Funkner *et al. Appl Netw Sci* (2021) 6:53
https://doi.org/10.1007/s41109-021-00395-2

Applied Network Science

RESEARCH Open Access

Citywide quality of health information system through text mining of electronic health records

Anastasia A. Funkner[1*], Michil P. Egorov[1], Sergey A. Fokin[2], Gennady M. Orlov[1,3] and Sergey V. Kovalchuk[1]

*Correspondence:
funkner.anastasia@itmo.ru
[1] ITMO University, Saint Petersburg, Russia
Full list of author information is available at the end of the article

Abstract

A system of hospitals in large cities can be considered a large and diverse but interconnected system. Widely applied in hospitals, electronic health records (EHR) are crucially different from each other because of the use of different health information systems, internal hospital rules, and individual behavior of physicians. The unstructured (textual) data of EHR is rarely used to assess the citywide quality of healthcare. Within the study, we analyze EHR data, particularly textual unstructured data, as a reflection of the complex multi-agent system of healthcare in the city of Saint Petersburg, Russia. Through analyzing the data collected by the Medical Information and Analytical Center, a method was proposed and evaluated for identifying a common structure, understanding the diversity, and assessing information quality in EHR data through the application of natural language processing techniques.

Keywords: Health information system, Electronic health record, Unstructured data, Natural language processing, Data completeness, Machine learning

Introduction

A system of hospitals in a big city can be considered a large and diverse but interconnected system. The interconnection of hospitals and physicians comes from operating in a common information space, same legislative environment, and providing health services to the population of the same city. Currently, electronic health records (EHR) (Nguyen et al. 2014) are widely adopted in healthcare organizations and provide improving consistency and interoperability of health-related information. Citywide EHR integration enables the implementation of large-scale analytical and clinical projects (see NYC Macroscope Newton-Dame et al. 2016 as an example). At the same time, in large cities there exist such factors as multiple levels of healthcare regulation (from government to hospital-level authorities) and diversity in clinicial staff experience, individual approaches, and patterns in clinical decision making, etc. This diversity is much more pronounced if an EHR is implemented in different health information systems (HIS) deployed in hospitals. In many cases, a significant portion of the information in an EHR is stored in the unstructured (textual) form. For example, anamnesis, diagnosis, epicrisis,

surgery protocols, and conclusions may be stored in such form. At the same time, this part of the information is often an important source for systematic analysis of health service quality. In such a situation, understanding and assessing complex health information operated in a citywide healthcare system can be quite challenging.

Within the presented research, we focused on the analysis and structuring of EHRs from hospitals in Saint Petersburg, Russia. The data was provided by the Medical Information and Analytical Center (MIAC)[1] responsible for monitoring and assessing the quality of healthcare service in the city. We analyze the EHR data, particularly textual unstructured data, as a reflection of the complex multi-agent system of healthcare in the city of Saint Petersburg. Such data may contain a lot of additional information that can not be displayed in a structured form (numbers, codes, enumerated concepts). Thus, this data falls out of the review by both analysts and doctors. However, unstructured data can be used to analyze the systemic quality in the city healthcare: to identify sources of uncertainty, build information-behavioral profiles of doctors, and assess influencing factors, the MIAC acts like a distributed heterogeneous information system. It contains people (doctors), individual HISs, and individual healthcare facilities. All this is connected implicitly though the general population of patients and the legal field, in which doctors work. Each "information agent" acts relatively independently, filling the system with information in the process of serving the patient flow. This way, we can watch this variety of connections through EHRs, the quality of which we strive to evaluate.

Within the presented study, we propose and elaborate an approach to unify EHR data to improve their structure using natural language processing (NLP) techniques. The approach is considered as a way to work automatically with diverse data (i.e., coming from different hospitals and HISes) to obtain and assess the implicit structure presented within the data. The approach may be used to structure the diverse data, improve the analysis and assessment procedures of large heterogeneous healthcare systems existing in big cities through the EHR data collection. Such improvement may increase the quality of both analytical procedures (e.g., implemented by the MIAC in Saint Petersburg) and hospital-level EHR interoperability characteristics.

The structure of the paper is as follows. The next section provides an overview of the related works in EHR structuring and quality analysis. A case study of citywide healthcare quality analysis in Saint Petersburg and the dataset used are described in "Case study" section. Next, the proposed method and implementation details are provided in "Implementation details" section. The obtained results are presented and analyzed in "Results" section. Finally, "Discussion" and "Conclusion and future work" sections provide a discussion and concluding remarks of the study respectively.

Related works

Structuring and assessing a medical text to determine its completeness and search for an ideal structure are still highly relevant to most hospitals and healthcare institutions.

In their research, Weiskopf and Weng (2013) provide a review of methods to assess the quality of EHRs. They define five criteria of quality (completeness, correctness,

[1] https://spbmiac.ru/ (in Russian).

concordance, plausibility, and currency) and seven methods that help to check EHRs according to one or more criteria: comparison with gold standards, data element agreement, data source agreement, distribution comparison, validity checks, log review, and element presence. Similar criteria (accuracy, correctness, validity, completeness, timeliness, usefulness, etc.) are identified by (St-Maurice and Burns 2017). One of the most popular methods is gold standard compliance. Based on this paper, other researchers expand the list of criteria. For example, Batini collects criteria from many papers and there are new ones among them: usefulness, cost-effectiveness, and confidentiality (Batini and Scannapieco 2016). However, it is complicated to interpret some of the criteria for assessing textual data. For our research, we aim to construct a gold standard based on a large amount of data and the doctors' experience that is invested in them and then to compare new records with this gold standard using the methods of data element agreement and element presence.

Another approach is to check for the presence of certain records in EHRs to assess their completeness and relevance. For example, van der Bij et al. (2017) estimate such parameters as a percentage of episodes that have a "meaningful" ICPC code, percentage of drugs linked to an episode of care, and others. Burke et al. identify 12 structures in EHR to assess its quality (Burke et al. 2014). In our case, we plan to check for the presence of certain information within one record. There are other methods for assessing the structure and completeness of records. Logan et al. (2001) try to find an acceptable way of EHR recording and assess the completeness and correctness by comparing the video of a patient's a doctor's meeting with the EHR data and their structure. Often, studies aiming to find the most complete and accurate format are based on surveys of a small number of doctors (Williams 2003). It is possible to develop such a model to extract specific information (Wang et al. 2012; Yehia et al. 2019). However, it is necessary to define such a list of questions and entities for each record type manually. So, it is less applicable for various real-world records that have accumulated in many HISs and hospitals.

Often, data semantics are presented in the context of data interoperability to transfer data between different MISs and clinical applications. Nguen et al. conducted a systematic review of EHR implementation with an assessment of information systems with DeLone and McLean's framework (Nguyen et al. 2014). Sun et al. present the architecture of their semantic processing approach where data is transmitted through the semantic layer with clinical ontologies inside (Sun et al. 2015). Most solutions for data interoperability are based on ontologies (Sun et al. 2015; Roberts and Demner-Fushman 2016; Freedman et al. 2020; Kersloot et al. 2020). Moreover, Kersloot et al. (2020) show the statistics for mapping clinical text fragments to ontology concepts that are described in reviewed papers. They conclude that 88% of the studies do not present any validation. Moreover, nowadays it is common to produce semantic interoperability between different MISs and clinical databases using developed libraries for database matching (Bruland et al. 2017). Also, semantic assessing and overview are often presented for academic texts (scientific papers, articles and books) that have specific preprocessing and methods for formal language that are much different from clinical records (Datta et al. 2019).

Another approach is not to try to structure the records, but to extract specific knowledge on demand. For example, Lamy et al. (2019) show an example of a pipeline for finding and extracting the necessary information for the Portuguese language: Portuguese

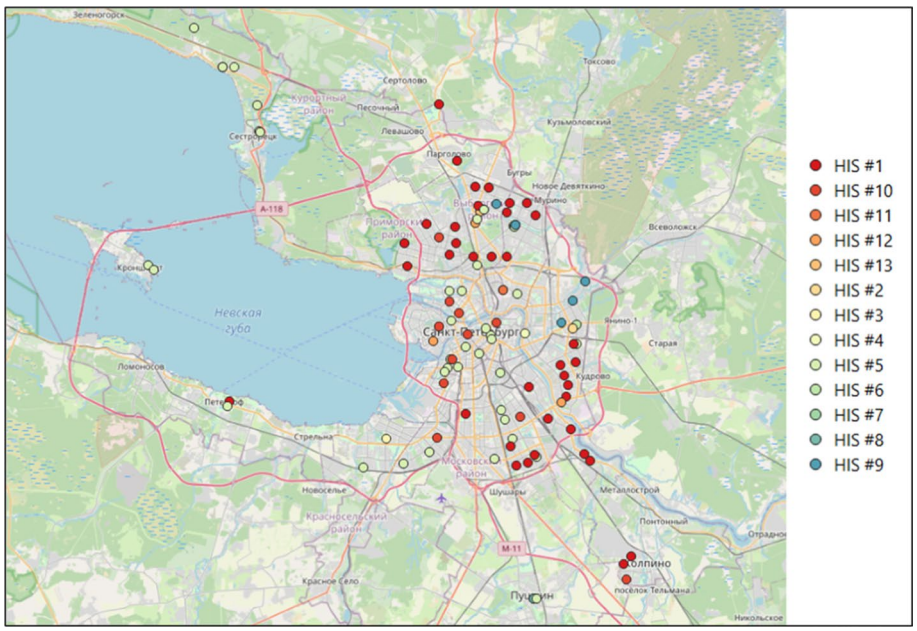


Fig. 1 Distribution of EHR providers (healthcare organizations) in Saint Petersburg with different HISs

EHRs are translated into English, and then ready-made and already well-proven tools for English are used. Recent studies often include machine learning approaches to retrieve information from EHRs and assess the completeness of the records. This way, works by Tang et al. (2013), Funkner and Kovalchuk (2020) investigate NLP within the task of reconstruction of temporal structures and events from EHRs.

Most of the authors conclude the importance of structuring and analysis of EHRs to improve the quality, interoperability, and integrability of both information and health service. One of the most important problem is improvement of structure and interoperability (as a consequence) of EHR data. Within our study, we focus on understanding the structural diversity and possible interpretation of EHRs in the healthcare system of large cities through the analysis of unstructured (free-form text) parts of EHRs collected from hospitals. The presented approach is aimed towards the automatic (unsupervised or semi-supervised) structuring procedures that can work for EHR weakly-structured data without predefined domain-specific unified structures, dictionaries, and semantics. Also, an important advantage of such an approach is possible translation to low-resource languages where domain-specific NLP tools arent' presented well.

Case study

For this study, we consider a set of 79,234 depersonalized records of patients with arterial hypertension (AH) and acute coronary syndrome (ACS) who applied to medical centers in St. Petersburg, Russia, in 2020. The data was collected by the MIAC for the analysis of EHR and health service quality. The records were provided by 107 institutions using 13 different HISs (Fig. 1). The selection of HISs was developed, provided, and supported by different vendors. Each HIS has its architecture and user interface. Thus, the common practice of EHR input varies significantly.

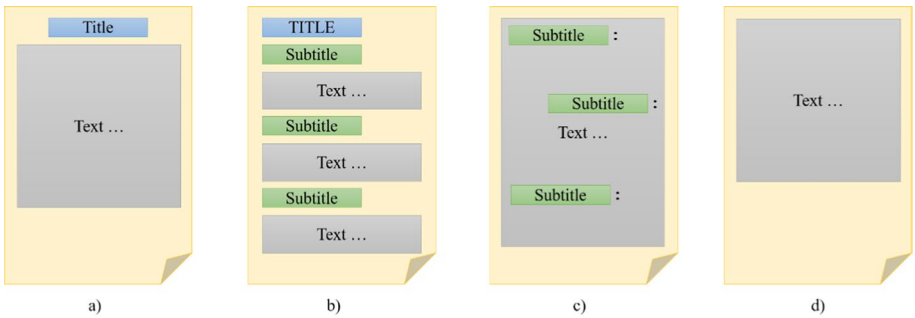


Fig. 2 Different forms of unstructured textual EHR data

The structured EHR data collected by the MIAC is widely used for monitoring and analytical purposes as well as for centralized development and regulation of informatization of the city healthcare system. However, even though the textual data in EHRs contains important information on the provision of health services, use of such data is significantly limited due to the lack of structuring and diversity in format. The practical goals of this study are analysis, structuring, and quality assessment of the EHRs. The results may be used by the MIAC to improve their analytical facilities as well as by hospitals to support better information processing, interoperability, and clinical decision support in their HISs.

Commonly, unstructured EHR data in Russian practice is a natural language text that contains many specific medical terms, abbreviations, words in Latin and, less often, English (names of equipment, drugs). Unfortunately, raw text often contains typos and other distortions caused by the data transfer between information systems (connected words, lack of separators between sentences, HTML and XML tags). Such texts can have some structural features: they contain subheadings or field names separated by colons inside them (see examples of possible textual EHR data structure in Fig. 2) Most often, an unstructured text presents a patient's life and illness history, discharges, records of consultations, and less often protocols of operations and other medical procedures. For example, records can have a title 'Anamnesis' and its text (Fig. 2a), or title 'Protocols of operation' and subtitles 'Type of the operation', 'Duration of the operation', etc. (Fig. 2b), or does not have any title inside, but has incorporated subtitles as 'Diagnosis', 'Vital signs', etc. (Fig. 2c). Also, records can be totally unstructured without any indicated titles and subtitles (Fig. 2d). However, each of these records has its format and features of the language structure that depend on the medical center and the HIS. The above problems have a critical impact on the speed and ability to automatically process such texts. It is also worth noting that the data is presented in Russian, which narrows down the range of available tools for language processing.

Implementation details

General processing phase

Currently, there is a lack of ready-to-go technologies available for domain-specific medical text analysis in a language other than English (Névéol et al. 2018). For the last year, our research team has been developing a set of tools for automatic processing of

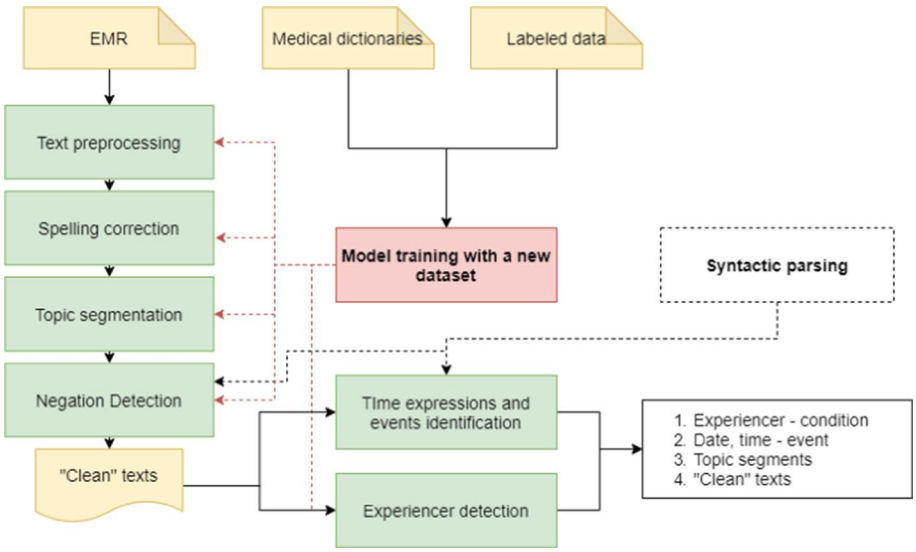


Fig. 3 NLP modules for processing medical texts in Russian

medical texts in Russian. These tools are implemented as extensible Python modules aimed at processing various types of medical texts. Currently, there are five tools at different stages of development (Fig. 3): spelling correction, negation detection for diseases, extracting the experiencer of the disease, topic segmentation, and extraction of temporal structures and events (Balabaeva and Kovalchuk 2020; Balabaeva et al. 2020; Funkner and Kovalchuk 2020; Shaikina and Funkner 2020; Funkner et al. 2020). Each tool solves a specific problem or helps with text preprocessing, but none of them determine the general structure of the text.

The current study uses and extends the implemented software for NLP, structuring textual data, and identifying basic elements of EHRs.

Methodology

This section describes methods for processing records, structuring them, and assessing their quality. Figure 4 shows the three main stages of record processing: identifying the type of record, substructure recognition, and assessing the quality. At each stage, models are trained (blue elements in Fig. 4), which can be used for new records. Besides, Fig. 4 shows the topics and record formats (yellow elements), which are defined in the training dataset and can be easily interpreted by a specialist.

EHR type detection

Within our study, we consider processing a dataset containing EHRs collected from different medical centers with different HISs. The first step in structuring such heterogeneous records is to identify the type of record (consultation with a doctor, test results, surgery protocol, etc.). Since records are collected in a large number of healthcare organization with their own rules and practices, the same type of record may have different names.

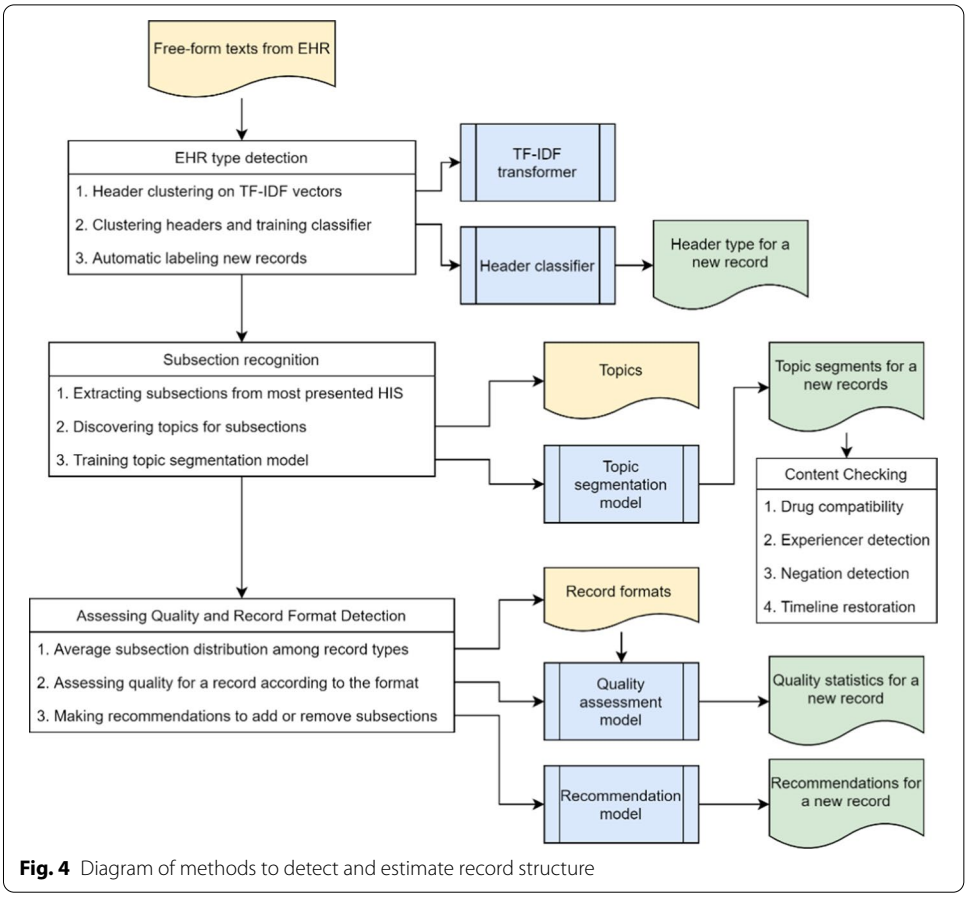
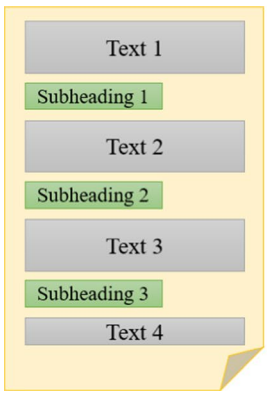


Fig. 4 Diagram of methods to detect and estimate record structure

We propose to use the classic approach for grouping text titles: preprocessing (removing extra characters, lemmatization), removing stop words (mainly prepositions, pronouns, conjunctions), TF-IDF (term frequency, inverse document frequency) transformation to reduce the weight of background words (for example, the word doctor or hospital in this context is a background, but in general are not stop words), clustering of TF-IDF vectors. With this study, we use only record titles to identify a record type. However, it is possible to add extra features of records, but model training complexity will increase.For clustering, there are many appropriate methods: hierarchical clustering can show the nesting of clusters one into another when using different thresholds, the k-means method is easily interpreted in terms of vectors and the "central" record can be found (closest to the cluster center), with which other records of the cluster are compared. We propose to use the OPTICS method, which allows finding clusters in the feature space based on density (Ankerst et al. 1999). After some experimentation, we notice that names of some types of records do not differ much between medical centers (for example, discharge reports). At the same time, there are types of records, for example, specialist consultations, which may have different lengths (institutions add the name of the medical department, type of specialist, doctor's name, etc.) and content names (synonyms: examination instead of consultation, etc.). Thus, groups of names of different density and size



Extraction with regular expressions

Subheading	Subsection text
#record_start	Text 1
Subheading 1	Text 2
Subheading 2	Text 3
Subheading 3	Text 4

Fig. 5 Example of identifying subsections in a record

are formed in the feature space, for the identification of which the OPTICS method is the most appropriate.

To determine the type of record with new data in the future, it is proposed to train a classifier for which the cluster number will be used as a class. One of the simplest and most suitable methods is the k-nearest classification since it is based on distribution of vectors in the feature space.

At this stage, it is possible to use any other methods of vector transformation, clustering, and classification that are most suitable for the peculiarities of the record language and the specificity of the recordset.

Subsection recognition

Finding substructures in records is highly dependent on the dataset. With our dataset, it was noticed that records of the most represented HIS contain subsections, whose names (subheadings) can be extracted using regular expressions. Thus, each record is divided into subsections with a subheading. Often, the beginning of a recording does not have a subheading, so it is given the service name #record_start. Figure 5 shows an example of splitting a record using regular expressions.

If records do not contain subheadings or other substructures, one can skip this step. If the texts of the record are quite long and consist of several paragraphs, then topic modeling can be carried out on these paragraphs. This will allow to divide the records into substructures and reveal their format.

Subsections are grouped by subheadings. For each type of subsection, separate topic modeling is carried out using the method of additive regularization. Additive regularization of topic models adds regularizers to the matrix decomposition, which help to highlight background topics, sparse the topic matrix for a clearer separation between topics, and automatically determine the number of topics (Vorontsov et al. 2015). In addition, this method shows the best modeling results on texts of different lengths, which is typical for EHRs from different medical centers. We carry out 50 iterations without any regularizers and extra 30 iterations with sparse regularizer (smooth sparse phi regularizer with tau = 1e6). However, tau value for regularizers depends on the number of input texts and their lengths, so for much larger or less corpus, the value should be tuned manually. Topic modelling on each sectionallows us to identify key terms (excluding

background words) for each type of subsections, compare them with each other and validate how much the declared subheading corresponds to the content of the subsections.

Based on the identified topics and their key terms, the topic segmentation model that we developed earlier is trained (Shaikina and Funkner 2020). This model calculates the frequency of topic terms in each sentence and adds the coefficient of the most frequent terms of the previous and next sentences. The trained topic segmentation model can be used to label other records that do not have substructures inside.

Accessing quality and record format detection

"Subsection recognition" section describes how to extract subsections from records and thereby identufy record structures from data. The next step is to identify the typical record structure for the extracted subsections. In this work, we calculate the frequency of subsections by type of records and, according to the selected threshold, determine the most appropriate subsections for each record type. The choice of the threshold can be manual or automatic with searching for a critical value: the threshold is calculated by determining the proportion of records that are considered "ideal".

"Ideal" post formats are stored as a dictionary, where the key is the record type and the

```
{
   'record type 1' : ['subheading 11', 'subheading 12', …],
   'record type 2' : ['subheading 21', 'subheading 22', …],
   …
   'record type n' : ['subheading n1', 'subheading n2', …],
}
```

value is a list of subheadings (Python programming language):

After determining the "ideal" format, we can calculate how many of the required subsections each record contains (after extracting subsections with regular expressions or topic segmentation, see "Subsection recognition" section). In addition, based on the "ideal" format, we can make recommendations about which sections to add, and which ones are better to transfer to other types of records.

Results

Preliminary data analysis

Each provided record includes metainformation (patient ID, specialist ID, institution ID and name, HIS ID, date, ICD-10 diagnosis, record name) and free-form text. Text can be presented in different forms and includes from 0 to 827 sentences (Fig. 6). Moreover, Fig. 6 shows how different HIS records are. For example, the most represented HISs (#1 and #5) have the same dispersion for text length and number of words and sentences. HIS #11 includes shorter texts, but has many more words and sentences on average. Probably, HIS #11 has more abbreviations and omitted words inside its records. HIS #3 provides long texts but the number of words is about zero. It means that the records are filled with special signs and HIS tags. HISs #10, #12, and #4 for all metrics have a median of about zero, so we suppose that most of the records are empty or filled with meaningless special signs.

We also compared unique words and their incidence rate in each HIS. HIS #1 and #5 contain the most unique words: 57,348 and 33,327 words, respectively. Also, they

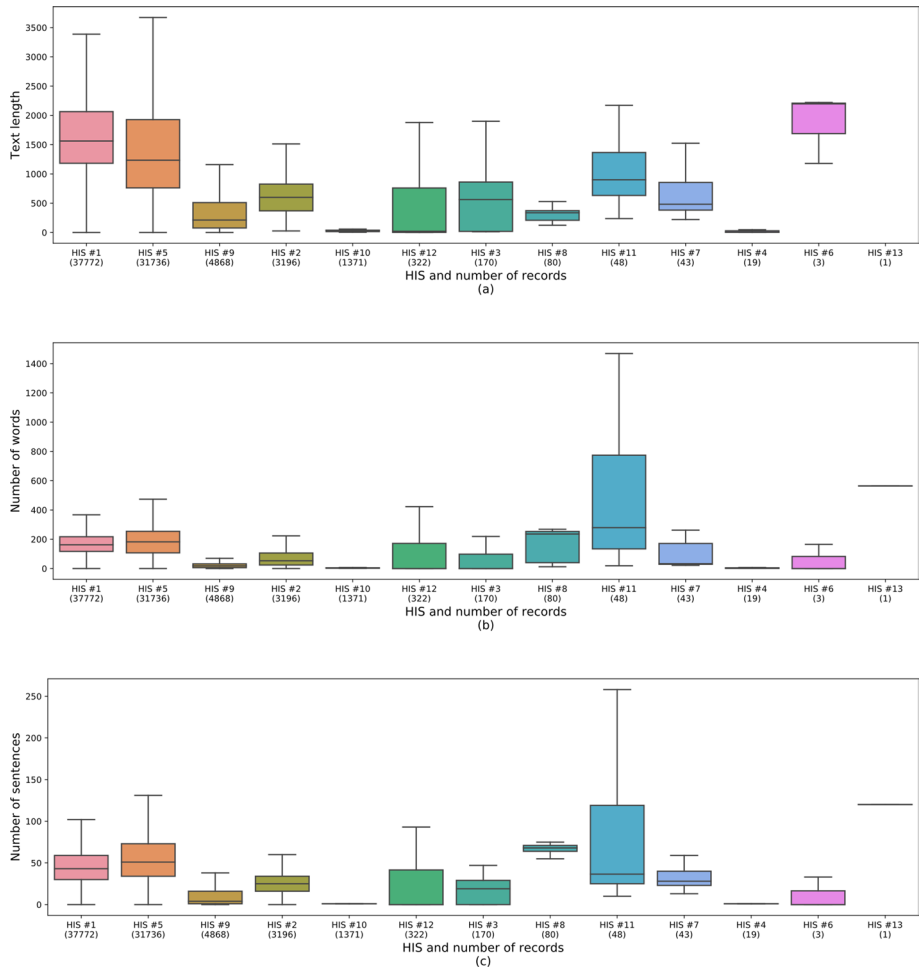


Fig. 6 Comparison of HIS corpora for text length (**a**), number of words in texts (**b**), and number of sentences in texts (**c**). HISs are sorted by the number of records

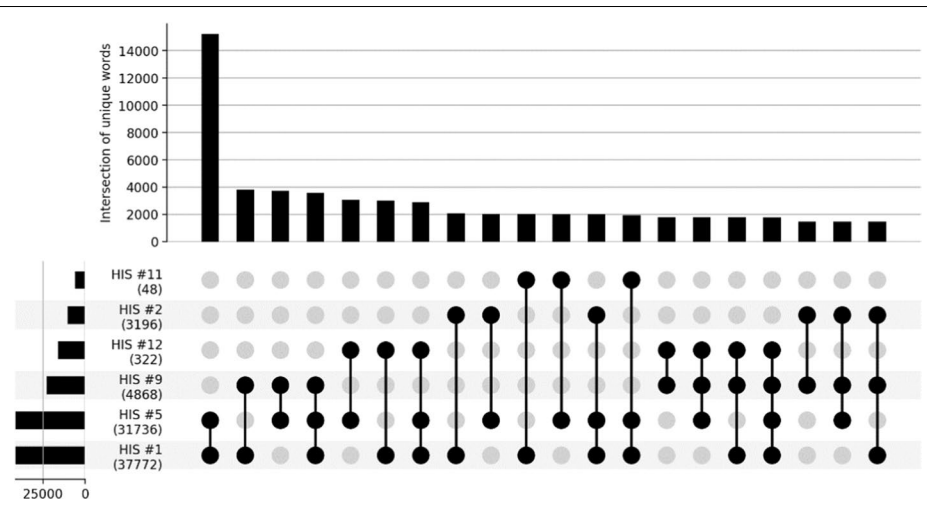


Fig. 7 Upset plot for the most common intersections of unique words of HISs

Table 1 Most common words in the most presented HISs

HIS #1			HIS #5			HIS #9		
Words in Russian	**English translation**	**Number of words**	**Words in Russian**	**English translation**	**Number of words**	**Words in Russian**	**English translation**	**Number of words**
мг	mg	98,027	мг	mg	68,987	мг	mg	5384
ад	BP	90,394	ст	Hg	65,451	день	day	2742
ст	Hg	88,399	ад	BP	45,731	ад	BP	2175
анамнез	Anamnesis	53,509	дыхание	Respiration	42,702	увеличить	To enlarge	1884
мин	min	47,708	анамнез	Anamnesis	41,038	дыхание	Respiration	1565
мм	mm	47,521	год	Year	40,625	состояние	State	1554
рт	Hg	45,146	безболезненный	Painless	37,736	утром	Morning	1456
год	Year	43,281	заболевание	Disease	37,289	заболевание	Disease	1276
тон	Tone[a]	39,009	диагноз	Diagnosis	37,164	мин	min	1268
состояние	State	37,662	состояние	State	35,302	таб	pill	1181
дыхание	Respiration	37,348	лечение	Treatment	34,871	ст	Hg	1079
чистый	Clear	36,812	мм	mm	34,674	безболезненный	Painless	1074
сердце	Heart	36,510	увеличить	Enlarge	34,286	мм	mm	1057
ритмичный	Rhythmic	35,984	сердце	Heart	33,590	контроль	Control	1052
удовлетворительный	Satisfactory	35,586	отрицать	Deny	30,403	дата	Date	1031
жалоба	Complaint	32,651	тон	Tone	30,187	вечером	Evening	989
живот	Stomach	32,153	жалоба	Complaint	29,902	пульс	Pulse	985
увеличить	To enlarge	32,125	литр	LITRE	29,558	рт	Hg	947
гб	HD	31,189	мин	min	28,938	диагноз	Diagnosis	918
мягкий	Tender[b]	31,011	удовлетворительный	Satisfactory	28,381	сердце	Heart	914

[a] From "heart tones"

[b] From "abdominal wall is tender"

Table 2 Metacharacteristics: misprints, terms, abbreviations according to dictionaries

HIS	Medical terms	Russian words	English words	Proper name	Abbreviations	Medical Drugs	Correct words	Number of records
HIS #1	0.21	0.33	0.0001	0.20	0.17	0.03	0.42	37,772
HIS #5	0.32	0.47	0.0001	0.14	0.20	0.04	0.57	31,736
HIS #3	0.47	0.64	0.0014	0.09	0.54	0.10	0.77	170
HIS #10	0.63	0.78	0.0000	0.24	0.61	0.00	0.78	1371
HIS #9	0.47	0.69	0.0000	0.18	0.38	0.08	0.79	4868
HIS #12	0.55	0.72	0.0007	0.15	0.35	0.06	0.81	322
HIS #11	0.57	0.74	0.0000	0.17	0.36	0.06	0.81	48
HIS #2	0.51	0.72	0.0000	0.15	0.41	0.08	0.82	3196
HIS #6	0.48	0.72	0.0000	0.20	0.44	0.07	0.82	3
HIS #8	0.41	0.80	0.0000	0.30	0.48	0.04	0.86	80
HIS #7	0.66	0.85	0.0000	0.11	0.37	0.05	0.91	43
HIS #4	0.53	0.89	0.0000	0.00	0.21	0.00	0.95	19
HIS #13	0.67	0.87	0.0234	0.16	0.72	0.12	0.96	1

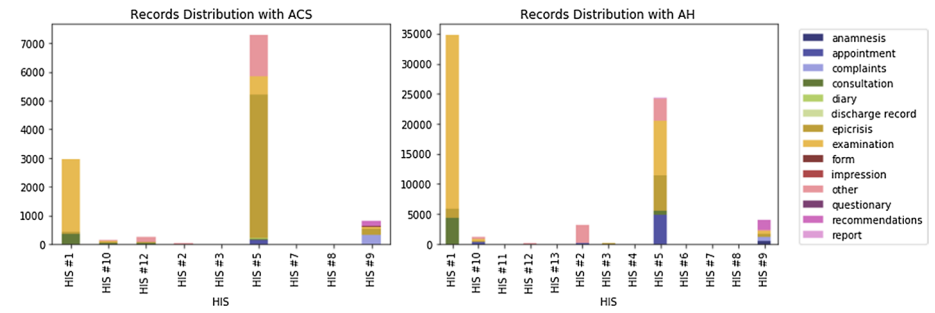


Fig. 8 Distributions of record types for patients with arterial hypertension (AH) and acute coronary syndrome (ACS) with manual labeling

have the largest intersection rate of unique words: 15,231 words (see Fig. 7). HIS #11 contains only 48 records (records are long and have many words and sentences, see Fig. 6), however, it has one of the largest intersection rates with HIS #1 and #5. This indicates the similarity of the content of records in HIS #11 to the most represented HISs.

Table 1 shows the most frequent words for HIS #1, #5, and #9, excluding stop words. The most common words for these HISs are units of measurement: doses of prescribed drugs and results of medical tests (mg), blood pressure measurements (mm, Hg), heart rate (min), frequency of drug intake (pill, day, morning, evening). Also, many words are associated with describing a patient's condition: history, state, breathing, heart, satisfactory, complaint, diagnosis, etc.

We also try to estimate the number of misprints and spelling errors in the texts (see Table 2). Using the module for correcting misprints (Balabaeva et al. 2020), we compare the unique words from the records with the corresponding dictionaries (dictionaries of medical terms, Russian spelling dictionary, English dictionary, dictionary of medicines). Also, based on these dictionaries, the proportion of correct (found in dictionaries) words is estimated. Most misprints in unique words

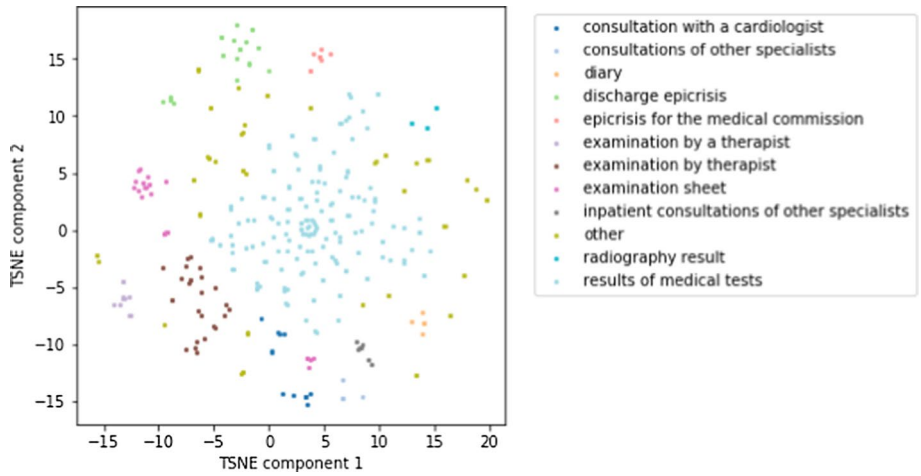


Fig. 9 t-SNE dimensionality reduction for record header space (the header points are colored according to OPTICS cluster)

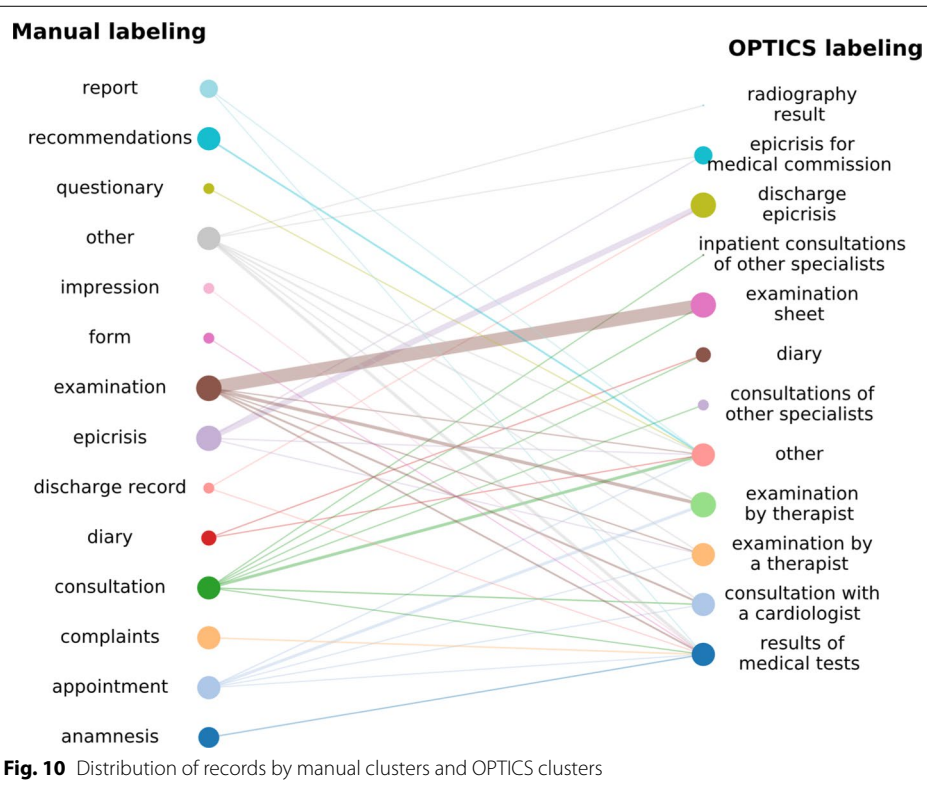


Fig. 10 Distribution of records by manual clusters and OPTICS clusters

Table 3 Comparison of labeling using clustering metrics

Labeling	Silhouette coefficient	Calinski-Harabasz Index	Davies-Bouldin Index
Manual	−0.18	30.57	2.77
OPTICS	0.11	44.07	2.69

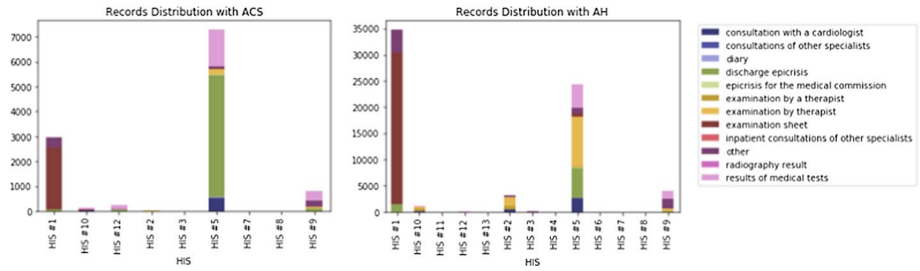


Fig. 11 Distributions of record types for patients with arterial hypertension (AH) and acute coronary syndrome (ACS) with OPTICS labeling

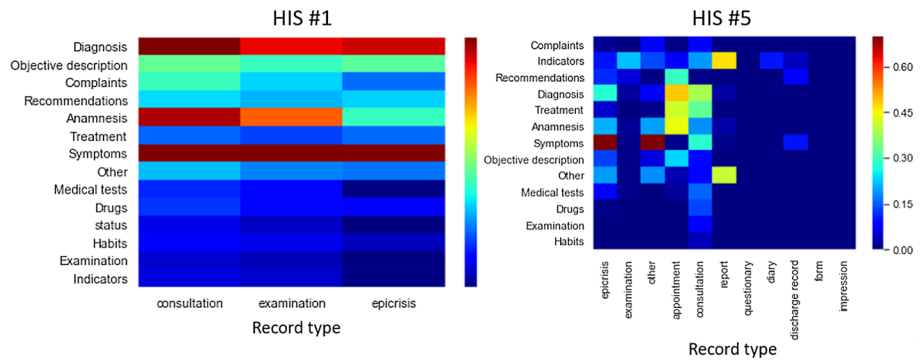


Fig. 12 Distribution of subsections among record types for two selected HISs

are contained in HISs #1 and #5, which is expected: the more words, the more mistakes. However, HIS #3 has a low share of correct words, although it contains only 170 entries. When processing the words of this system, it is necessary to correct misprints.

Structuring and analysis of EHR data

We have applied the methods described in "Implementation details" section to the MIAC dataset. As shown in Fig. 8, HISs #1 and #5 have the most records (88% of records in total) in the analyzed dataset. Therefore, we used the data of these HISs for further training and validation of all the models.

Figure 8 shows the manually labeled types of records for all HISs. As can be seen, HIS #1 contains only three types of records: examination (83% of all records), epicrisis and doctor's consultation. HIS #5 contains 10 types of records. As HIS #2 has specific names for its records, all of them are labeled as "other".

Furthermore, automatic labeling of records by types is carried out: the OPTICS method for clustering is applied to the preprocessed and vectorized headers of records (see "EHR type detection" section). After reviewing all the clusters, the names for each group were determined. Compared to manual labeling (Fig. 8), clusters have more specific names (Fig. 9). Also, we compare how manual and OPTICS clusters are related by records (Fig. 10). The main share of records from the manual category "examination" transfers into the category of "examination sheet", and so on with "epicrisis" and "statement epicrisis". However, the manual categories "consultation", "appointment" and

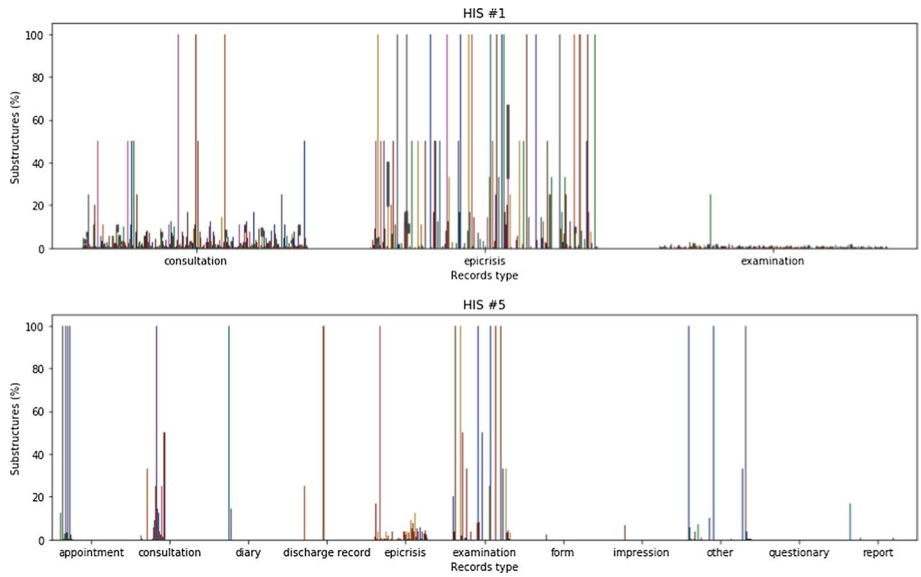


Fig. 13 Percentage distibutions of subsections among records type for HIS #1 and HIS #5 (different colors for different subsections)

Table 4 Automatically retrieved accurate records for HIS #1 and HIS #5 according to subsection representations

	HIS #5	HIS #1
Record type	**Accurate records**	
Appointment	1	–
Consultation	5	9
Diary	7	–
Discharge record	7	–
Epicrisis	33	–
Examination	2	4
Other	1	–
Report	25	–

"other" are divided into consultation and examination groups by different specialists. Besides, Table 3 shows clustering metrics for both labeling systems: silhouette coefficient (the closer to 1, the better), Calinski-Harabasz index (the higher the better), and Davies-Bouldin index (the lower, the better). OPTICS clustering shows better results according to all calculated metrics. Figure 11 shows how OPTICS records are distributed across HISs. HIS #2 now has several types of records, not just "others" as in Fig. 8. HIS #1 still consists of three types of records.

At the next stage, we extract the subheadings and subsections (see "Subsection recognition" section) for HIS #1 and HIS #5: 225 and 150 subsections, respectively. Subheadings are grouped manually to simplify visualization and primary analysis. The distribution of subsections is shown in Fig. 12. For some record types (questionnaire, form, and impression), no subheadings were found using regular expressions. In general,

Table 5 Comparison of EHR structure indices by HIS and records with AH and ACS diagnoses

HIS	AH	ACS	Total
HIS#1	0.0285	0.0225	0.0239
HIS#5	0.0142	0.0164	0.0163

subsections are well related to what should be in the record type: "report" contains the largest proportion of subsections related to health indicators (blood pressure, heart rate, etc.); "epicrisis" contains the largest proportion of "symptoms", as it describes what happened to the patient during the hospitalization.

In addition, topic modeling and segmentation were carried out on the texts of the subsections (see "Subsection recognition" section). The training, validation, and test set of each HIS is 50%, 20%, and 30% of the entire set, respectively. Thus, for HIS #1, automatic segmentation to predict the type of subsection shows the result of 0.51 F1-sore (38 classes), and for HIS #5 it is 0.11 F1-score (19 classes). The low metric is due to the large variability of subheadings (it is necessary to carry out careful grouping and processing to find the same names and combine subsections more correctly). Similar clustering and classification methods can be applied as in "EHR type detection" section.

"Accessing quality and record format detection" section describes methods for assessing the quality of records. Figure 13 shows the percentage distribution of subsections by record type. As can be seen, the types of records are characterized by different types of subsections and their number. To assess the quality, a threshold of 10% was chosen: if at least 10% of the texts of the considered subsection are contained in this type of record, then this subsection is typical for this type. Based on this threshold, the records of HISs #1 and #5 are assessed for each medical center separately. Table 4 shows that less than 1% of the records are found to be reasonably accurate (containing more than 80% of the required subsections). In total, there are more than 30 thousand records for each of these two HISs.

Interpretation and evaluation

This section provides an expert evaluation of the quality of records based on the extracted data about the record type and subsections within the record. The evaluation was performed together with specialists of MIAC who are involved in monitoring and assessing the information quality of hospitals in Saint Petersburg. The goal of the evaluation was twofold. First, the analysis of the EHR structure and completeness was performed for selected hospitals. Second, the comparison to existing assessing procedures applied in MIAC was performed to consider possible extension and updating of them. To reach the goal, the current study was focused on indices based on EHR structure which relatively reflect the completeness of EHR. At the same time it enables close comparison to existing official measures of EHR implemented in MIAC.

Based on the processed information, an index of the EHR structure in a healthcare facility was constructed with an assessment of the presence of the necessary subheadings for patients with arterial hypertension and acute coronary syndrome (AH and ACS) for 23 most represented institutions using HIS#1 and HIS#5.

Table 6 Comparison of EHR structure indices by outpatient and inpatient clinics

HIS	Outpatient	Inpatient
HIS#1	0.0239	–
HIS#5	0.0086	0.0218

Lets a record from dataset be r and each recod includes subtitles s: $r_i = \{s_1, s_2, \ldots, s_i\}$, where i the number of records in dataset. To calculate the EHR structure index, the various types of EHR were grouped into three groups: (1) epicrisis ($G_1 = \{r_1, r_2, \ldots, r_{G_1}\}$); (2) examination, appointment, consultation ($G_2 = \{r_1, r_2, \ldots, r_{G_2}\}$); (3) other records, including reports ($G_3 = \{r_1, r_2, \ldots, r_{G_3}\}$). For each group, the possible set of subtitles T_{G_i} are calculated with the formula:

$$T_{G_i} = \left\{ s_1, s_2, \ldots, s_{k_{G_i}} | s_m \in \bigcup_{j=1}^{n_i} r_j, m = \overline{1, k_{G_i}} \right\}, \quad i = \{1, 2, 3\}.$$

The total frequency of occurrence of subheadings s in these groups G_i relative to the total number of records $n_i = |G_i|$ of each type is calculated. The average sum of occurrence rates for each record type, normalized by the total number of subheading types, can take a value from 0 if there is no information for all subheadings in all record types to 1 in the opposite case when all subheadings are filled:

$$SI(G_i) = \frac{\sum_{s_j \in T_{G_i}} \sum_{r_l \in G_i} \Delta(s_j, r_l)}{n_i k_{G_i}}, \quad n_i = |G_i|, \quad k_{G_i} = \left|T_{G_i}\right|,$$

$$\Delta(s, r) = \begin{cases} 1, \ if & s \in r \\ 0, \ if & s \notin r \end{cases}.$$

This value is proposed to be taken as an index of the EHR structure.

The results of calculating the EHR structure index are presented in Table 5.

To conclude, from Table 5, the EHRs are better managed in HIS#1. This may indicate both a more developed functionality and a deeper level of implementation in this type of HIS. HIS#1 is implemented only in outpatient clinics, whereas HIS#5 is implemented in both outpatient and inpatient clinics. To compare different types of HISs under the same conditions and to exclude the impact of different EHR requirements in the case of outpatients and inpatients, an index for each HIS was calculated, see Table 6.

Based on the index of the EHR structure for each medical organization, a rating of healthcare facilities was formed, which allows to position healthcare facilities from more structured management of EHRs to a less structured and less detailed one. This rating allows to objectively compare different clinics and apply administrative or incentive measures to equalize the quality of EHR management. In addition, analysis of the dynamics of changes in the EHR structuring index in Saint Petersburg as a whole can allow us to draw conclusions about the effectiveness of the use of organizational and financial incentives and to forecast the achievement of target levels of EHR structuring.

In Saint Petersburg, the measurement of EHR completeness and quality index introduced by the MIAC for assessing the city's healthcare facilities (hereinafter the MIAC

Table 7 Overall rating of healthcare facilities by the EHR structure index in comparison with the MIAC index

Institution #	EHR structure index	MIAC index
1	0.0667	45.08
2	0.0530	23.85
3	0.0356	16.63
4	0.0346	47.06
5	0.0341	61.32
6	0.0239	33.42
7	0.0220	20.1
8	0.0209	13.28
9	0.0176	28.32
10	0.0177	33.6
11	0.0170	35.57
12	0.0152	22.7
13	0.0139	26.87
14	0.0132	51.39
15	0.0130	45.07
16	0.0128	35.55
17	0.0099	20.83
18	0.0085	10.96
19	0.0074	27.62
20	0.0043	22.25
21	0.0039	26.5
22	0.0019	44.82
23	0.0002	25.25

index) is already in use. The MIAC indices are calculated according to officially approved methodology for assessing the EHR completeness in different hospitals. The index is based on presense of explicit records and documents provided by the hospital within the integrated HIS. The calculated indices for different hospitals in Saint Petersburg are published periodically on the official MIAC site both for city level and for different city hospitals (MIAC 2021). The correlation between the MIAC index and the EHR structure index in healthcare facilities was calculated. It turned out to be low (equal to 0.231), which practically means a weak relationship between these indices. This can be explained by the fact that the MIAC index characterizes the completeness of the transmission of records, while the EHR structure index characterizes the completeness of EHRs themselves (see Table 7).

Discussion

Textual data in an EHR contains an important portion of the information regarding the healthcare service provided to the patient. Structuring of such information plays an important role in multiple tasks including improvement of information consistency and interoperability, clinical decision support, and healthcare facility assessment. Within our case study, the relatively low correlation between the informativeness of structured and unstructured parts of EHRs through the presented indices was observed. After the detailed analysis of existing structures and variation in EHRs discovered during this

study the discussion was initiated with the MIAC experts. One of the reached conclusions was that the existing indices used for assessing completeness and quality of EHR applied in practice need to be extended with deeper analysis of EHR with the developed procedures. Thus, structuring and analysis play an important role in the improvement of EHRs both when they are collected in the MIAC and inside the HISs of the hospitals. Also, the possible update of existing indices can improve the assessing the quality of information management in hospitals and organization ranking over the city by making the indices more detailed and well-grounded.

Another important result that can be seen through the diversity and structuring analysis are the behavioral patterns of physicians who input EHR data. The patterns can be seen through the structure and meta-characteristics of the text. They reflect the principles and practices in clinical decision making, as well as the experience of a physician. Moreover, additional closeness of EHRs within a single hospital can be further explained through the "common information space" within a hospital where general rules and practices are implemented in a unified way. Further analysis and interpretation of the diversity can be considered a source for identification of hospitals' and physicians' profiles within a complex citywide healthcare system.

The proposed method may be considered a general way for analysis and structuring of EHR data in diverse datasets. The approach enables a deeper understanding of the sources of diversity and differentiates particular structures in EHR data in an automatic or semi-automatic way. Considering EHR data as a reflection of the real-world healthcare system and the processes in it, a possible application is assessing and improving healthcare service through identification and sharing of the best practices both in terms of clinical decision making and information structuring. We believe that the proposed approach may be used to structure EHR data for better understanding and analysis of distributed healthcare systems.

Conclusion and future work

Within the proposed work, we introduce an approach for structuring and analysis of EHR data in a distributed complex healthcare system. The proposed method can be applied in diverse applications including assessment, improvement of information in EHR systems, and extending the healthcare service with additional clinical decision support and analytical services. Within this research, we consider a case study of the city healthcare system in Saint Petersburg, Russia, to introduce additional structuring, analysis, and assessment of healthcare facilities. The obtained results were used by the MIAC for further improvement of the citywide healthcare monitoring and assessment system. The listed problems of processing unstructured records and the absence of a unified HIS in Saint Petersburg are the basis for a new large-scale project for the analysis, unification, and standardization of the accumulated data in the MIAC to analyze the citywide quality of healthcare. We believe that the proposed approach may be applied in different cases where diverse EHR data is processed and analyzed (e.g., data collected on the level of large cities or even a country).

Further development of the approach includes several directions. First, the approach can be extended with a deeper interpretation of the diversity in EHRs (including

personal experience, local policies, common information in hospitals, etc.). Second, the multiscale information sharing between physicians, hospitals, HISs can be estimated and analysed. Third, physician profiling and personal practices can be identified, structured, and assessed to correct (as bad practices) or share (as good practices). Finally, information exchange in a global diverse environment can be optimized to improve both clinical practices and information interoperability.

Abbreviations

ACS: Acute coronary syndrome; AH: Arterial hypertension; BP: Blood pressure; EHR: Electronic health records; HD: Hypertension disease; Hg: Hydrargyrum (Mercury); HIS: Health information system; HTML: HyperText markup language; ICD: International classification of diseases; ID: Identifier; mg: Milligrams; MIAC: Medical information and analytical center; min: Minutes; mm: Millimeters; NLP: Natural language processing; OPTICS: Ordering points to identify the clustering structure; t-SNE: T-distributed stochastic neighbor embedding; TF–IDF: Term frequency–inverse document frequency; XML: Extensible markup language.

Acknowledgements

Not applicable.

Authors' contributions

All authors have contributed equally.

Funding

This work was supported by the Ministry of Science and Higher Education of Russian Federation, goszadanie no. 2019-1339.

Availability of data and materials

The data that supports the findings of this study is available from the MIAC (https://spbmiac.ru/o-miac) but restrictions apply to the availability of this data, which was used under license for the current study, and so is not publicly available.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]ITMO University, Saint Petersburg, Russia. [2]Medical Information and Analytical Center, Saint Petersburg, Russia. [3]Sokolov North-Western District Scientific and Clinical Center, Saint Petersburg, Russia.

Received: 18 March 2021 Accepted: 29 June 2021
Published online: 17 July 2021

References

Ankerst M, Breunig MM, Kriegel HP, Sander J (1999) OPTICS: ordering points to identify the clustering structure. ACM Sigmod Rec 28:49–60. https://doi.org/10.1145/304181.304187

Balabaeva K, Kovalchuk S (2020) Experiencer detection and automated extraction of a family disease tree from medical texts in russian language. In: Lecture Notes in Computer Science, vol 12140, pp 603–612. https://doi.org/10.1007/978-3-030-50423-6_45

Balabaeva K, Funkner A, Kovalchuk S (2020) Automated spelling correction for clinical text mining in Russian. Digit Pers Heal Med Proc MIE 270:43–47

Batini C, Scannapieco M (2016) Data and information quality: dimensions, principles and techniques, 1st edn. Springer Publishing Company Incorporated

Bruland P, Doods J, Storck M, Dugas M (2017) What information does your EHR contain? Automatic generation of a clinical metadata warehouse (CMDW) to support identification and data access within distributed clinical research networks. Stud Health Technol Inform 245:313–317. https://doi.org/10.3233/978-1-61499-830-3-313

Burke HB, Hoang A, Becher D et al (2014) QNOTE: an instrument for measuring the quality of EHR clinical notes. J Am Med Inform Assoc 21:910–916. https://doi.org/10.1136/amiajnl-2013-002321

Datta S, Bernstam EV, Roberts K (2019) A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. J Biomed Inform 100:103301. https://doi.org/10.1016/j.jbi.2019.103301

Freedman HG, Williams H, Miller MA et al (2020) A novel tool for standardizing clinical data in a semantically rich model. J Biomed Informatics X 8:100086. https://doi.org/10.1016/j.yjbinx.2020.100086

Funkner AA, Kovalchuk SV (2020) Time expressions identification without human-labeled corpus for clinical text mining in Russian. In: Lecture Notes in Computer Science, vol 12140, pp 591–602. https://doi.org/10.1007/978-3-030-50423-6_44

Funkner A, Balabaeva K, Kovalchuk S (2020) Negation detection for clinical text mining in Russian. In: Studies in health technology and informatics, vol 270, pp 342–346. https://doi.org/10.3233/SHTI200179

Kersloot MG, van Putten FJP, Abu-Hanna A et al (2020) Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. J Biomed Semant 11:1–21. https://doi.org/10.1186/s13326-020-00231-z

Lamy M, Pereira R, Ferreira JC, et al (2019) Extracting clinical information from electronic medical records. In: Advances in intelligent systems and computing. In: Advances in Intelligent Systems and Computing, vol. 806, pp 113–120. https://doi.org/10.1007/978-3-030-01746-0_13

Logan JR, Gorman PN, Middleton B (2001) Measuring the quality of medical records: a method for comparing completeness and correctness of clinical encounter data. In: Proc AMIA Symp, pp 408–412. https://pubmed.ncbi.nlm.nih.gov/11825220/

MIAC (2021) Medical organizations ranking: St. Petersburg citizen's EHR (Рейтинги медицинских организаций: ЭМК петербуржца) - in Russian. https://spbmiac.ru/ehlektronnoe-zdravookhranenie/rejjtingi-e-zdravookhraneniya/rejtingi-mo-emk-peterburzhca/

Névéol A, Dalianis H, Velupillai S et al (2018) Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semant 9:1–13. https://doi.org/10.1186/s13326-018-0179-8

Newton-Dame R, McVeigh KH, Schreibstein L et al (2016) Design of the New York city macroscope: innovations in population health surveillance using electronic health records. Gener Evid Methods Improv Patient Outcomes 4:26. https://doi.org/10.13063/2327-9214.1265

Nguyen L, Bellucci E, Nguyen LT (2014) Electronic health records implementation: an evaluation of information system impact and contingency factors. Int J Med Inform 83:779–796. https://doi.org/10.1016/j.ijmedinf.2014.06.011

Roberts K, Demner-Fushman D (2016) Annotating logical forms for EHR questions. In: Proceedings of the 10th international conference on language resources and evaluation, LREC 2016. pp 3772–3778. https://pubmed.ncbi.nlm.nih.gov/28503677/

Shaikina AA, Funkner AA (2020) Medical corpora comparison using topic modeling. Procedia Comput Sci 178:244–253. https://doi.org/10.1016/j.procs.2020.11.026

St-Maurice J, Burns C (2017) An exploratory case study to understand primary care users and their data quality tradeoffs. J Data Inf Qual 8:1–24. https://doi.org/10.1145/3058750

Sun H, Depraetere K, De Roo J et al (2015) Semantic processing of EHR data for clinical research. J Biomed Inform 58:247–259. https://doi.org/10.1016/j.jbi.2015.10.009

Tang B, Wu Y, Jiang M et al (2013) A hybrid system for temporal information extraction from clinical text. J Am Med Inform Assoc 20:828–835. https://doi.org/10.1136/amiajnl-2013-001635

van der Bij S, Khan N, ten Veen P et al (2017) Improving the quality of EHR recording in primary care: a data quality feedback tool. J Am Med Inform Assoc 24:81–87. https://doi.org/10.1093/jamia/ocw054

Vorontsov K, Frei O, Apishev M et al (2015) Bigartm: open source library for regularized multimodal topic modeling of large collections. Commun Comput Inf Sci 542:370–381. https://doi.org/10.1007/978-3-319-26123-2_36

Wang Z, Shah AD, Tate AR et al (2012) Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLOS ONE 7:e30412. https://doi.org/10.1371/journal.pone.0030412

Weiskopf NG, Weng C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 20:144–151. https://doi.org/10.1136/amiajnl-2011-000681

Williams JG (2003) Measuring the completeness and currency of codified clinical information. Methods Inf Med 42:482–488. https://doi.org/10.1055/s-0038-1634243

Yehia E, Boshnak H, AbdelGaber S et al (2019) Ontology-based clinical information extraction from physician's free-text notes. J Biomed Inform 98:103276. https://doi.org/10.1016/j.jbi.2019.103276

Publisher's Note