Risk Factors Extraction from Clinical Texts based on Linked Open Data

Svetla Boytcheva[1] and **Galia Angelova**[1] and **Zhivko Angelov**[2]
[1] Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Sofia, Bulgaria
svetla.boytcheva@gmail.com, galia@lml.bas.bg
[2] ADISS Lab Ltd,
4 Hristo Botev blvd., 1463 Sofia, Bulgaria
angelov@adiss-bg.com

Abstract

This paper presents experiments in risk factors analysis based on clinical texts enhanced with Linked Open Data (LOD). The idea is to determine whether a patient has risk factors for a specific disease analyzing only his/her outpatient records. A semantic graph of "meta-knowledge" about a disease of interest is constructed, with integrated multilingual terms (labels) of symptoms, risk factors etc. coming from Wikidata, PubMed, Wikipedia and MESH, and linked to clinical records of individual patients via ICD–10 codes. Then a predictive model is trained to foretell whether patients are at risk to develop the disease of interest. The testing was done using outpatient records from a nation-wide repository available for the period 2011-2016. The results show improvement of the overall performance of all tested algorithms (kNN, Naïve Bayes, Tree, Logistic regression, ANN), when the clinical texts are enriched with LOD resources.

1 Motivation

Recently, with the improving quality of Natural Language Processing (NLP), it is increasingly recognized as the most useful tool to extract clinical information from the free text of scientific medical publications and clinical records. In this way NLP becomes an instrument supporting biomedical research and new application scenarios are sought to reveal patterns and dependencies expressed by medical texts. Open-source NLP software appears, tailored to clinical text, and this increases NLP dissemination and acceptance. The construction of language resources for biomedical NLP goes in parallel to technology development. The large variety of medical terminology systems is continuously transformed and integrated into standardized, structured repositories of Linked Open Data[1]; de-identified data sets of electronic health records (EHRs) are made available as open resources[2]. Current hype in open linked data and collective efforts for their generation allow to benefit from the multilingual versions of some encyclopedic datasets like Wikidata and Wikipedia. Still there is a lack of NLP tools and linguistic resources with sufficient quality for processing medical texts in languages other than English but the interest to process such texts increases too.

Our goal is to determine whether a patient has risk factors for a specified disease, according to the information in his/her outpatient record. We suggest to enrich patient-related clinical narratives with additional information sources in order to enable a deeper investigation of dependencies between diseases and risk factors. In general it is difficult to predict the risk of a certain disease from the text of a clinical record only. Patient history contains numerous facts that are documented within a series of records but most often the medical expert reads them in isolation. In addition, many symptoms might signal various diseases. We propose to construct semantic graphs of "meta-knowledge" about diseases of interest, to integrate there multilingual terms (labels) of symptoms, risk factors etc., and to link clinical records of individual patients to this construction with the hope to discover new hints and interrelations that are not contained in the primary documents.

In the experiments presented here, patient records in Bulgarian language are enhanced with

[1] https://lod-cloud.net/

[2] E.g. at BioPortal https://bioportal.bioontology.org/ and at DBMI Data Portal https://portal.dbmi.hms.harvard.edu/

Proceedings of Recent Advances in Natural Language Processing, pages 161–167,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_019

semantic information provided by medical ontologies and other resources in English, like scientific publications and encyclopedic data. Several data mining experiments were run on datasets containing outpatient records of diabetic patients in Bulgarian linked to encyclopedic extracts and Life Sciences LOD in English. The results show that LOD infuse some relations that are not found by standard text mining techniques of clinical narratives, and thus enable the discovery of associations hinting to further risk factors for diabetes mellitus.

2 Related Work

Mining of inter-related collections of clinical texts and LOD is still rare. On the one hand, with hundreds of open biomedical ontologies and numerous biomedical datasets made available as LOD, there is a salient opportunity to integrate clinical and biomedical data to better interpret patient-related texts and to uncover associations of biomedical interest. On the other hand, such mining experiments require significant efforts to make clinical data interoperable with standardized health terminologies, biomedical ontologies and growing LOD repositories. One of the earliest papers in this direction is (Pathak et al., 2013) which describes how patient EHRs data at Mayo Clinic are represented as Resource Description Framework (RDF) in order to identify potential drug-drug interactions for widely prescribed cardiovascular and gastroenterology drugs. Some drug-drug interactions of interest were identified which suggest lack of consensus on practice guidelines and recommendations. The authors of (Odgers and Dumontier, 2015) describe how they transformed a de-identified version of the STRIDE[3] EHRs into a semantic clinical data warehouse containing among others annotated clinical notes. They showed the feasibility of using semantic web technologies to directly exploit existing biomedical ontologies and LOD. As far as NLP is concerned, an open-source tool (NegEx) is used in the EHR transformation to recognize negated terms. The integrated search in EHR data and LOD is not yet considered as a popular trend in the secondary use of clinical narratives (Meystre et al., 2017) and is still an emerging direction of research mostly due to the complex data preparation.

Information Extraction (IE) refers to the automatic extraction of concepts, entities and events as well as their relations and associated attributes from free text. A recent review of clinical IE applications (Wang et al., 2018) notes the increasing interest to NLP but lists only 25 IE systems which were used multiple times, outside the labs where they were created. Isolated attempts exist to apply IE in the context of EHR processing in frameworks for semantic search, for instance SemEHR deployed to identify contextualized mentions of biomedical concepts within EHRs in a number of UK hospitals (Wu et al., 2018). We mention the following research prototypes as experimental developments, based on some sort of IE: (Shi et al., 2017) reports about a system extracting textual medical knowledge from heterogeneous sources in order to integrate it into knowledge graphs; (Hassanpour and Langlotz, 2016) describes a machine learning system that annotates radiology reports and extracts concepts according to a model covering most clinically significant contents in radiology; (Jackson et al., 2018) presents the information extraction and retrieval architecture CogStack, deployed in the King's College Hospital. CogStack has functionality to transform records into de-identified text documents and applies generic clinical IE pipelines to derive additional structured data from free texts.

Most of the successful systems listed above work for clinical narratives in English. All major resources, ontologies and terminology classifications like UMLS[4] and MESH[5] are available in English. The comprehensive ontology SNOMED CT[6] was developed initially in English and then translated to other languages. Progress in biomedical NLP for languages other than English will catalyze the development of tools in the respective languages and will enable access to medical data presented in a variety of languages (Névéol et al., 2018). In Europe, the European commission supports the development of multilingual platforms like SEMCARE which performs queries on unstructured medical data in English, German, and Dutch (López-García et al., 2016).

Using Big Data (nowadays - millions of EHRs) to advance medical research and health care prac-

[3] Stanford Translational Research Integrated Database Environment including a repository for EHR data, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815452/

[4] https://www.nlm.nih.gov/research/umls/

[5] https://meshb.nlm.nih.gov/search

[6] http://www.snomed.org/

tices is now on the rise (Kessel and Combs, 2016). Core NLP components are already embedded in general clinical platforms similar to CogStack (Jackson et al., 2018). Development of high quality corpora and terminology is a key factor for NLP progress in smaller languages. Here we employ English terminology in data mining tasks concerning EHRs in Bulgarian language.

3 Materials

The datasets used in this study are a blend between LOD and clinical texts in Bulgarian language that belong to the Repository underpinning the Bulgarian Diabetes Register.

The Register was automatically generated in 2015 from 260 million pseudonymized outpatient records (ORs) provided by the National Health Insurance Fund (NHIF) for the period 2011–2014 for more than 5 million citizens yearly, more than 7 million citizens in total (Boytcheva et al., 2017). Updated twice with data about 2015 and 2016, today the Register is maintained by the University Specialized Hospital for Active Treatment of Diabetes (USHATE) - Medical University Sofia. At present the Repository of ORs, which underpins the Register, contains about 262 million records. These are reimbursement requests submitted by General Practitioners and Specialists from Ambulatory Care after every contact with a patient. The average number of patients with Diabetes Mellitus Type 2 (T2DM) per year is about 450,000.

In the primary database, from where we extract our datasets, the ORs are stored as semi-structured files with predefined XML-format. Administrative information is structured: visit date and time; pseudonymized personal data and visit-related information, demographic data etc. All diagnoses are given by ICD–10[7] codes and location names are specified in Bulgarian according to a standard nomenclature. However much information is provided as free text: anamnesis (case history, previous treatments, often family history, risk factors), patient status (summary of patient state, height, weight, body-mass index, blood pressure etc.), clinical tests (values of clinical examinations and lab data listed in arbitrary order) as well as prescribed treatment (codes of drugs reimbursed by NHIF, free text descriptions of other drugs).

To enhance clinical information with semantic data related to diagnoses, risk factors and symptoms, the following open datasets are selected:

- Wikidata[8] - contains multilingual encyclopedic information. Wikidata is a trusted resource, providing multilingual terminologies, their association with MESH codes, and complex relations between diagnoses, risk factors, and symptoms. Currently Wikidata contains descriptions of 5,227 items included in ICD–10 and 10,517 descriptions of items included in ICD–10–CT. The main problem is that many duplicated entities exist. For instance, for ICD–10 code I20 there are two items "angina pectoris (Q180762)" and "ischaemic heart disease (Q1444550)". Using SPARQL[9] queries, from Wikidata we collect for a given diagnosis all risk factors related to it as well as the associated MESH codes. From the list of risk factors that is originally in English we produce also a list in Bulgarian for the corresponding terms.

- PubMed[10] – the largest collection of scientific publications in the area of biomedicine and life sciences. From Pubmed we automatically extract publication abstracts and related MESH terms via advanced queries[11] through API. The search is limited to 10,000 abstracts in order to keep balance between the amounts of clinical narratives and texts of scientific publications.

- Wikipedia – from Wikipedia we extract automatically Wikipedia pages' summaries for a specified query via MediaWiki RESTful web service API[12]. The information in Wikipedia is encyclopedic and more broader, thus the semantic information there is too vague and shallow, in contrast to PubMed abstracts.

- MESH ontology – this ontology is chosen because both Pubmed publications and Wikidata contain references to it. In addition a mapping between MESH and SNOMED CT is available.

[7] http://apps.who.int/classifications/icd10/browse/2016/en#/

[8] https://www.wikidata.org/wiki/Wikidata:Main_Page

[9] https://www.w3.org/TR/rdf-sparql-query/

[10] https://www.ncbi.nlm.nih.gov/pubmed/

[11] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2651214/

[12] https://www.mediawiki.org/wiki/API:Tutorial

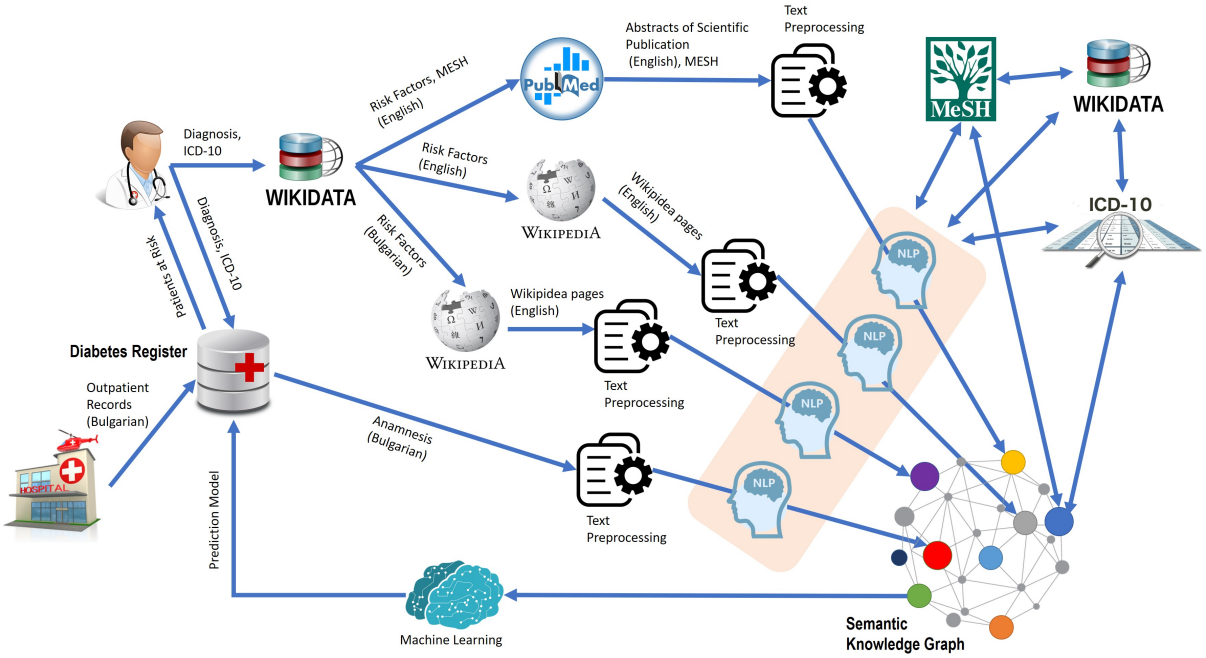

Figure 1: Pipeline for identification of patients at risk

4 Methods

The proposed method for risk factors identification is based on LOD and benefits from mapping multilingual data and using their vocabularies.

The data flow diagram is shown on Fig. 1. The process starts with selection of a diagnosis D, according to ICD–10 or ICD–10–CD by a medical expert who is looking for patients at risk in the Diabetes Register. The next step is to extract corresponding risk factors and symptoms for D in English $RE = \{re_1, re_2, ..., re_n\}$ from Wikidata, their equivalent terms in Bulgarian $RB = \{rb_1, rb_2, ..., rb_n\}$ and the MESH codes $M = \{m_1, m_2, ..., m_n\}$.

For each term re_i in English and its corresponding rb_i in Bulgarian, summaries of the respective Wikipedia pages are extracted automatically.

For each Mesh code m_i of risk factor r_i are extracted up to 10,000 abstracts of Pubmed publications and their annotations with MESH codes. For Pubmed the advanced search is done using an automatically generated query in the form:

```
(D/pc [majr] OR D/di [majr] OR
   D/ep [majr] D [mh]) AND
( re1 [mh] OR ... OR ren [mh])
```

where the MeSH qualifiers for subheadings are: "pc" refers to "prevention and control"; "di" means "diagnosis"; "ep" is "epidemiology", "mh" - MeSH heading, and "majr" - to search MeSH heading that is a major topic of an article.

From the Bulgarian Diabetes Register a dataset is excerpted for patients with the diagnosis D. For those with recent D onset, ORs for previous periods are collected (only within 2011–2016).

4.1 Text Pre-Processing

The main transformations are done stepwise:

- tokenization - for Bulgarian language we used the UDPipe tokenizer[13].
- convertion of all words to lower case;
- removal of all punctuation marks;
- removal of all numbers;
- application of a stemmer and lemmatizer - for Bulgarian the UDPipe lemmatizer, for English Porter Stemmer (Porter, 2006);
- filtering stopwords - both for Bulgarian[14] and English;
- application of text vectorization based on TFIDF.

[13] https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-train.html#support_in_text_mining

[14] http://bultreebank.org/wp-content/uploads/2017/04/BTB-StopWordList.zip

4.2 Semantic Model

Semantic Knowledge Graphs are used recently as powerful representation of entities and relations between them (Paulheim, 2017). Often knowledge graphs are generated automatically from semi-structured resources or from the documents underpinning various ontologies, through terms/words they contain, by a combination of linguistic and statistical methods (Grainger et al., 2016).

In our experiment, all data are interlinked in a Semantic Knowledge graph via MESH codes and ICD–10 codes. Wikidata is the mediator between all resources providing cross-lingual ontology information and mapping between MESH and ICD-10 codes. Term mappings from MeSH to ICD-10 are 1,535 and to ICD-10-CM are 2,127. In addition Wikidata provides multilingual vocabulary for symptoms and diseases, and Pubmed publications are annotated by Mesh codes.

For each symptom and risk factor, related to the selected diagnosis D, the system identifies the most significant words related to D from the Wikipedia and Pubmed datasets respectively, using p-value as a measure for their significance. The knowledge graph is enriched with relations between these terms. The main relation between clinical texts and other resources is based on ICD-10 codes and some symptoms and risk factors that are presented in the anamnesis section of ORs.

4.3 Predictive Model

Two types of clinical texts are used as training datasets - for patients that have the diagnosis D, and for patients that do not have D. After text preprocessing, semantic hashing of all clinical texts in both datasets is done for predefined size of the hash. Two predictive models are applied:

- Based on the ORs information only
- Based on the ORs information enhanced with semantic data for symptoms and risk factors. In this case the vector space is extended; not only the dimensions of semantic hash vectors are used, but also additional dimensions for all symptoms, risk factors and the most significant terms related to them.

Several machine learning techniques were used to train the predictive model, including Naïve Bayes (NB) (McCallum et al., 1998), kNN, Tree, Logistic Regression and Artificial Neural Networks. (Dreiseitl and Ohno-Machado, 2002).

5 Experiments and Results

The diagnosis with ICD–10 code I20 "Ischaemic heart disease" is chosen for experiments because the Diabetes Register contains a plenty of clinical descriptions about this case. Patients with T2DM are at higher risk for developing I20 which is one of the T2DM complications. Sets of symptoms and risk factors for I20, both in Bulgarian and English, and English MeSH codes are automatically extracted by Wikidata queries. Seven symptoms are extracted: angina pectoris, nausea, dyspnea, lightheadedness, unstable angina, neck pain, fatigue. Only for three of them there are labels in Bulgarian, and for five of them there are MESH codes. In addition, there are 18 risk factors and for 11 of them labels in Bulgarian exist. Multiple MeSH codes are associated to some risk factors but there diagnoses without associated MeSH code. The final cardinality of the generated term set is $|RE| = 24$, $|RB| = 14$ and $|M| = 34$.

The Wikipedia API extracts 87 documents for a query with the RB terms in Bulgarian and we limit the set to the top 5 related documents. They contain 14,600 tokens from 3,627 types. Although only the top 5 most related documents are taken into consideration, some of the extracted Wikipedia pages are not directly related to the symptoms and risk factors, as they discuss e.g. herbs and medications for treatment. But some very related symptoms are included: for example for "nausea" the Wikipedia page about "vomiting" is extracted, and for "smoking" pages about "tobacco" and "pipe" are found. In addition, some barely related pages are extracted – mainly about some famous people, who suffer from the diseases in question and have related symptoms. Unfortunately the information in Wikipedia categories "Medical conditions" and "Diseases" for Bulgarian is too limited. For "Medical conditions" there are only 41 pages, and for many diseases the articles in the Bulgarian Wikipedia are stubs or some pages are not tagged in the respective categories.

For the query with RE terms 146 documents are identified that contain 40,099 tokens from 3,803 types. The extracted Wikipedia pages in English are also sometimes noisy and unrelated mainly due to the ambiguity e.g. pages about "Nausea(novel)" and "Nausea(band)", or "Insomnia (2002 film)" are extracted too. Other pages, indirectly related to the risk factors, contain information about Health organizations for treatment

of the respective diseases, about diagnostic procedures, medication for the treatment etc.

Using the Pubmed advanced search, the generated query with RE terms identified 67,103 related scientific papers, from which we retrieved a subset of 2,000 abstracts only. The MeSH headings only contain 104,363 tokens from 2,509 types. Both MeSH heading and Pubmed abstracts contain 546,772 tokens from 12,684 types.

Despite the imperfection of all texts extracted from Wikipedia and PubMed, the most significant terms related to the predefined subsets of symptoms and risk factors are sound and correct. Their identification is based on bag–of–words and calculation of p-value ($p \leq 0.01$) and False Discovery Rate ($FD \leq 0.2$). For example, for тютюнопушене (tobacco smoking) the following words are identified as relevant: тютюнопушене (tobacco smoking), дим (smoke), пушене (smoking), тютюнев (tobacco), практикувам (practice), пристрастяване (addiction), пушач (smoker), цигара (cigar). These words were selected among 3,588 words in the text of related Wikipedia pages in Bulgarian. From those words 754 were filtered as relevant, and the final selection contain 9 words as most significant.

Two subsets of ORs (Anamnesis section) are extracted from the Repository behind the Diabetes Register: $S1$ for 36,580 patients with diagnosis I20 and $S2$ for 86,000 patients without diagnosis I20. All clinical texts are preprocessed. The total number of tokens in $S1$ and $S2$ is 123,258 from 25,086 types. For experiments are used kNN (5 neighbours, Mahalanobis metric), Tree (Pruning: at least 2 instances in leaves, at least 5 instances in internal nodes, maximum depth 100; Stop splitting when majority reaches 95%), Neural Network (30 hidden layers, Rectified Linear Activation Function (ReLu) (Nair and Hinton, 2010), stochastic gradient-based optimizer Adam (Kingma and Ba, 2014)), $\alpha = 0.0004$, Max iterations 200, replicable training) , NB and Logistic Regression (Ridge L2). The baseline results of the prediction model based on ORs only are presented in Table 1.

The results of prediction models trained with enhanced LOD data (Table 2) show improvement of the overall performance of all algorithms on this task, especially NB and Logistic Regression, when the clinical texts are enriched with additional information provided by open data resources and medical terminologies.

Model	F1	P	R
kNN	0.796	0.795	0.801
Tree	0.778	0.776	0.783
Neural Network	0.705	0.713	0.735
NB	0.588	0.615	0.701
Logistic Regression	0.581	0.640	0.703

Table 1: Baseline: Performance of employed ML algorithm using semantic hashing over ORs only.

Model	F1	P	R
kNN	0.819	0.752	0.899
Tree	0.893	0.858	0.932
Neural Network	0.743	0.746	0.760
NB	0.823	0.704	0.989
Logistic Regression	0.825	0.703	0.999

Table 2: Performance of employed ML algorithm using semantic hashing of ORs enhanced with semantic model data.

6 Conclusion and Further Work

The proposed approach shows how clinical texts can be enhanced with additional information about the diseases, their symptoms and risk factors. The experimental results show promising improvement of the risk factors prediction accuracy. Still there is a problem with Latin medical terminology that is often used in the Bulgarian clinical texts. Another issue is the imperfection of the additional terms provided by the LOD resources, due to many ambiguous terms included there. As future work we plan to apply word sense disambiguation to the texts extracted from open resources and more precise methods for constructing the relations in the semantic knowledge graphs. As future work we are planning to do deep analysis of the individual contribution of each new term added to the clinical texts. Another direction for future work is to use some transfer learning methods like UMLfit (Howard and Ruder, 2018), BERT (Devlin et al., 2018) and XLnet (Yang et al., 2019) to train models for word embedding on clinical texts in Bulgarian.

Acknowledgments

This research is partially funded by the Bulgarian Ministry of Education and Science, grant DO1-200/2018 'Electronic health care in Bulgaria' (e-Zdrave) and the Bulgarian National Science Fund, grant DN-02/4-2016 'Specialized Data Mining Methods Based on Semantic Attributes' (IZIDA). We are grateful to anonymous reviewers for useful comments and suggestions.

References

Svetla Boytcheva, Galia Angelova, Zhivko Angelov, and Dimitar Tcharaktchiev. 2017. Integrating data analysis tools for better treatment of diabetic patients. *CEUR Workshop Proceedings* 2022:229–236.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Stephan Dreiseitl and Lucila Ohno-Machado. 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics* 35(5-6):352–359.

Trey Grainger, Khalifeh AlJadda, Mohammed Korayem, and Andries Smith. 2016. The semantic knowledge graph: A compact, auto-generated model for real-time traversal and ranking of any relationship within a domain. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pages 420–429.

Saeed Hassanpour and Curtis Langlotz. 2016. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine* 66:29–39.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* .

Richard Jackson, Ismail Kartoglu, Clive Stringer, et al. 2018. Cogstack - experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital. *BMC Medical Informatics and Decision Makingvolume* 18.

Kerstin A Kessel and Stephanie E Combs. 2016. Review of developments in electronic, clinical data collection, and documentation systems over the last decade–are we ready for big data in routine health care? *Frontiers in oncology* 6:75.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Pablo López-García, Markus Kreuzthaler, Stefan Schulz, Daniel Scherr, Philipp Daumke, Kornél G Markó, Jan A Kors, Erik M van Mulligen, Xinkai Wang, Hanney Gonna, et al. 2016. Semcare: Multilingual semantic search in semi-structured clinical data. In *eHealth*. pages 93–99.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*. Citeseer, volume 752 (1), pages 41–48.

SM Meystre, Christian Lovis, T Bürkle, G Tognola, A Budrionis, and CU Lehmann. 2017. Clinical data reuse or secondary use: current status and potential future progress. *Yearbook of medical informatics* 26(01):38–52.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. pages 807–814.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics* 9(1):12.

David J Odgers and Michel Dumontier. 2015. Mining electronic health records using linked data. *AMIA Summits on Translational Science Proceedings* 2015:217.

Jyotishman Pathak, Richard C Kiefer, and Christopher G Chute. 2013. Using linked data for mining drug-drug interactions in electronic health records. *Studies in health technology and informatics* 192:682.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web* 8(3):489–508.

Martin F Porter. 2006. An algorithm for suffix stripping. *Program* .

Longxiang Shi, Shijian Li, Xiaoran Yang, Jiaheng Qi, Gang Pan, and Binbin Zhou. 2017. Semantic health knowledge graph: Semantic integration of heterogeneous medical knowledge and services. *BioMed research international* 2017.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics* 77:34–49.

Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, et al. 2018. Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association* 25(5):530–537.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* .