Data Mining from Free-Text Health Records: State of the Art, New Polish Corpus

Krištof Anetta[1,2]

[1] Natural Language Processing Centre, Faculty of Informatics, Masaryk University
Botanická 68a, Brno, Czech Republic
xanetta@fi.muni.cz
[2] Faculty of Arts, University of Ss. Cyril and Methodius
Nám. J. Herdu 2, Trnava, Slovakia
kristof.anetta@ucm.sk

Abstract. This paper deals with data mining from free-form text electronic health records both from global perspective and with specific application to Slavic languages. It introduces the reader to the promises and challenges of this enterprise and provides a short overview of the global state of the art and of the general absence of this kind of research in Central European Slavic languages. It describes pl_ehr_cardio, a new corpus of Polish health records with 18 years' worth of medical text. This paper marks the beginning of a pioneering research project in medical text data mining in Central European Slavic languages.

Keywords: EHR, electronic health records, named entity recognition, text data mining, NLP, natural language processing, Slavic languages, Polish.

1 Introduction

In recent years, as the performance of deep learning NLP approaches skyrocketed, a distinct niche of research has been gaining momentum: data mining from free-form text health records. In its short lifespan, it has already generated promising results when applied to English. This research extends the reach of this niche into Central European Slavic languages, where it has been largely absent. A large dataset of Polish health records spanning 18 years of data has been acquired and processed, forming the *pl_ehr_cardio* corpus. Apart from reviewing the general promises of data mining from health records and the global state of the art, this text also serves as an introduction to this cornerstone Polish corpus.

2 The Transformative Potential of Text-Mining Health Records

"A wealth of clinical histories remains locked behind clinical narratives in free-form text" [1] – this succinct sentence shows the key motivation behind text-mining health records. All over the world, there exist billions over billions of

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2020, pp. 13–22, 2020. © Tribun EU 2020

free-form text records of patient visits, hospitalizations, and other doctor-patient encounters, but most of this data is unstructured and allows very limited searching and processing (one estimate claims that 85 percent of actionable health information is stored in an unstructured way [2]) – all the valuable insights that big-data approaches can yield are inaccessible by default. When you consider the vastness of this database of human health right at our fingertips, the impulse to mine and structure the data is only natural (researchers have been urging development and collaboration in this niche [3]) – and with the rapid improvement of natural language processing using deep learning in the past years, mankind might finally possess the tools to do it. Civilization has always advanced on the shoulders of accumulated structured knowledge, and in the same vein, getting a grip on the world's health by leveraging decades of data with billions of cases might effect profound changes in medical science and global medical practice.

2.1 Statistics and Correlation

If properly processed into structured knowledge, databases of patient records would reveal crucial statistical information on diseases, including their early signs and effects of medicines, but also on various lifestyle-health correlations. Sample sizes, even as subsets after filtering for specific characteristics, would be orders of magnitude greater than in many contemporary clinical studies, and the researchers would be able to draw a dense network of interrelations between patient variables, symptoms, medications, and diagnoses. From the perspective of global health, standardized structured knowledge about various populations would make it easier for researchers to compare health and medical practice across the globe.

2.2 Evaluation and Prediction

An expert system leveraging health record databases as structured knowledge could also run calculations over the data and come up with entirely new judgments and estimations. It could:

- identify outliers and inconsistencies, which could discover either human error in diagnosis and prescription, or notable cases worthy of closer examination
- use deep learning to find patterns capable of predicting future outcomes of individual cases, which would open up avenues for risk group selection and better, more cost-efficient aiming of specific preventive measures.

3 Challenges

Free-form text health records exhibit several characteristics that make them more difficult to process than usual speech and writing. This is due to the hybrid nature of the text – it mixes codes and does not aspire to grammatical correctness or ease of comprehension, instead, efficiency is key and the requirement

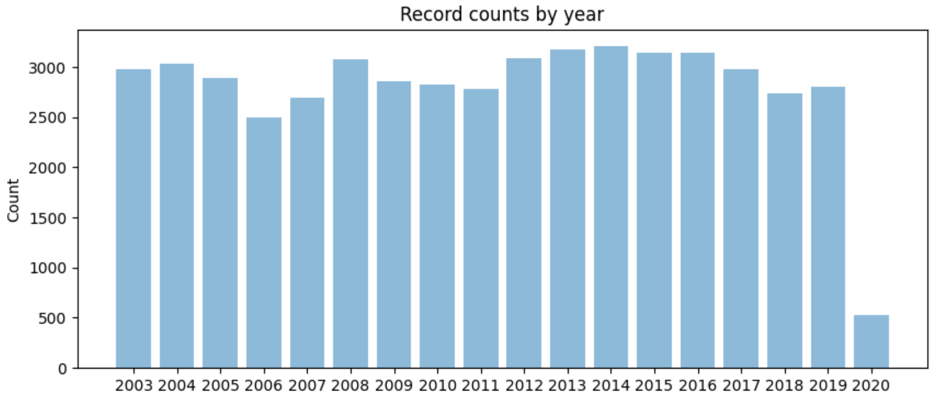

Fig. 1. Health records count distribution (*pl_ehr_cardio*)

for transparency is satisfied by being decipherable for a small group of medical experts in the respective language. The challenges involved in extracting structured data from health records include:

- Incompleteness: sentences in medical records may not correspond to standard sentence structure, missing essential syntactic elements or simply separating bits of information in a telegraphic fashion.
- Abbreviations: due to typing efficiency, abbreviating is very common, with many cases in which multiple abbreviated versions correspond to the same word.
- Bilingual text including Latin: since the meaning of medical text relies heavily on Latin words, the lexicon of the base language used for analysis needs to be extended with medical Latin. This issue has been described in [4].
- Numbers and codes: crucial information is encoded in measurements and symbolic representations, and a knowledge extraction system must be able to either determine their meaning based on form, unit or surrounding characters (such as ”=” linking it to a variable), or at least recognize to which parts of surrounding language they are connected.
- Shifted or altered word meaning because of medical context: many words from natural language change their meaning in medical text or represent categorical variables, both nominal and ordinal.
- Typing errors: since electronic health records are often produced fast and interactively, many tokens are deteriorated, and the challenge in cases such as mistyped abbreviations is often too difficult even for human readers, requiring well-trained context-based solutions.

The above clearly demonstrates that for reliable and meaningful data extraction, existing NLP tools have to be heavily customized and very specifically trained.

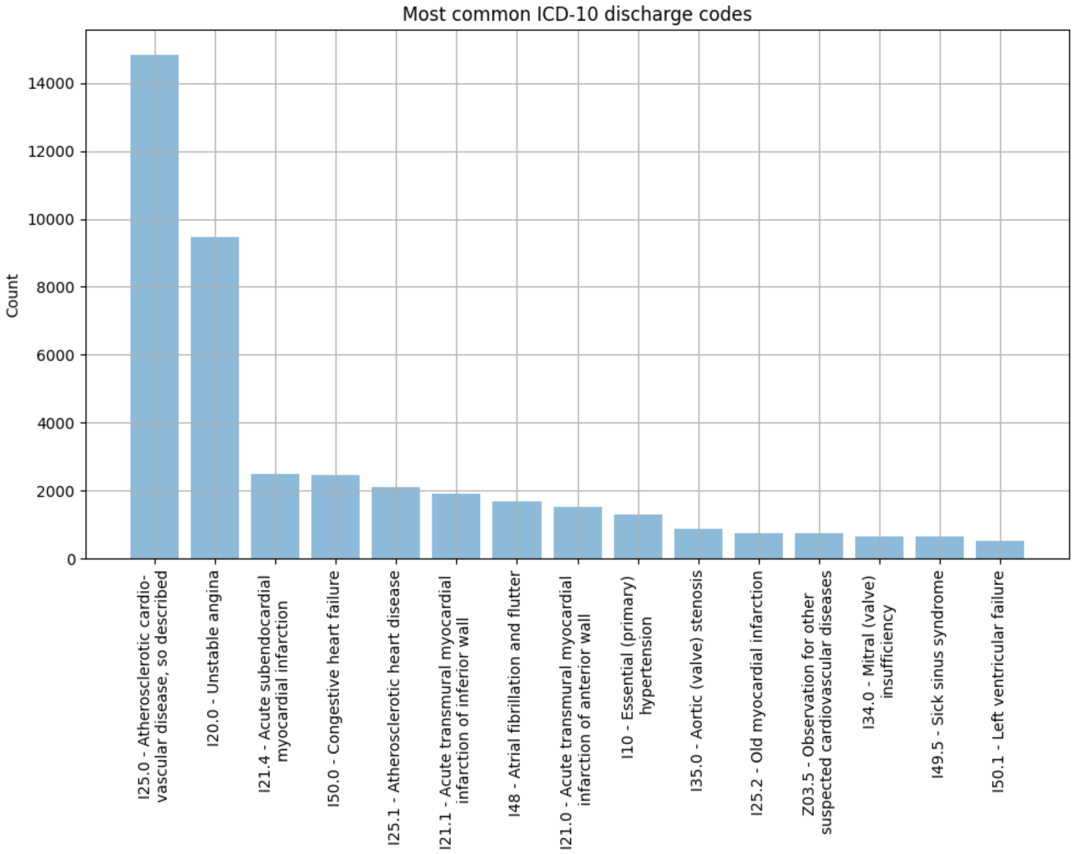

Fig. 2. Most common ICD-10 discharge codes (*pl_ehr_cardio*)

4 State of the Art, Prevailing Technology

4.1 Available Standard Frameworks

For English, several tools have been developed that directly or indirectly aid data extraction from free-form text health records:

- Apache cTAKES [5,6] is an open-source NLP system designed for extracting clinical information from the unstructured text of electronic health records. It is built using the UIMA (Unstructured Information Management Architecture) framework and Apache OpenNLP toolkit. Version 4.0 of this system was released in 2017. There have been attempts to adapt cTAKES for languages other than English [4], specifically Spanish [7] and German [8].
- MetaMap [6,9,10] is a program that maps biomedical text to the UMLS (Unified Medical Language System) Metathesaurus.

Standardized medical ontologies and term databases are an essential step towards comparability of results and interoperability – notable examples include:

- SNOMED CT [11,12] – multilingual clinical healthcare terminology containing codes, terms, synonyms, and definitions, considered to be the most comprehensive in the world. It has been employed in data extraction from health records [13].

4.2 Current Methods

Recent studies employing deep learning approaches have demonstrated that unstructured clinical notes improve prediction when added to structured data [14], similarly, the Deep Patient project has successfully included unstructured notes in its analyses [15]. Free-form text notes have proved especially useful for patient phenotyping [16]. Deep learning methods were also utilized in health record text mining for specific groups, such as those at risk of youth depression [17], prostate cancer [18], and the group of smokers [19], and also for adverse drug event detection [20,21,22].

Apart from various custom applications, convolutional neural networks [18,19], recurrent neural networks [20,22] and both uni- and bidirectional Long short-term memory (LSTM) [19,20,21,22] are notable candidates for the most widely used techniques. Some researchers also adapted BERT for clinical notes [23]. These deep learning architectures are frequently supplemented by the usage of conditional random fields (CRF) [21,22].

Overall, the tasks attempted with the above center around named entity recognition (NER) and relation extraction.

4.3 In Slavic Languages

Due to the relatively smaller size of Slavic languages, research related to them has been lagging global progress considerably, but there were notable attempts in Russian [24], Bulgarian [25], and Polish – since Polish is the subject of this research, it is worth noting the continuous efforts of a particular team [26,27,28], the most recent findings demonstrated on a corpus of more than 100,000 patient visits.

Table 1. Corpus statistics (*pl_ehr_cardio*)

Tokens	34,315,153
Words	23,831,785
Sentences	2,583,087
Average sentence length	9.226
Unique word forms	160,042
Unique word forms (lowercase)	141,685
Unique lemmas	124,727
Unique lemmas (lowercase)	114,556

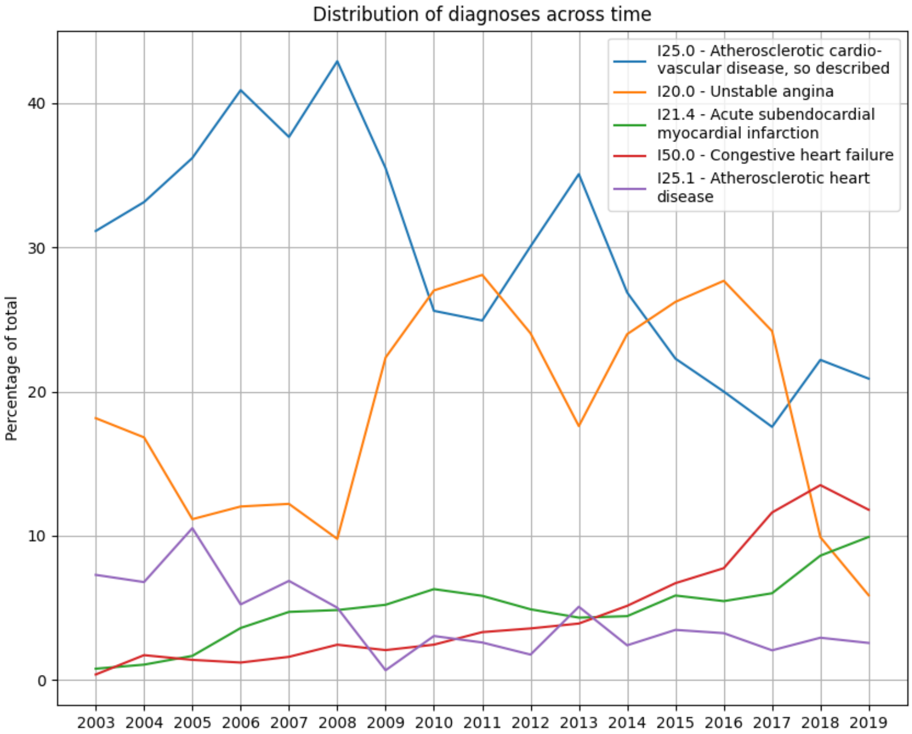

Fig. 3. Distribution of 5 most common discharge codes between 2003 and 2019 (*pl_ehr_cardio*)

5 New Corpus of Polish Health Records

The newly acquired dataset of Polish health records that forms the *pl_ehr_cardio* corpus consists of 50,465 recorded hospitalizations of cardiology patients, evenly distributed across the 17-year period between 2003 and 2019, also including partial data from 2020. Figure 1 demonstrates that years 2003 to 2019 are easily comparable in that no single year is overrepresented. After tagging these cardiology health records using the corpus management software tool SketchEngine [29][3], basic statistics were documented (Table 1)

Each record contains an ICD-10 discharge diagnosis code, which is a useful starting characteristic of the data. Figure 2 shows the most common discharge codes. Although the total number of records is quite evenly distributed over individual years, there is considerable variation in the proportions between discharge codes (Figure 3) – presumably in large part due to changes in diagnosing practice (e.g. preferred degree of specificity), although a more sensitive analysis

[3] `http://www.sketchengine.eu`

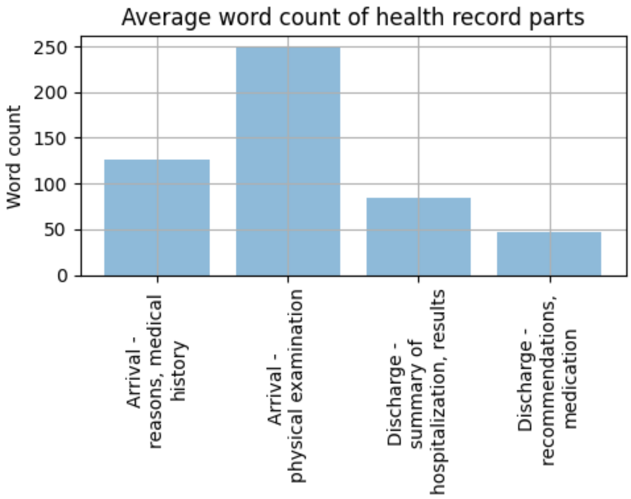

Fig. 4. Average word count of health record parts (*pl_ehr_cardio*)

might discover actual shifts in the occurrence of cardiological issues caused by shifting demographics and lifestyles.

Each record consists of 4 parts:

- Arrival: reasons, medical history (*Wywiad – Początek choroby*)
- Arrival: physical examination (*Wywiad – Badanie przedmiotowe*)
- Discharge: summary of hospitalization, results (*Epikryza – Badanie fizykalne*)
- Discharge: recommendations, medication (*Epikryza – Zalecenia lekarskie*)

Every part is written in a different style and concentrating on different concepts, and will require custom-tailored attention. Figure 4 shows that roughly half of the available text is concentrated in part 2, which is concerned with the physical examination after the patient's arrival. Word count may not exactly correspond to the amount of information present, but it gives a rough indication of the profile of the data, among others that there is ample information about symptoms and physical examination findings, which is especially valuable when correlated with diagnoses. Also, part 4 containing recommendations and medication prescriptions is usually written in a much more condensed fashion, which means that its relatively lower average word count still provides generous amounts of data on medication.

From the various challenges mentioned in the first sections, this Polish corpus does not suffer from too much Latin usage or abbreviation, but the syntax of its sentences very often leaves out elements (notably verbs) and punctuation, which complicates dependency parsing. After preliminary processing using the spaCy [30] library with pl_core_news_lg model (500,000 unique vectors), it has become obvious that named entity recognition trained on regular Polish corpora yields no useful results and it will require specific training. On the other hand, spaCy's dependency parsing correctly identified a large portion of dependencies such as nominal subject, nominal modifier, or adjectival modifier,

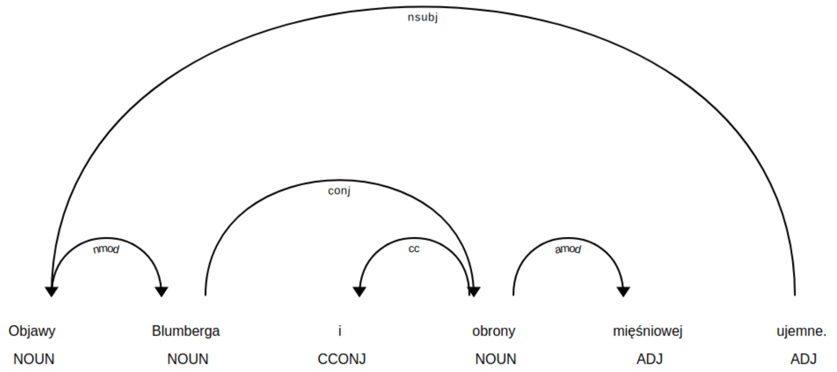

Fig. 5. Example sentence with dependencies shown in spaCy's displaCy visualizer

which will be crucial in extracting information about physical examination findings.

6 Conclusion

This paper's purpose has been twofold. First, it aimed to briefly introduce the growing niche of data mining from the unstructured text of health records including the promises, challenges, and current state of the art in this area. Arguably, this niche's growth is still in the beginnings, given the magnitude of existing data and the centrality of big data approaches to a case study-based science like medicine. For decades, this enterprise has been viewed as a major opportunity for the expansion of medical knowledge and practice, but only the advent of highly effective deep learning NLP methods did bring sufficient power to fully leverage the heaps of unstructured content.

Second, this paper used the opportunity to describe a newly formed corpus of Polish health records and thereby demonstrate some of the ideas and considerations in beginning such research on a concrete example. This dataset detailing more than 50,000 cardiology hospitalizations over 18 years will be the subject of subsequent studies, in which it will both pioneer a topic rarely broached in Slavic languages and contribute valuable descriptive and correlational information about cardiology patients, their symptoms, procedures, medication, and diagnoses to the medical community.

Acknowledgments. The Polish data was provided by the Medical University of Silesia in Katowice, Poland, with the help of Prof. Krystian Wita, MD PhD, Prof. Wojciech Wojakowski, MD PhD, Prof. Wacław Kuczmik, MD PhD, Tomasz Jadczyk, MD PhD, and Marcin Wita, MD PhD.

This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101.

References

1. Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., Osmani, V.: Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics* 7(2):e12239, (2019). https://doi.org/10.2196/12239
2. Nenadic, G.: Key Patient Information Stored in Routinely Collected Healthcare Free-text Data is Still Untapped. *Open Access Government* (2019).
3. Ohno-Machado, L.: Realizing the Full Potential of Electronic Health Records: The Role of Natural Language Processing. *Journal of the American Medical Informatics Association* 18(5), 539 (2011). https://doi.org/10.1136/amiajnl-2011-000501
4. Névéol, A., Dalianis, H., Velupillai, S. et al.: Clinical Natural Language Processing in Languages Other than English: Opportunities and Challenges. *Journal of Biomedical Semantics* 9(12), (2018). https://doi.org/10.1186/s13326-018-0179-8
5. Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., Chute, C. G.: Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications. *Journal of the American Medical Informatics Association* 17(5), 507–513 (2010). https://doi.org/10.1136/jamia.2009.001560
6. Reátegui, R., Ratté, S.: Comparison of MetaMap and cTAKES for Entity Extraction in Clinical Notes. *BMC Medical Informatics and Decision Making* 18(74), (2018). https://doi.org/10.1186/s12911-018-0654-2
7. Costumero, R., García-Pedrero, A., Gonzalo-Martín, C., Menasalvas, E., Millan, S.: Text Analysis and Information Extraction from Spanish Written Documents. In: Slezak, D., Tan, A. H., Peters, J., Schwabe, L. (eds.) *Brain Informatics and Health. BIH 2014. Lecture Notes in Computer Science*, vol 8609, pp. 188–197. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09891-3_18
8. Becker, M., Böckmann, B.: Extraction of UMLS® Concepts Using Apache cTAKES™ for German Language. *Studies in Health Technology and Informatics* 223, 71-76 (2016).
9. Aronson, A. R.: Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proceedings, AMIA Symposium*, pp. 17–21 (2001).
10. Aronson, A. R., Lang, F.: An Overview of MetaMap: Historical Perspective and Recent Advances. *Journal of the American Medical Informatics Association* 17(3), 229–236 (2010). https://doi.org/10.1136/jamia.2009.002733
11. Donnelly, K.: SNOMED-CT: The Advanced Terminology and Coding System for eHealth. *Studies in Health Technology and Informatics* 121, 279–290 (2006).
12. Benson, T., Grieve, G.: *Principles of Health Interoperability: SNOMED CT, HL7 and FHIR*. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30370-3
13. Peterson, K. J., Liu, H.: Automating the Transformation of Free-Text Clinical Problems into SNOMED CT Expressions. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 497–506 (2020).
14. Zhang, D., Yin, C., Zeng, J. et al. Combining Structured and Unstructured Data for Predictive Models: A Deep Learning Approach. *BMC Medical Informatics and Decision Making* 20, 280 (2020). https://doi.org/10.1186/s12911-020-01297-6
15. Miotto, R., Li, L., Kidd, B. et al.: Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports* 6, 26094 (2016). https://doi.org/10.1038/srep26094
16. Yang, Z., Dehmer, M., Yli-Harja, O. et al.: Combining Deep Learning with Token Selection for Patient Phenotyping from Electronic Health Records. *Scientific Reports* 10, 1432 (2020). https://doi.org/10.1038/s41598-020-58178-1

17. Geraci, J., Wilansky, P., de Luca, V. et al.: Applying Deep Neural Networks to Unstructured Text Notes in Electronic Medical Records for Phenotyping Youth Depression. *Evidence-Based Mental Health* 20, 83-87 (2017).
18. Leyh-Bannurah, S., Tian, Z., Karakiewicz, P. I., Wolffgang, U., Sauter, G., Fisch, M., Pehrke, D., Huland, H., Graefen, M., Budäus, L.: Deep Learning for Natural Language Processing in Urology: State-of-the-Art Automated Extraction of Detailed Pathologic Prostate Cancer Data From Narratively Written Electronic Health Records. *JCO Clinical Cancer Informatics* 2, 1-9 (2018) https://doi.org/10.1200/CCI.18.00080
19. Rajendran, S., Topaloglu, U.: Extracting Smoking Status from Electronic Health Records Using NLP and Deep Learning. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, pp. 507–516 (2020).
20. Wunnava, S., Qin, X., Kakar, T., Sen, C., Rundensteiner, E. A., Kong, X.: Adverse Drug Event Detection from Electronic Health Records Using Hierarchical Recurrent Neural Networks with Dual-Level Embedding. *Drug Safety* 42(1), 113-122 (2019). https://doi.org/10.1007/s40264-018-0765-9
21. Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., Ananiadou, S.: Adverse Drug Events and Medication Relation Extraction in Electronic Health Records with Ensemble Deep Learning Methods. *Journal of the American Medical Informatics Association* 27(1), 39-46 (2020). https://doi.org/10.1093/jamia/ocz101
22. Yang, X., Bian, J., Gong, Y. et al.: MADEx: A System for Detecting Medications, Adverse Drug Events, and Their Relations from Clinical Notes. *Drug Safety* 42, 123–133 (2019). https://doi.org/10.1007/s40264-018-0761-0
23. Huang, K., Altosaar, J., Ranganath, R.: ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. (2019) arXiv:1904.05342
24. Gavrilov, D., Gusev, A., Korsakov, I.: Feature Extraction Method from Electronic Health Records in Russia. In: *Proceeding of the 26th Conference of FRUCT Association*, pp. 497-500 (2020).
25. Zhao, B.: Clinical Data Extraction and Normalization of Cyrillic Electronic Health Records Via Deep-Learning Natural Language Processing. *JCO Clinical Cancer Informatics* 3, 1-9 (2019). https://doi.org/10.1200/CCI.19.00057
26. Mykowiecka, A., Marciniak, M., Kupsc, A.: Rule-based Information Extraction from Patients' Clinical Data. *Journal of Biomedical Informatics* 42(5), 923-936 (2009). https://doi.org/10.1016/j.jbi.2009.07.007
27. Marciniak, M., Mykowiecka, A.: Terminology Extraction from Medical Texts in Polish. *Journal of Biomedical Semantics* 5(24), (2014). https://doi.org/10.1186/2041-1480-5-24
28. Dobrakowski, A. G., Mykowiecka, A., Marciniak, M., Jaworski, W., Biecek, P.: Interpretable Segmentation of Medical Free-Text Records Based on Word Embeddings. In: Helic, D., Leitner, G., Stettinger, M., Felfernig, A., Raś, Z. W. (eds.) *Foundations of Intelligent Systems. ISMIS 2020. Lecture Notes in Computer Science*, vol 12117. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59491-6_5
29. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: Ten Years On. *Lexicography* 1, 7-36 (2014).
30. Honnibal, M., Montani, I.: *spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing.* (2017)