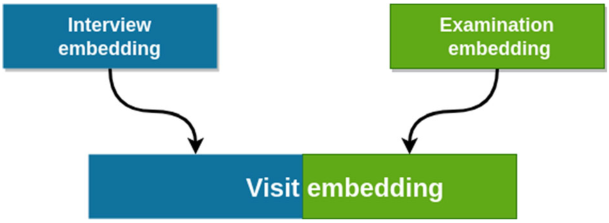


Fig. 1 Visit embedding is the concatenated vector of two descriptions' embeddings which are simple averages of concepts' embeddings

difficulty of choosing ϵ). The algorithms were memory and time efficient, so we did not need to use more advanced methods.

For both clustering algorithms it is crucial to choose a valid distance measure. We decided to use the Euclidean distance between vector representations of visits.

The similarity of the obtained clusterings was measured by the adjusted Rand index (Rand, 1971). For the final results, we chose the hierarchical clustering algorithm due to easier reproducibility of the clustering.

For clustering, we selected visits where the description of a recommendation and at least one of an interview and an examination were not empty (some concepts were recognized). It significantly reduced the number of considered visits. Table 4 gives basic statistics of obtained clusters. The second to last column contains the adjusted Rand index. It can be interpreted as a measure of similarity between two clusterings. The higher similarity measure of obtained clusterings the more consistent are obtained results. Segmentation is an ill-defined problem, often small changes in parameters lead to completely different results, so it is important to verify the stability of the obtained clusterings.

For determining the optimal number of clusters, we considered the number of clusters between 2 and 15 for each specialty. We chose the number of clusters so that adding another cluster did not give relevant improvement of a sum of differences between elements and clusters' centers (according to the so-called *Elbow method*).

Table 4 The statistics of clusters for selected domains

Domain	# clusters	# visits	Custers' size	k-means - hclust	Sil
Cardiology	6	1201	428, 193, 134, 303, 27, 116	0.87	0.27
Dermatology and vener.	6	1204	455, 89, 176, 30, 391, 63	0.64	0.17
Endocrinology	5	1510	389, 412, 208, 183, 318	0.8	0.28
Family medicine	6	11230	3108, 2353, 601, 4518, 255, 395	0.69	0.28
Gynecology	4	3456	1311, 1318, 384, 443	0.8	0.21
Internal medicine	5	6419	1954, 1993, 1343, 874, 255	0.76	0.25
Orthopedics	4	1869	360, 1257, 102, 150	0.19	0.14
Pediatrics	5	4742	1751, 658, 666, 715, 952	0.46	0.14
Psychiatry	5	1012	441, 184, 179, 133, 75	0.81	0.28

The second to last column shows the adjusted Rand index between k-means and hierarchical clustering and the last column is the mean silhouette value

4 Results

4.1 Analogies in medical concepts

To better understand the structure of concept embeddings and to determine the optimal dimension of embedded vectors, we used the word analogy task introduced in Mikolov et al. (2013) and examined in a medical context in Newman-Griffis et al. (2017). In the former work, the authors defined five types of semantic relationship and nine types of syntactic relationship.

We proposed our own relationships between concepts that was more closely related to the medical language. We exploited the fact that we had a lot of multiword concepts in the corpus and very often the same words were included in different terms. We would like the embeddings to be able to catch relationships between terms. A question in the term analogy task is the computing of a vector: $vector(left\ foot) - vector(foot) + vector(hand)$ and checking if the correct $vector(left\ hand)$ is in the neighborhood (measured as the cosine of the angle between the vectors) of this resulting vector.

We defined seven types of such semantic questions and computed accuracy of the answers in a similar way as in Mikolov et al. (2013): we manually created a list of similar term pairs and then we formed a list of questions by taking all two-element subsets of the pairs list. Table 5 shows the created categories of questions.

We 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). We checked if the embeddings of such words are similar.

We computed the embeddings for terms occurring at least 5 times in the descriptions of the selected visits. The number of chosen terms in the interview descriptions was 3,816 and 3,559 in the examination descriptions. Among these were 2,556 common terms for interview and examination. Embeddings of a size from 10 to 200 were evaluated. For every embedding of interview terms, the accuracy of all eight tasks was measured. Table 6 shows the mean of eight task results. The second column includes the results of the most restrictive rule: a question is assumed to be correctly answered only if the closest term of the vector computed by operations on related terms is the same as the desired answer. The total number of terms in our data set (about 900,000 for interviews) was many times lower than sets examined in Mikolov et al. (2013). Furthermore, because the language in medical free-text records is very specific, we do not know if fulfilling all the analogies is possible. Taking this into account, the accuracy of about 0.17 is very high and better than we expected. We then

Table 5 The categories of questions in term analogy task with example pairs

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 6 Mean accuracy of correct answers on term analogy tasks

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

Rows show different embedding sizes and columns correspond to the size of neighborhoods. The highest value for each column is bolded

checked the closest 3 and 5 words to the computed vector and assumed a correct answer if the correct vector was there. In the biggest neighborhood the majority of embeddings returned an accuracy higher than 0.5.

For computing visit embeddings, we chose term embeddings of dimension 20, since this resulted in the best accuracy of the most restrictive analogy task and it allowed us to perform more efficient computations than higher dimensional representations. Figure 2 illustrates the PCA projection of term embeddings from four categories of analogies.

4.2 Comparison with pretrained embeddings

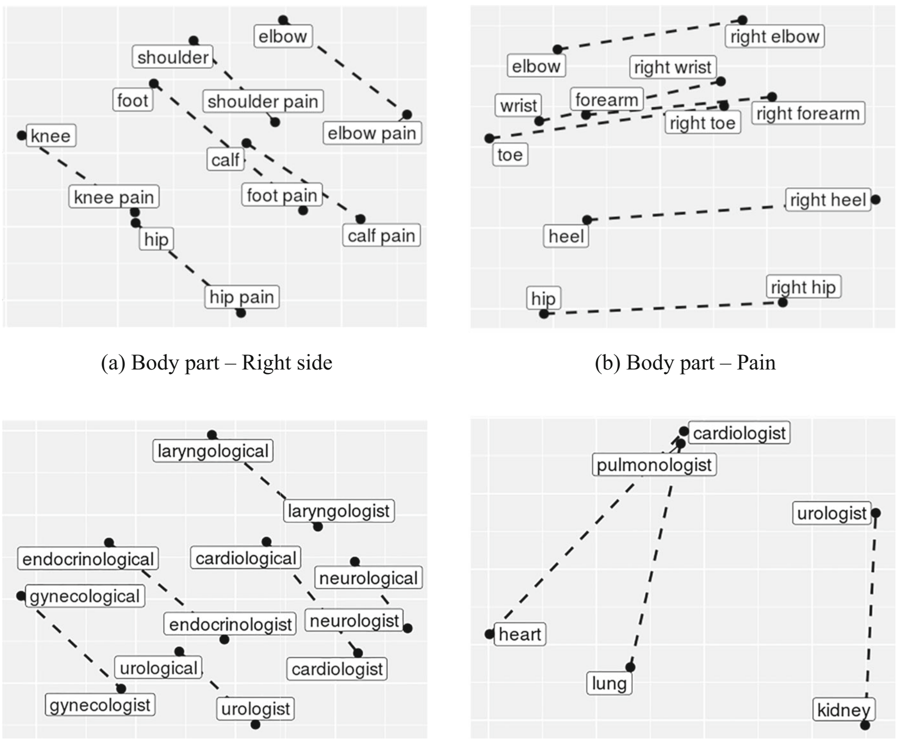
We compared some of our embeddings with two sets of embeddings obtained in Pennington et al. (2014) on Wikipedia 2014 + Gigaword 5 and Common Crawl corpora (hence, not trained on medical data). Despite having a many times lower corpus of texts (Table 7), we obtained better analogies on the example pairs of related medical terms, compare Figs. 2 and 3. A comparison with all our term analogy tasks was impossible because pretrained embeddings do not contain multiword expressions.

4.3 Visit clustering

Clustering was performed separately for each specialty of doctors. Figure 4 illustrates two-dimensional t-SNE projections of visit embeddings colored by clusters (Maaten & Hinton, 2008). Some domain clusters are very clear and separated (Fig. 4a and i). This corresponds with the high stability of the clustering measured using the Rand index.

The first method for evaluating the quality of clusters was computing silhouette values (Rousseeuw, 1987). The mean silhouette values for all visits from domains are shown in the last column of Table 4. For all medical domains, the mean silhouette is markedly greater than 0, which suggests that the obtained clusters are not accidental but result from the characteristics of visit descriptions. An example of a more detailed insight into silhouettes for internal medicine is shown in Fig. 5.

We also evaluated how clear derived segments are when it comes to medical diagnoses (ICD-10). No information about recommendations or diagnosis was used in the phase of clustering to prevent data leakage. To find similarities between clusters and ICD-10 codes,



(a) Body part – Right side

(b) Body part – Pain

(c) Specialty – Adjective

(d) Specialty – Body part

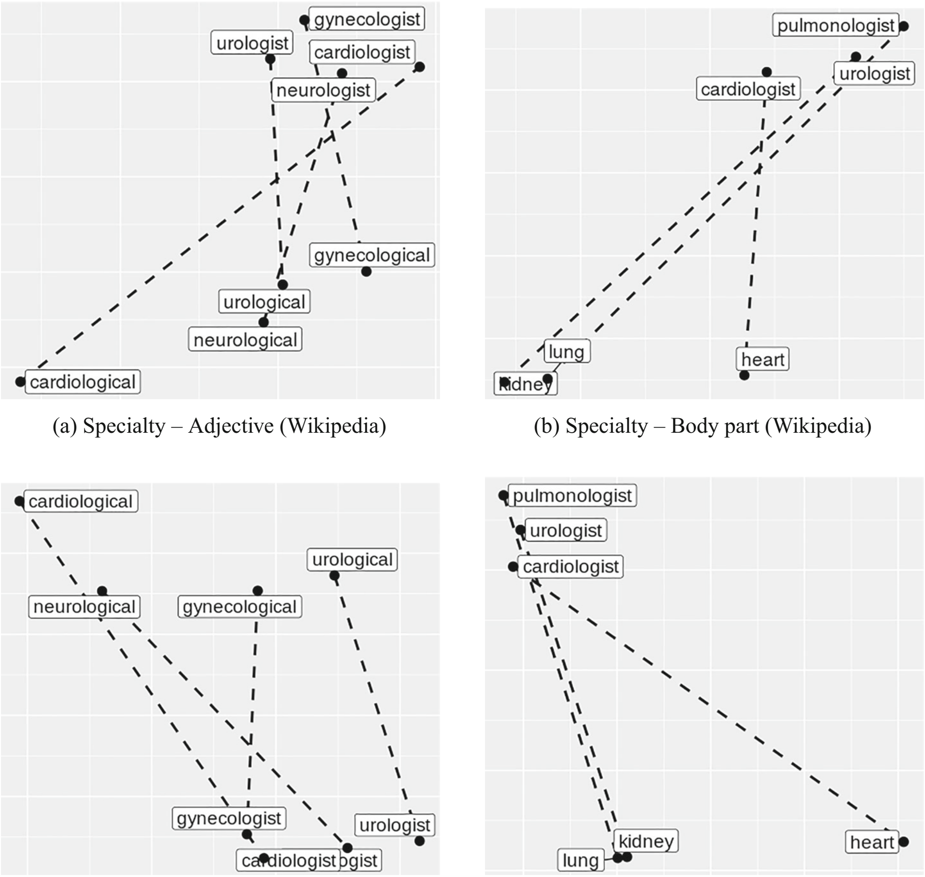
Fig. 2 Visualization of analogies between terms. The pictures show term embeddings projected into 2d-plane using PCA. Each panel shows a different type of analogy. See Dobrakowski et al. (2020)

we look at correspondence analysis (CA) plots. Figure 6a shows the CA plot between clusters and ICD-10 codes for family medicine clustering. Two large groups of codes appeared: 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 the Z76 ICD-10 code (encounter for issuing a repeat prescription).

In the clustering of gynecology (Fig. 6b), we also have two groups: the diseases of the genitourinary system (N), connected with Clusters 1 and 3; and pregnancy, childbirth

Table 7 Comparison of our embeddings with two sets of embeddings from Pennington et al. (2014)

Dataset	#tokens	#vocabulary	Vector dim.
Interviews descriptions	900K	3816	20
Examinations descriptions	1.1M	3559	20
Wikipedia 2014 + Gigaword 5	6B	400K	300
Common Crawl	840B	2.2M	300



(a) Specialty – Adjective (Wikipedia)

(b) Specialty – Body part (Wikipedia)

(c) Specialty – Adjective (Common Crawl)

(d) Specialty – Body part (Common Crawl)

Fig. 3 PCA visualization of analogies between terms for GloVe pre-trained embeddings

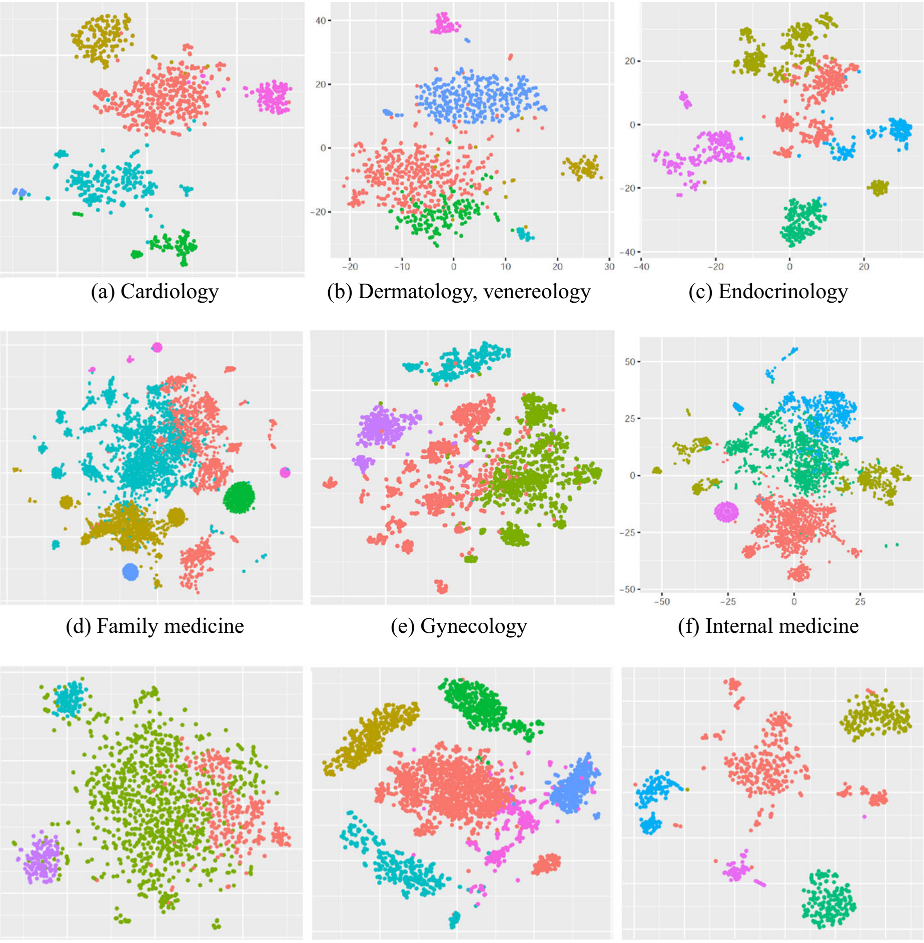
and the puerperium (O), connected with Cluster 2. The presented methodology therefore allowed us to obtain groups of visits with similar diagnoses expressed in ICD-10 codes.

We also examined the distribution of doctors' IDs in the obtained clusters. It turned out that some clusters almost exactly covered the descriptions written by one doctor. This happened in the specialties where clusters were separated by large margins (e.g. psychiatry, pediatrics, and cardiology). Figure 7 shows correspondence analysis between doctors' IDs and clusters for psychiatry clustering.

4.4 Recommendations in clusters

According to the main goal of our clustering described in the Introduction, we would like to obtain similar recommendations inside every cluster. We therefore examined the frequency of occurrence of the recommendation terms in particular clusters.

We examined terms of recommendations related to one of five categories: procedure to be carried out by patient, examination, treatment, diet and medicament. Table 8 shows an



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

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

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

Fig. 4 Clusters of visits for selected domains. Each dot corresponds to a single visit. Colors correspond to segments. Visualization created with t-SNE. It is an extended version of Fig. 2 presented in Dobrakowski et al. (2020)

example of an analysis of the most common recommendations in clusters in gynecology clustering. In order to find only characteristic terms for clusters we filtered the terms which belong to one of the 15 most common terms in at least three clusters.

5 Software for interpretable segmentation of medical free-text records

The methods presented in this paper are implemented in the `memr` package for `R` (R Core Team, 2020). The name is an acronym for Multisource Embeddings for Medical Records. The package can be installed from the GitHub repository https://github.com/MI2DataLab/memr available under MIT license.

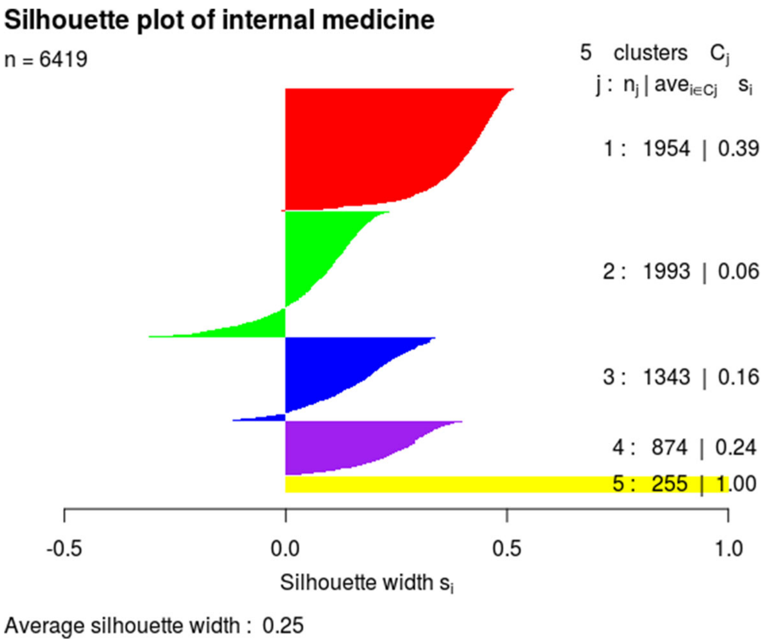


Fig. 5 Silhouette plot for internal medicine clustering. Most visits are well fitted to their clusters

The package allows for creating embeddings of medical free-text records written by doctors and provides a wide spectrum of tools for data visualization and segmentation of medical visits. These tools are intended to develop computer-supported medicine by facilitating medical data analysis and interpretation. The package can be exploited for many applications such as the recommendation prediction, patients' clustering etc. that can aid doctors in their practice.

The core function in this package is `embed_terms()` which creates medical term embeddings based on the GloVe algorithm implemented in the `text2vec` (Selivanov & Wang, 2018) package. To validate the quality of computed embeddings one can run the visual *word analogy task* with the function `visualize_analogies()`. It produces PCA plots (with the `ggplot2` package (Wickham, 2016)) of given pairs of terms. Examples of resulting plots are shown in Figs. 2 and 3.

Function `embed_list_visits()` aggregate embeddings created for various types of input into a single embedding for a visit. If we have more information about visits, such as doctors' specialties or ICD-10 codes, `memr` can help with data analysis and visualization. We can perform visit clustering of a specified doctor's specialty by the k-means algorithm with the function `cluster_visits()`. With the `visualize_visit_embeddings()` function we can visualize the visits by the t-SNE algorithm (Maaten & Hinton, 2008) with the use of the `Rtsne` package (Krijthe, 2015). The resulting plot is similar to Fig. 4. Using the recommendations in our data, we can show the most popular recommendations for each cluster using the function `get_cluster_recommendations()`.

The `memr` package also allows for visualization of ICD-10 codes. For every ICD-10 code the function `visualize_icd10()` computes an average of embeddings of all visits assigned by the doctor to this code and plot t-SNE visualization of embeddings. The

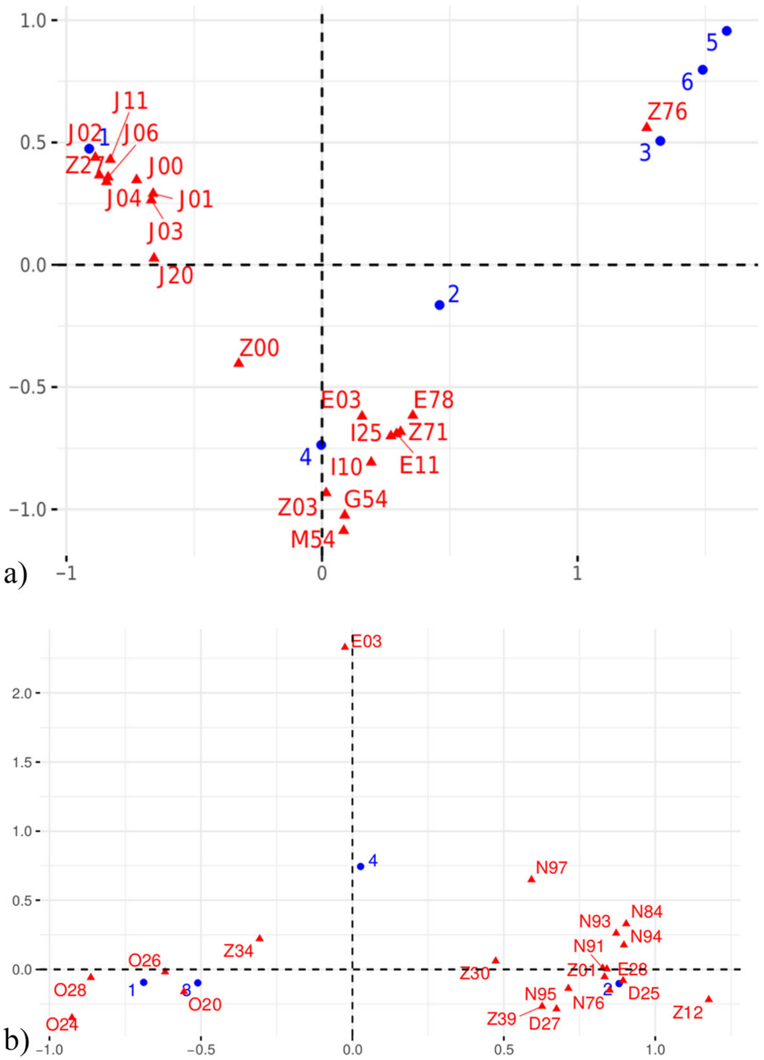


Fig. 6 Correspondence analysis between clusters and ICD-10 codes for gynecology clustering (panel a) and for family medicine clustering (panel b). Similar ICD codes are grouped near the same clusters. See Dobrakowski et al. (2020)

resulting plot is shown in Fig. 8. To sum up, the open-source package `memr` facilitate computations and analysis of visits embeddings.

6 Conclusions, applications and future works

In this paper we proposed a new method for clustering of visits in health centers based on descriptions written by doctors. The method is implemented in the `R` open-source package `memr`. We validated the method on a large corpus of Polish free-text medical records. We

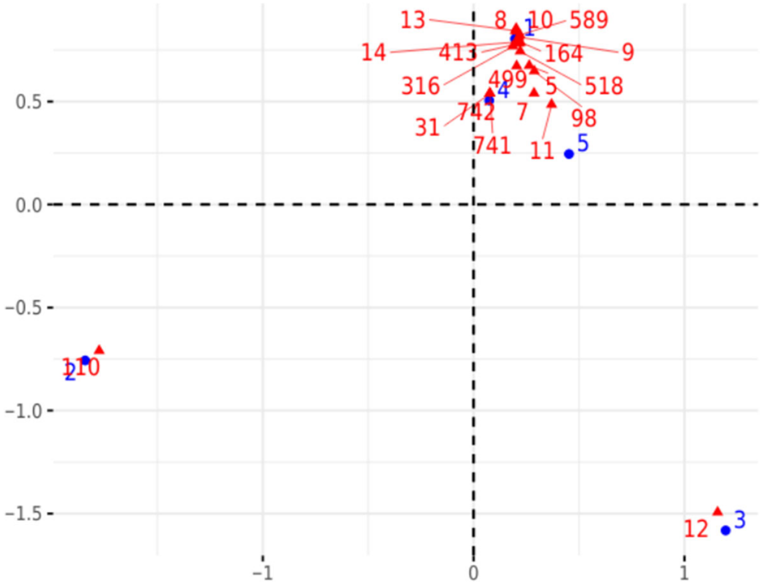


Fig. 7 Correspondence analysis between clusters and doctors' IDs for psychiatry clustering. Clusters 2 and 3 are perfectly fitted to a single doctor

identified important medical concepts in the corpus and converted texts into sets of semantic labels. For languages for which SNOMED CT is available it is possible to skip the step of terminology extraction and to use existing resources.

The visit embedings are based on concept embeddings created with the GloVe algorithm. The quality of the embeddings was measured and confirmed by the analogy task designed specifically for this corpus. It turns out that analogies work well (over 60% of analogies are fulfilled when we look at the 5 closest terms), which ensures that concept embeddings store some useful information.

Clustering was performed on the embedding of visits created based on word embedding, so the original texts which may have included sensitive data were unnecessary. Visual and numerical examination of derived clusters showed an interesting structure among visits.

Table 8 The most common recommendations for each segment derived for gynecology

Cluster	Size	Most frequent recommendations
1	1311	recommendation (16.6%), general urine test (5.6%), diet (4.1%), vitamin (4%), dental prophylaxis (3.4%)
2	1318	therapy (4.1%), to treat (4%), cytology (3.8%), breast ultrasound (3%), medicine (2.2%)
3	384	acidum (31.5%), the nearest hospital (14.3%), proper diet (14.3%), health behavior (14.3%), obstetric control (10.2%)
4	443	to treat (2%), therapy (2%), vitamin (1.8%), diet (1.6%), medicine (1.6%)

The percentage of visits in this cluster which contain a specified term are shown in brackets. We skipped terms common in many clusters, such as: *treatment, ultrasound treatment, control, morphology, hospital, lifestyle,* and *zus* (Social Insurance Institution)

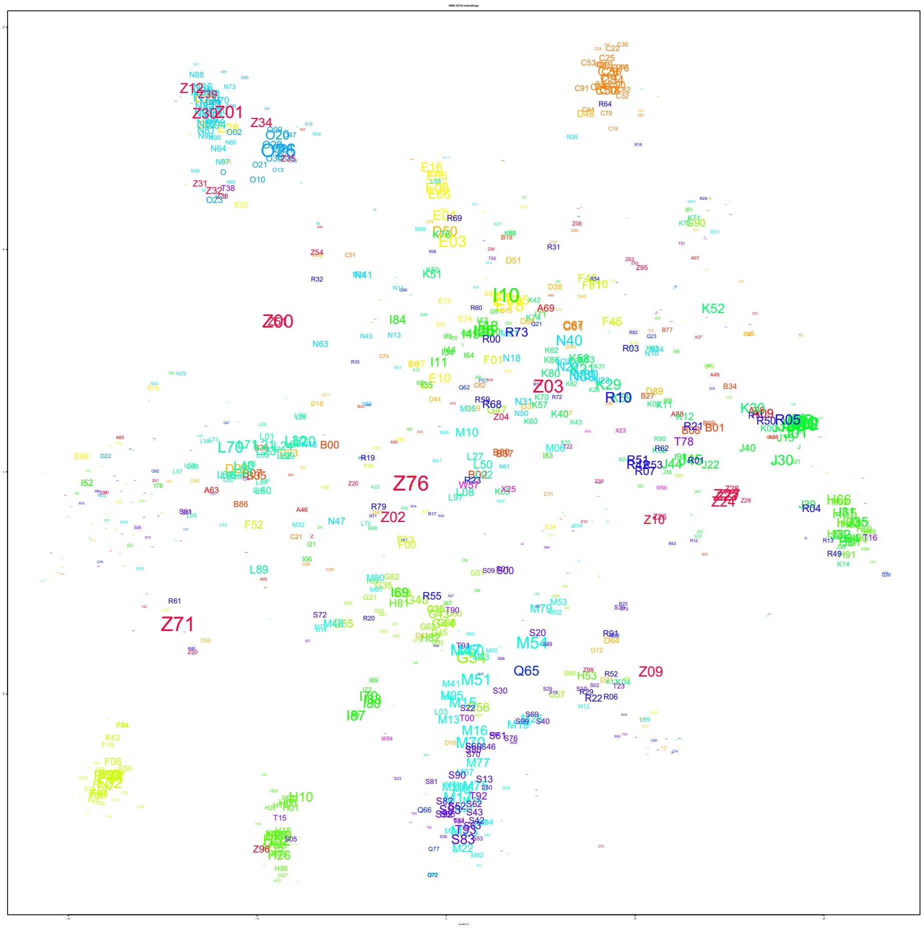


Fig. 8 Map of ICD-10 codes in the space of embeddings for all visits. More popular codes have larger labels. To see the map in higher resolution please visit https://github.com/MI2DataLab/memr/blob/master/ICD10_embeddings.pdf

As we have shown, the obtained segments were linked with medical diagnosis, even when the information about recommendations or diagnosis was not used for the clustering. This additionally convinced us that the identified structure was related to some subgroups of medical conditions.

The obtained clustering have many applications. For example, they can be used to assign new visits to already derived clusters. Based on descriptions of an interview or a description of patient examination, we can identify similar visits and show corresponding recommendations.

In the future work it could be valuable to investigate varying interrelationships among the identified clusters. As we have said in the Introduction, a single patient can belong to several clusters when there are many visits related to the patient. We could use the information about succeeding visits and join this information with the clusters. This way maybe we could uncover sequential relationships among some clusters. When we have these, we will be able to output the subsequent cluster of a new visit to predict the progression of a disease or a treatment.

Author Contributions A.G.D. designed and implemented the methodology for clustering of visits and the term analogy task, and performed the visualization of the results. A.M., M.M and W.J. designed and implemented tools for extraction of medical concepts. P.B. advised on the project. All authors contributed to the writing.

Funding Open access funding provided by University of Warsaw. This work was financially supported by the National Centre for Research and Development in Poland, Grant POIR.01.01.01-00-0328/17. PBi was supported by The National Science Centre in Poland, Opus grant 2017/27/B/ST6/0130t.

Availability of data and material (data transparency) Data is not available due to its high sensitivity.

Code Availability The methodology is implemented into the `R` open-source package `memr` (Multisource Embeddings for Medical Records). The package is available at GitHub https://github.com/MI2DataLab/memr under MIT licence.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: Ordering points to identify the clustering structure. *ACM Sigmod Record*, *28*(2), 49–60.

Apostolova, E., Channin, D. S., Demner-Fushman, D., Furst, J., Lytinen, S., & Raicu, D. (2009). Automatic segmentation of clinical texts. In *Proceedings of the 31st annual international conference of the IEEE engineering in medicine and biology society* (pp. 5905–5908). IEEE.

Banea, C., Chen, D., Mihalcea, R., Cardie, C., & Wiebe, J. (2014). Simcompass: Using deep learning word embeddings to assess cross-level similarity. In P. Nakov, & T. Zesch (Eds.) *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 560–565). ACL and Dublin City University.

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, *19*(84), 1–5.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl_1), D267–D270.

Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to train good word embeddings for biomedical NLP. In *Proceedings of BioNLP* (pp. 166–174).

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J., & Sun, J. (2016a). Multi-layer representation learning for medical concepts. In *KDD '16: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1495–1504). ACM.

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

Choi, Y., Chiu, C. Y.-I., & Sontag, D. (2016c). Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science* pp. 41–50.

Dalianis, H. (2018). *Clinical text mining*. New York: Springer.

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

De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., & Bruza, P. (2014). Medical semantic similarity with a neural language model. In *Proceedings of CIKM* (pp. 1819–1822). ACM.

Dobrakowski, A. G., Mykowiecka, A., Marciniak, M., Jaworski, W., & Biecek, P. (2020). Interpretable segmentation of medical free-text records based on word embeddings. In D. Helic, G. Leitner, M. Stettinger, A. Felfernig, & Z. Raś (Eds.) *Foundations of intelligent systems. ISMIS Lecture Notes in Computer Science*, (Vol. 12117 p. 2020). Cham: Springer.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96: proceedings of the second international conference on knowledge discovery and data mining* (pp. 226–231). AAAI Press.

Fetter, R. B., Shin, Y., Freeman, J. L., Averill, R. F., & Thompson, J.D. (1980). Case mix definition by diagnosis-related groups. *Medical Care*, *18*(2), (iii):1–53.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value Method. *International Journal on Digital Libraries*, *3*, 115–130.

Ganesan, K., & Subotin, M. (2014). A general supervised approach to segmentation of clinical texts. In *IEEE international conference on big data* (pp. 33–40).

Gordon, L., Grantcharov, T., & Rudzicz, F. (2019). Explainable artificial intelligence for safe intraoperative decision support. *JAMA Surgery*, *154*(11), 1064–1065.

Jaworski, W., & Kozakoszczak, J. (2016). ENIAM: Categorial syntactic-semantic parser for Polish. In *Proceedings of COLING* (pp. 243–247).

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

Kobylińska, K., Mikołajczyk, T., Adamek, M., Orłowski, T., & Biecek, P. (2019). Explainable machine learning for modeling of early postoperative mortality in lung cancer. In *Artificial intelligence in medicine: Knowledge representation and transparent and explainable systems* (pp. 161–174). Springer.

Krijthe, J. H. (2015). Rtsne: T-distributed stochastic neighbor embedding using barnes-hut implementation. R package version 0.13.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Marciniak, M. (2015). Domain corpora as a source of information, volume 4 of Monograph Series. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Marciniak, M., Mykowiecka, A., & Rychlik, P. (2016). TermoPL — a flexible tool for terminology extraction. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (pp. 2278–2284). Portorož, Slovenia: ELRA.

Masarie, F. E., & Miller, R. A. (1987). Medical subject headings and medical terminology: An analysis of terminology used in hospital charts. *Bulletin of the Medical Library Association*, *2*(75), 89–94.

Masciari, E., Mazzeo, G. M., & Zaniolo, C. (2013). A new, fast and accurate algorithm for hierarchical clustering on euclidean distances. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 111–122). Springer.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Minarro-Giménez, J. A., Marin-Alonso, O., & Samwald, M. (2014). Exploring the application of deep learning techniques on medical text corpora. *Studies in Health Technology and Informatics*, *205*, 584–588.

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

Névéol, A., Dalianis, H., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantic*, *9*(12), 1–13.

Orosz, G., Novák, A., & Prószéky, G. (2013). Hybrid text segmentation for Hungarian clinical records. In *Proceedings of MICAI* (pp. 306–317). Springer.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP* (pp. 1532–1543).

R Core Team (2020). R: A Language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

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

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Ruffini, M., Gavaldà, R., & Limón, E. (2017). Clustering patients with tensor decomposition. arXiv:1708.08994.

Selivanov, D., & Wang, Q. (2018). text2vec: Modern Text Mining Framework for R, R package version 0.5.1.

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 COLING* (pp. 2789–2804).

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). Birch: an efficient data clustering method for very large databases. *ACM Sigmod Record*, *25*(2), 103–114.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Adam Gabriel Dobrakowski[1] · **Agnieszka Mykowiecka**[2] · **Małgorzata Marciniak**[2] · **Wojciech Jaworski**[1] · **Przemysław Biecek**[1,3]

Adam Gabriel Dobrakowski
ad359226@students.mimuw.edu.pl

Agnieszka Mykowiecka
agn@ipipan.waw.pl

Małgorzata Marciniak
mm@ipipan.waw.pl

Wojciech Jaworski
W.Jaworski@mimuw.edu.pl

[1] University of Warsaw, Banacha 2, Warsaw, Poland

[2] Institute of Computer Science Polish Academy of Sciences, Jana Kazimierza 5, Warsaw, Poland

[3] Warsaw University of Technology, Koszykowa 75, Warsaw, Poland