



# Machine Learning Based Text Mining in Electronic Health Records: Cardiovascular Patient Cases

Sergey Sikorskiy<sup>1</sup>(✉), Oleg Metsker<sup>1</sup>, Alexey Yakovlev<sup>2</sup>,  
and Sergey Kovalchuk<sup>1</sup>

<sup>1</sup> ITMO University, Saint Petersburg, Russia  
sikorskiy.s@hotmail.com, olegmetsker@gmail.com,  
sergey.v.kovalchuk@gmail.com

<sup>2</sup> Almazov National Medical Research Centre, Saint Petersburg, Russia  
alex.yakovlev@mail.ru

**Abstract.** This article presents the approach and experimental study results of machine learning based text mining methods with application for EHR analysis. It is shown how the application of ML-based text mining methods to identify classes and features correlation to increases the possibility of prediction models. The analysis of the data in EHR has significant importance because it contains valuable information that is crucial for the decision-making process during patient treatment. The preprocessing of EHR using regular expressions and the means of vectorization and clustering medical texts data is shown. The correlation analysis confirms the dependence between the found classes of diagnosis and individual characteristics of patients and episodes. The medical interpretation of the findings is also presented with the support of physicians from the specialized medical center, which confirms the effectiveness of the shown approach.

**Keywords:** Text mining · Cardiology · Machine learning · Decision support  
Treatment process analysis · Acute coronary syndrome

## 1 Introduction

The developing methods of electronic health record (EHR) analysis is a relevant scientific problem because EHR contains critical information about the treatment process. In this regard, the automation of EHR texts processing could significantly expand and improve the diseases classification, the treatment duration, and cost prediction as well as prediction other important events. However, the difficulties of EHR analysis are associated with the semi-structuring, specific stylistic and grammatical features, abbreviations, clinician terms, etc. Pattern-based methods are often used to solving this problem [1]. These methods are characterized by high accuracy but they do not provide high flexibility and identification of various combinations of speech turns which limits their application. On the other hand, the machine learning (ML) based methods show high flexibility in the analysis of EHR and have a relatively good accuracy [2–4].

This research is demonstrating the capabilities of using ML-based methods of text mining to extend EHR analysis. The experimental study is focused on data extraction and classification of EHR of cardiovascular patients in Almazov Centre (one of the leading cardiological centers in Russia). The proposed approach enables improvement of the accuracy of events and parameters predicting in the course of patient treatment.

## 2 Related Work

The text-mining methods are widely used in a various domains [5–9]. Also, there are a lot of applications in healthcare. For example, a task of extraction of associations of diseases and genes associated with them is considered in [10]. The proposed approach is based on the method of named object recognition (NOR) using a dictionary of names and synonyms. Also, to analyze text on associations the analysis of the EHR semantics was applied to identify the links between medical terms with disease and treatment. For example, semantic analysis based on the ML method to identify clinical terms is presented in [11]. There are known problems of determining the sequence of events (treatment, examination, operation) and identifying cause-and-effect relationships in semi-structured EHR in the healthcare domain. The semantic analysis allows defining the relationship between events and causes. For example, solving this problem is possible with the semantic analysis using name recognition, Bayesian statistics, and logical rules to identify the risk factors [12]. It is possible to extract cause-and-effect relationships from EHR with using text mining based on ML methods. For example, a random forest is used to classify the descriptions of allergic reactions to drugs. Such method provides sufficient accuracy (90%) in comparison with the pattern based method (82%) [13].

Considering healthcare in Russia, it worth to mention that the Russian language has own features. To analyze EHR in the Russian language, the pattern based methods [1] are used more often. The pattern based methods demonstrate good accuracy in EHR analysis, but it does not give flexibility comparable to ML methods.

## 3 EHR Analysis: Cardiovascular Patients

Within the study, the analysis of diagnoses performed to identify implicit connections between patients EHR that do not fit into the basic clinician categories. The analyzed EHR dataset contains information about 8865 cardiovascular patients, different age categories from 25 to 96 years and various outcomes of treatment (died or not), which is not reflected in the texts of diagnoses. The diagnoses were taken from electronic medical records of the patients of the Almazov Centre treated from 2010 to 2017. The general study group ( $n = 8865$ ) included several categories of the patients:

- Patients with the acute coronary syndrome (ACS), admitted to hospital urgently by Ambulance or by transfer from other hospitals ( $n = 4955$ ).
- Patients with a stable course of coronary heart disease, which is a routine performed angioplasty and stenting of coronary arteries ( $n = 2760$ ).

- Patients with heart rhythm disorders, which was performed by the arrhythmic interventions in emergency and planned order or cardioversion (n = 619).
- Other patients with emergency cardiac conditions requiring treatment in intensive care units and surgery patients (n = 531).

The study analyzed cases of medical care, which accumulated a sufficient amount of text data (5500 patients) containing a detailed description of clinical diagnosis (underlying disease, associated conditions, complications, comorbidities with detailed characteristics and stages, the name and date of invasive interventions and surgical operations, ICD-10 codes). Analyzed EHRs include only manual entry of the text (without templates). The comparison of the selected group with the general database of cardiovascular patients by average indicators (gender, age, indicators, analyses, death rate, the average length of stay, etc.) do not reveal differences. It can be concluded that the selected cases are representative.

EHR preprocessing involves bringing the text to a more readable form because the diagnoses contain an abundance of abbreviations and various stylistic methods of writing medical texts. The preprocessing is to find and unify terms and acronyms. Regular expressions and other procedural methods collected in the Python script enabled unification of terms within the texts with replacement of various occurrence of the terms with a single one. The complete list contains more than 1000 regular expressions in 12 main groups: (1) abbreviation of diseases and synonyms; (2) ranks, stages of disease, functional classes; (3) removing unnecessary words; (4) general condition of the patient; (5) drugs; (6) locations and type of pain; (7) breathing and dyspnea; (8) edemas; (9) urogenital system; (10) blood pressure; (11) numerals; (12) removing special characters.

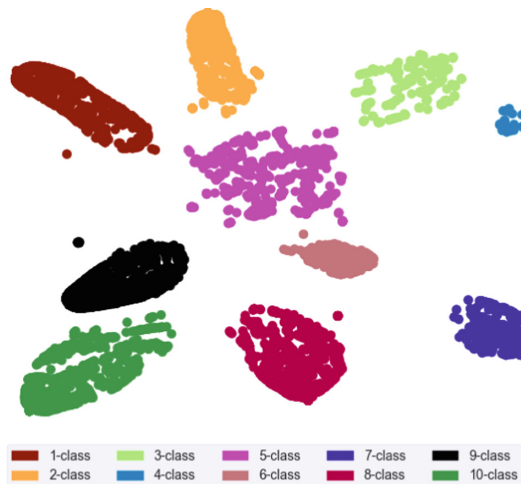
Further EHR processing is based on the methods of texts vectorization and bringing them to binary values and identify groups of the patients. This stage includes converting of texts to the matrix of token (keywords) counts using CountVectorizer from scikit-learn<sup>1</sup> library. Within the analysis 1060 keywords were used. Top keywords include: disease (6064); ischemic (5912); heart (5463); acute (3607); coronary (2477); heart attack (2301); myocardium (1739) heart (1502); angina (1265); hypertonic (1084).

The developed pre-processing routines were generalized to enable reusing of the script in various analysis scenarios.

On the next step ML methods were hired to get deeper understanding of patient variability. The reducing of dimensions of the binary vectors using the t-SNE [14] method from the scikit-learn package discovers the 10 clusters (Fig. 1). The configuration of the t-SNE method to obtain clusters of patients eliminates the random initialization of the model and the coefficient of perplexity 30 which gives reasonable distinction between clusters. t-SNE method gives clearer cluster boundaries even in 2d space, while, for example, principal component analysis (PCA) widely used for dimension reduction shows a weak distinction between clusters with 2 principal components. As a result, t-SNE was selected for further analysis and interpretation.

---

<sup>1</sup> <http://scikit-learn.org/>.



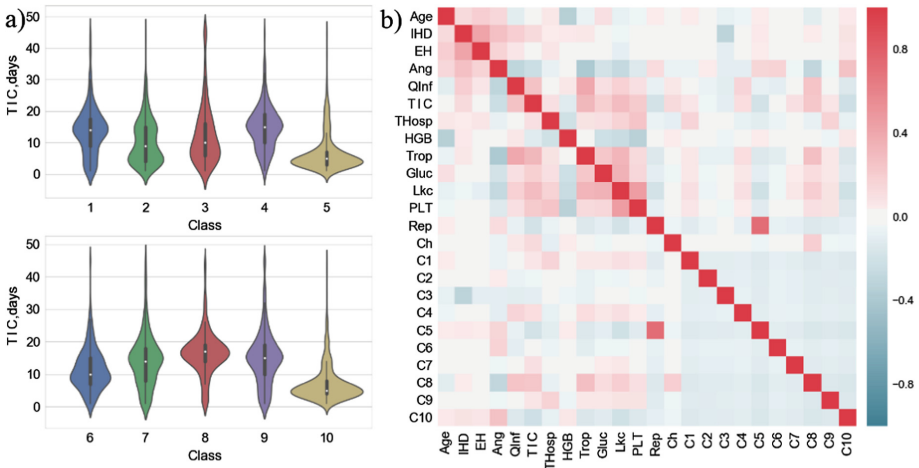
**Fig. 1.** The clusters of diagnoses descriptions.

From a practical point of view, interpretation of different clusters can have different meanings. Class of the patient can be considered as a predictor of an event (e.g., re-hospitalization). To get deeper insight on clusters interpretation it is needed to analyze the patient data and individual characteristics in depth. The obtained 10 classes of patients during the analysis of individual parameters of patients and episodes revealed various meaning of classes. For example, Class 5 and Class 10 differ in the duration of treatment comparing to other classes (Fig. 2a). Also, the correlation analysis showed correlations by class variable (C1–C10, Boolean) with individual characteristics of patients and episodes: age; ischemic chronic disease (IHD); essential hypertension (EH); angina (Ang); Q-wave infarction (QInf); treatment duration in intensive care (TIC); treatment duration in hospital (THosp); hemoglobin minimum (HGB); troponin maximum (Trop); glucose maximum (Gluc); leukocytes maximum (Lkc); thrombocytes maximum (PLT); Repetition (Rep); cholesterol (Ch) (Fig. 2b).

E.g., there is notable correlation between C1 and duration of treatment in the hospital; C3 and presence of hypertension in patients; C4 and myocardial infarction, C5 and repeatability, duration of treatment in intensive care, angina; C8 and cholesterol, maximum leukocyte level, maximum troponin level, q heart attack, angina pectoris, and duration; C10 and leukocytes, angina pectoris, and duration.

It is possible to expand the idea that a combination of text characteristics associated with repeated hospitalizations has been found. E.g., repetition of hospitalization is an essential indicator of value-based medicine related to high cost of treatment.

In general, the diagnosis is based on individual symptoms, syndromes, laboratory and instrumental studies, as well as the results of assessing the dynamics of these manifestations. During the observation of the patient, a clinician forms an idea of the nature of the disease and formulates a diagnosis. The presence of the disease implies the probable presence of certain risk factors that contributed to its development. The role of well-known pathogenetic mechanisms in its occurrence and progression,



**Fig. 2.** EHR analysis (a) treatment duration of the patients in discovered classes; (b) correlation matrix of text classes with individual characteristics of the patients

characteristics of its course, the need for and a specific effect of therapy, the probable prognosis, is considered as a single logical structure of the patient with a particular disease. The classes obtained earlier were analyzed with using the clinical method associated with indication of most frequent tokens in EHR texts:

- **Class 1.** (n = 592) Patients with ACS in the duration of hospital stay is slightly above average (14 K/d), in-hospital mortality of 1.8%. **Tokens:** acute, coronary.
- **Class 2.** Lighter, sharper, more heavy routine, or massive intervention in stable patients. Hospitalization 10 days, mortality of 1.2%. **Tokens:** ischemic, chronic, rheumatic.
- **Class 3.** Mostly scheduled patients with significant heart failure complicating various cardiac diseases, duration of admission is above average and relatively high mortality (2.7%). **Tokens:** atherosclerosis, mitral.
- **Class 4.** Patients with the most severe myocardial infarction, including transmural (with Q), with prolonged hospitalization (15/d), relatively high mortality of 3.1%. **Tokens:** ischemic, acute, infarction.
- **Class 5.** These are planned subsequent stages of revascularization in ACS or planned interventions (including stenting, reconstructive vascular surgery) with a favorable course, but requiring a subsequent step of early hospital rehabilitation, almost no mortality, the duration of treatment is small (mortality 0.4%, bed-day 7). **Tokens:** ischemic, hypertension, angina.
- **Class 6.** Planned stenting or patients with unstable angina, mortality is almost there, stay short light patients. **Tokens:** ischemic, unstable.
- **Class 7.** A variety of emergency patients with stay above average, but almost without lethality and severe emergency pathology with a favorable prognosis. **Tokens:** coronary, acute, syndrome.

- **Class 8.** Young patients with severe transmural heart attacks, with pronounced risk factors, even longer hospitalization than in 4 Class, but with a more favorable prognosis. **Tokens:** infarction, myocardium, acute.
- **Class 9.** Patients with ACS, hospitalized at a later stage of the disease. **Tokens:** coronary, ischemic, acute.
- **Class 10.** The most stable planned patients. **Tokens:** heart, failure, cardiac.

Based on this, one can draw conclusions about the impact of certain words in the diagnosis on the length of stay and the probability of death (for example, ischemic, action, infarction to increase the mortality rate). In the same time, this classification does not correspond to the nosological forms or type of operation or category of patients. It reflects the severity of the patient's condition, the seriousness of intervention, unique features, and overall treatment strategy. Also, it may coincide in patients with various diseases and vary in patients with the same diseases (depending on the characteristics of the patient and the course of the disease). The classification does not coincide with the principles of construction of medical nosological classifications, closer to the prognostic scales, but more fully and multifaceted (with characteristics' combinations). From the cost analysis or predict future outcomes should be more accurate than the tradition nosological classification and the scale. Clinical statistical groups (flexible classes) must form not on nosology but on clusters with cost accounting.

## 4 Discussion and Conclusion

Medical processes have many aspects to be considered. For example, the diagnosis of anamnesis and patient characteristics may be different from pathway or event classification. So one need to understand how the different classifications may be compared from the point of view of medical knowledge. For this purpose, it is necessary to have accurate predictive models. From medical point of view, this approach does not allow to plan the patient's treatment strategy for the long term, but enables assessing the characteristics of the episode of care and outcomes in the short term. It is relevant for a large medical center where a lot of patients receiving different treatment, which may have a similar set of random and unrelated characteristics. The interpretation of these combinations to a certain extent conditional, in fact, is the search for matches between clusters and groups selected by the clinical approach.

Using ML methods for data processing enables selection of combinations of features to be divided into classes outside the logical framework of the clinical presentation of the patient. It needs extraordinary efforts to integrate this information into the structure of the representation of the patient, with similar interpretation, some non-obvious characteristics of the groups identified in the analysis may be lost. The logical structure of the clinical presentation of the patient is the basis for the formation of treatment strategies and clinical decision-making in the conditions of data shortage.

The current study is still in progress. It includes development of a generalized toolbox for classifying, processing, and analyzing of EHR text data. Based on the early results of the analysis of texts, it can be concluded that the proposed approach to

clustering of text data is useful. Future work includes several directions of development. For example, having information about which group a patient belongs to, it may be possible to efficiently solve the problem of prognosis (required tests, duration or type of disease), as well as to improve the model of prediction of treatment processes.

**Acknowledgements.** This work financially supported by Ministry of Education and Science of the Russian Federation, Agreement #14.575.21.0161 (26/09/2017). Unique Identification RFMEFI57517X0161.

## References

1. Metsker, O., Bolgova, E., Yakovlev, A., Funkner, A., Kovalchuk, S.: Pattern-based mining in electronic health records for complex clinical process analysis. *Procedia Comput. Sci.* **119**, 197–206 (2017)
2. Rakocevic, G., Djukic, T., Filipovic, N., Milutinović, V.: *Computational Medicine in Data Mining and Modeling*. Springer, New York (2013). <https://doi.org/10.1007/978-1-4614-8785-2>
3. Thompson, P., Batista-Navarro, R.T., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., Timmermann, C., Worboys, M., Ananiadou, S.: Text mining the history of medicine. *PLoS ONE* **11**, e0144717 (2016)
4. Pereira, L., Rijo, R., Silva, C., Martinho, R.: Text mining applied to electronic medical records. *Int. J. E-Health Med. Commun.* **6**, 1–18 (2015)
5. Gupta, A., Simaan, M., Zaki, M.J.: Investigating bank failures using text mining. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–8. IEEE (2016)
6. Suh-Lee, C., Jo, J.-Y., Kim, Y.: Text mining for security threat detection discovering hidden information in unstructured log messages. In: 2016 IEEE Conference on Communications and Network Security (CNS), pp. 252–260. IEEE (2016)
7. Septiana, I., Setiowati, Y., Fariza, A.: Road condition monitoring application based on social media with text mining system: case study: East Java. In: 2016 International Electronics Symposium (IES), pp. 148–153. IEEE (2016)
8. Landge, M.A., Rajeswari, K.: GPU accelerated Chemical Text mining for relationship identification between chemical entities in heterogeneous environment. In: 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–6. IEEE (2016)
9. Mahmoud, M.A., Ahmad, M.S.: A prototype for context identification of scientific papers via agent-based text mining. In: 2016 2nd International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR), pp. 40–44. IEEE (2016)
10. Pletscher-Frankild, S., Pallejà, A., Tsafo, K., Binder, J.X., Jensen, L.J.: DISEASES: text mining and data integration of disease–gene associations. *Methods* **74**, 83–89 (2015)
11. Bino Patric Prakash, G., Jacob, S.G., Radhameena, S.: Mining semantic representation from medical text: a Bayesian approach. In: 2014 International Conference on Recent Trends in Information Technology. pp. 1–4. IEEE (2014)
12. Urbain, J.: Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models. *J. Biomed. Inform.* **58**, S143–S149 (2015)
13. Casillas, A., Gojenola, K., Perez, A., Oronoz, M.: Clinical text mining for efficient extraction of drug-allergy reactions. In: Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016, pp. 946–952 (2017)
14. Platzter, A.: Visualization of SNPs with t-SNE. *PLoS ONE* **8**, e56883 (2013)