Enrichment of EHR with Linked Open Data for Risk Factors Identification

Svetla Boytcheva
svetla.boytcheva@gmail.com
Institute of Information and
Communication Technologies,
Bulgarian Academy of Sciences
Sofia

Galia Angelova
galia@lml.bas.bg
Institute of Information and
Communication Technologies,
Bulgarian Academy of Sciences
Sofia

Zhivko Angelov
angelov@adiss-bg.com
Adiss Lab Ltd.
Sofia

Dimitar Tcharaktchiev
dimitardt@gmail.com
Medical University Sofia, University
Specialized Hospital for Active
Treatment of Endocrinology
Sofia

Vlayko Vodenicharov
vlayko_vodenicharov@abv.bg
Medical University Sofia, Faculty of
Medicine, Department of Hygiene
Sofia

ABSTRACT

This paper presents experiments in risk factors analysis based on clinical texts and related Linked Open Data. Enhancements with additional data sources can enrich patient data and allow for a deeper investigation of correlations. In order to explore the potential of this approach several experiments were run on data collections, extracted from a large, nationwide repository of outpatient records. Subclouds from the multilingual Geonames, Life Sciences Linked Open Data and DBpedia are used as additional sources. Sophisticated data mining in knowledge graphs finds frequent patterns of linked data items from the in–house clinical repository and the publicly available sources. The results show that Linked Open Data 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.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; • **Applied computing → Health informatics**.

KEYWORDS

Data Mining, Text Mining, Health Informatics, Big Data Analytics, Linked Open Data, Semantic Web, Knowledge Discovery

ACM Reference Format:
Svetla Boytcheva, Galia Angelova, Zhivko Angelov, Dimitar Tcharaktchiev, and Vlayko Vodenicharov. 2019. Enrichment of EHR with Linked Open Data for Risk Factors Identification. In *Proceedings of the 20th International Conference on Computer Systems and Technologies (CompSysTech '19), June 21–22, 2019, Ruse, Bulgaria.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3345252.3345290

CompSysTech '19, June 21–22, 2019, Ruse, Bulgaria
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7149-0/19/06…$15.00
https://doi.org/10.1145/3345252.3345290

1 MOTIVATION

Biomedicine is known as a data-intensive domain using multiple types of data in heterogeneous formats and from different sources, such as Electronic Health Records (EHR), clinical images and reports, genome data and others. Advanced semantic technologies are employed for the organization of available data in well-defined structures with the aim to facilitate storing, integrating and sharing data and knowledge. Linked Data (LD) as well as Linked Open Data (LOD)[1] provide a standardized means to define the association and characterization of any kind of data in the form of links reinforcing our tools to represent and manage knowledge. However, the systematic usage of LD and LOD requires a strong commitment because creating linked data resources with sound and comprehensive description of their meaning is a complex and subjective process. According to the best practices, publishers should refer to terms from widely-used vocabularies in order to ease the interpretation of their data [18]. On the other hand, recent reviews report that in life sciences most vocabulary terms (66.67%) are not dereferencable [2], which hinters the linkage of data items. Despite all complications the number of linked datasets is growing and huge Biomedical Sciences Research Infrastructures emerge. The appearance of large LOD resources further accelerates the development because the researchers see clearly the benefits of enhancing biomedical datasets with publicly available information.

This paper considers how integrated LOD resources help to overcome the language barrier and better explicate risk factors. It is well known that only small part of the risk factors (approx. 10%) is described in clinical narratives – mainly clinical factors. But more significant risk factors are the genetic ones (approx. 30%) and the exogenous factors (approx. 60%). Unfortunately information about the latter two is rarely discussed in clinical texts. Here we show how linking names of geographic locations from patient records to names of geographic locations in Geonames LOD adds new features describing the environment where patients live. The patient records

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

in this experiment come from a Repository of outpatient records submitted to the Bulgarian National Health Insurance Fund in 2016 by all General Practitioner and Specialists from the Ambulatory Care in Bulgaria. In this study we deal with some disorders related to the respiratory system.

The paper is structured as follows. Section 2 summarizes some related work about knowledge graphs and application of LOD as a novel tool for improving biomedical analytics. Section 3 presents the materials we use and Section 4 describes the methods. The experiments and results are discussed in Section 5. Section 6 contains the conclusion and plans for future work.

2 RELATED WORK

Knowledge graphs (KG) were introduced by Google in 2012 as a less formal representation of interlinked data which significantly enhances the search queries [20]. As in typical semantic networks, the KG nodes represent labeled concepts and the edges represent semantic relations between nodes. KGs encode world knowledge and can help to automatically identify the entities and relations in a natural language (NL) text. Existing KGs are the largest open conceptual resource that provides support for semantic text interpretation. Some researchers propose to use KGs as a representation model of medical information. For instance, [19] suggests to automatically retrieve entities from Electronic Health Records (EHRs) texts to knowledge graphs. The authors introduce a contextual inference pruning algorithm to explore complex semantics between entities in a chain inference. The results have relatively low accuracy due to the lack of standard Chinese medical terminology. Recently KGs were deployed as a technology that improves patient care and revolutionizes prediction and prevention [1]. More than 180 different life science and health care taxonomies and ontologies are interlinked at the core of the Knowledge Graph. Various raw data about patients are integrated in the platform as well. The claim is that KGs provide a much more efficient way to find patterns and use them for improving patient outcomes, in some cases prediction of failures with more than 70% accuracy, 2 days in advance.

Linked Open Data are Linked Data released under an open license, which does not impede its reuse for free, such as DBpedia, Wikidata and others. The topic is relatively new so many papers discuss the available resources with focus on LOD and the process of their development. The Life Sciences Linked Open Data (LSLOD) cloud was created in 2004 using the terminology of the Unified Medical Language System (UMLS) [3]. LSLOD currently contains 1,234 datasets with 16,136 links. Recently this Cloud provides basic knowledge for a variety of research experiments. In 2011 the Health Care and Life Science Interest Group of W3C[2] supported the development of a data infrastructure for pharmaceutical research, called Linked Open Drug Data, which enables links and easy search across open data sources in order to identify novel and meaningful correlations and mechanisms. Twelve open-access datasets relevant to pharmaceutical research were made available as Linked Data [17]. In general Biomedical Linked Data has now more than 10 billion links [3] connecting entities in diverse topics, including medicine, drug, symptom, gene, and others (although the limited amount of links is viewed as a weakness together with the difficulty to implement federated queries).

As shown above, extensive drug-related resources were recently transformed to LD. Due to this reason many research experiments deal with drugs. The paper [15] presents a proof-of-concept system that transforms patient data stored in Mayo's clinic enterprise warehouse to RDF and provides federated querying to this data integrated with Drug-Drug Interaction information from DrugBank. The results demonstrate the benefits of interlinking and querying multiple, heterogeneous public Web sources with private, institution-specific patient information. The authors of [11] study how negative food-induced interactions with drugs vary from one part of the world to another. Two datasets (drug data and recipes data) are transformed and connected as LD. The results show that North American and most European cuisines have negative interactions with drugs from the category "Antiinfectives for systemic use" while the cuisines from Southern Europe, Asia, Latin America and Africa negatively interact mostly with drugs from the categories "Blood and blood forming organs" and "Various". A graph analytics method, inspired by the Apriori algorithm for association mining, is presented in [12]. It identifies frequent substructures in the narratives in the Adverse drug reactions (ADR) reporting system of the US Federal Drug Agency. In order to provide explanations about the discovered drug–ADR associations in a systematic manner, the authors integrate data from four different sources including LSLOD and evaluate the proposed approach against existing pharmacovigilance methods for three different validation sets. Some of the discovered substructures are known while other findings need to be further validated by a domain expert. The authors claim that the pattern-based querying can bring together pharmacological knowledge existing in isolated, heterogeneous sources (e.g. spontaneous reporting systems) and provide mechanistic explanations behind the detected drug-drug interactions and related ADRs.

Interesting technological considerations demonstrate how to employ LOD in biomedical domain. The paper [9] uses linked data from the BioPortal system to create a navigation structure within the patterns obtained from sequential pattern mining, thus supporting the exploration of trajectories of diagnoses and treatments according to different medical classifications. Another proposal is that clinical data in hospital and medical centers can be made interoperable by using standardized health terminologies, biomedical ontologies, and growing networks of Linked Open Data [14]. The authors transformed a de–identified version of the Stanford's STRIDE database into a semantic LD clinical data warehouse containing visits, labs, diagnoses, prescriptions, and annotated clinical notes and demonstrate its utility. The paper [13] presents a fact repository for causal chains of diseases, based on a disease ontology and abnormality ontology, which is developed as linked data (1,554 diseases and 7,080 abnormal states). A navigation system called Disease Compass provides browsing the causal chains as well as general linked data such as DBpedia and 3D anatomical images. Thus the disease definition answers questions such as what abnormal state causes a disease or how might the disease advance, and what symptoms may appear.

Finally we note that all initiatives related to the development of biomedical linked data are based on English biomedical terminology.

[2] https://www.w3.org/2011/09/HCLSIGCharter

[3] http://linkedlifedata.com/sources.html

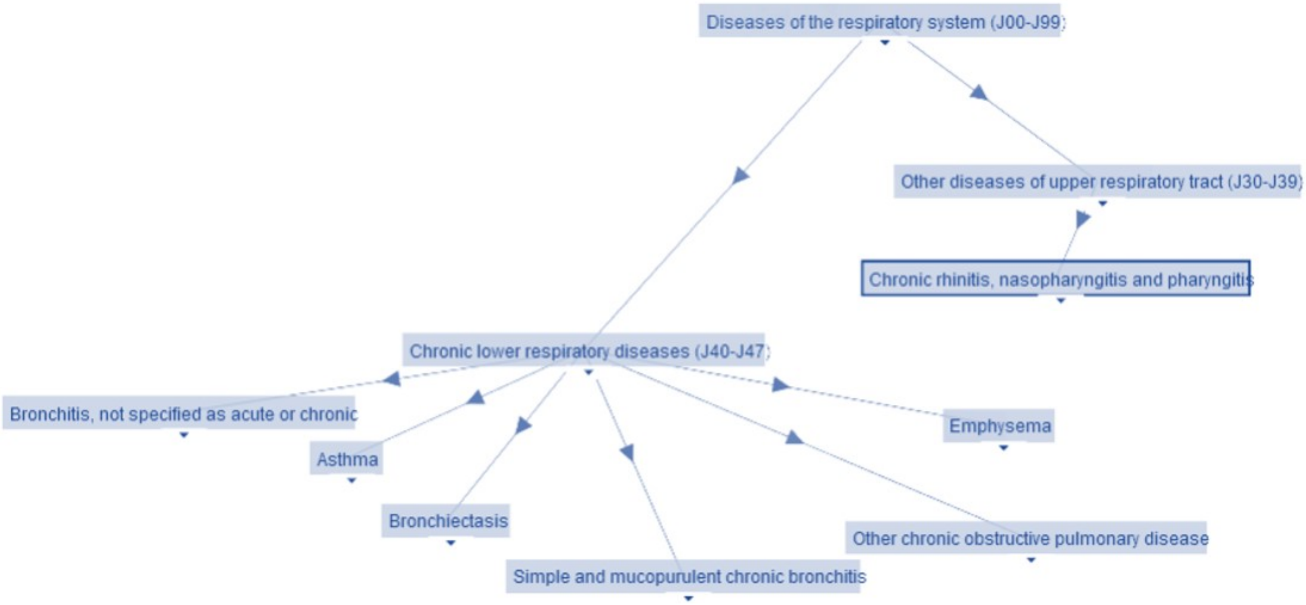

Diseases of the respiratory system (J00-J99)
Other diseases of upper respiratory tract (J30-J39)
Chronic rhinitis, nasopharyngitis and pharyngitis
Chronic lower respiratory diseases (J40-J47)
Bronchitis, not specified as acute or chronic
Asthma
Bronchiectasis
Simple and mucopurulent chronic bronchitis
Other chronic obstructive pulmonary disease
Emphysema

Figure 1: ICD-10 hierarchy of some diseases of respiratory system (J30 and J40-J45) generated by BioPortal visualization tool

Special efforts are required to involve these resources in research tasks handling terminology in languages other than English.

3 MATERIALS

In this study we use a Repository of about 262 million pseudonymized outpatient records (ORs) submitted to the Bulgarian National Health Insurance Fund (NHIF) in the period 2010–2016 for more than 7 million Bulgarian citizens in total, about 5 million citizens yearly. The NHIF collects for reimbursement purpose all ORs produced by General Practitioners and Specialists from Ambulatory Care for every patient visit.

This Repository was provided to the University Specialized Hospital for Active Treatment of Endocrinology – Medical University Sofia, as a primary dataset for automatic generation of the Bulgarian National Diabetes Register [4]. In the primary archive, ORs are stored as semi-structured files with predefined XML-format. Information needed for health care management is structured: visit date and time; pseudonymized personal data and visit-related information, demographic data (age, gender, and demographic region), locations etc. All diagnoses are given by ICD–10[4] codes and location names are specified according to a standard nomenclature. Here we consider only ORs for visits in 2016 (in total the ORs of 5,187,207 citizens). The selected subset of patients with disorders of the respiratory system contains 427,160 patients. Relevant branches of the ICD–10 classification are shown in Figure 1.

Two LOD resources are used to enrich the information available in the Repository of ORs. The first one is the multilingual Geonames[5] with rich information about geo locations, which is important for prediction and exploration of risk factors triggered by environmental characteristics. In our case the location names in the ORs are given in Bulgarian language, so via Geonames we can relate the location to its English description and access all attributes available there. Another LOD resource is Bioportal[6] which contains many ontologies related to the human body and health. One of them is integrated in our experiment – the classification of diseases - ICD10CM[7]. ICD is also multilingual, so the primary Bulgarian codes in the input ORs are easily transferable to the English version. Practically the LOD resources help linking patient records in Bulgarian language to English biomedical terminology. Further public sources are Wikipedia and DBPedia.

4 METHODS

The proposed method follows an established pipeline for knowledge discovery in data bases (KDD) [16] shown in Figure 2. The pipeline starts with data selection, then data preprocessing and transformation into RDF triples, linkage to LOD and continues with data mining of the enriched EHRs, and finally risk alerts are triggered on the basis of patterns interpretation.

The implementation is based on the integration of several tools and algorithms. A free version of GraphDB[8] is used as a framework for the first three steps. GraphDB provides a tool called Ontorefine for internal and external datasets integration and reconciliations into semantic knowledge graph. The algorithm FP-Growth for frequent patterns mining [7] is applied at the data mining step. The final interpretation and selection of the risk factors is made by applying semi-automated process for hypothesis generation.

[4] http://apps.who.int/classifications/icd10/browse/2016/en#/
[5] https://www.geonames.org/
[6] http://sparql.bioontology.org/
[7] http://bioportal.bioontology.org/ontologies/ICD10CM
[8] https://www.ontotext.com/products/graphdb/

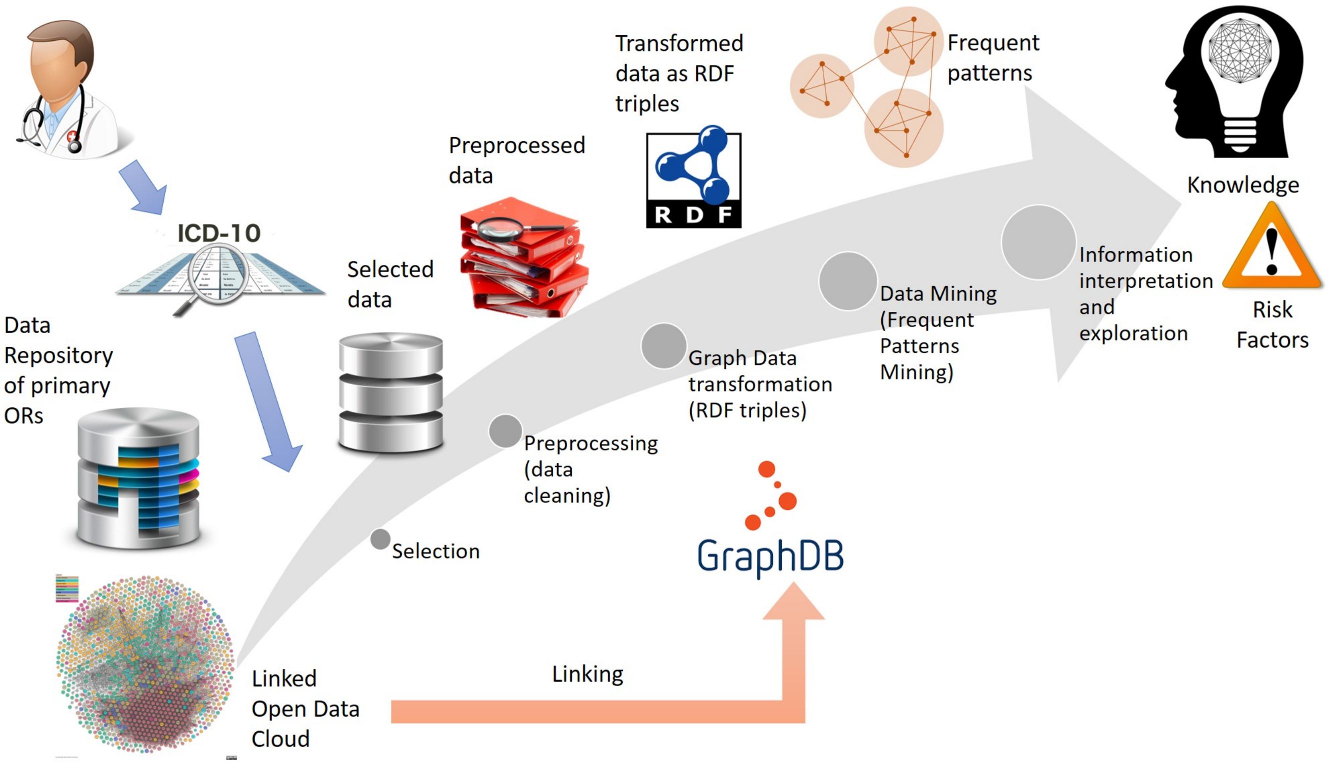
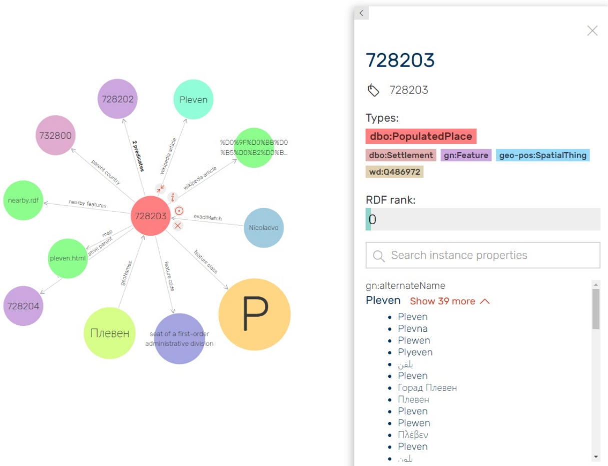

Transformed data as RDF triples
Frequent patterns
Knowledge
Preprocessed data
RDF
ICD-10
Selected data
Information interpretation and exploration
Risk Factors
Data Mining (Frequent Patterns Mining)
Data Repository of primary ORs
Graph Data transformation (RDF triples)
Preprocessing (data cleaning)
Selection
GraphDB
Linked Open Data Cloud
Linking

Figure 2: Risk Factors Identification Pipeline



728203

728203

Types:
dbo:PopulatedPlace
dbo:Settlement gn:Feature geo-pos:SpatialThing
wd:Q486972

RDF rank:
0

Search instance properties

gn:alternateName
Pleven Show 39 more

- Pleven
- Plevna
- Plewen
- Plyeven
- بلفن
- Pleven
- Горад Плевен
- Плевен
- Pleven
- Plewen
- Πλέβεν
- Pleven
- بلوِن

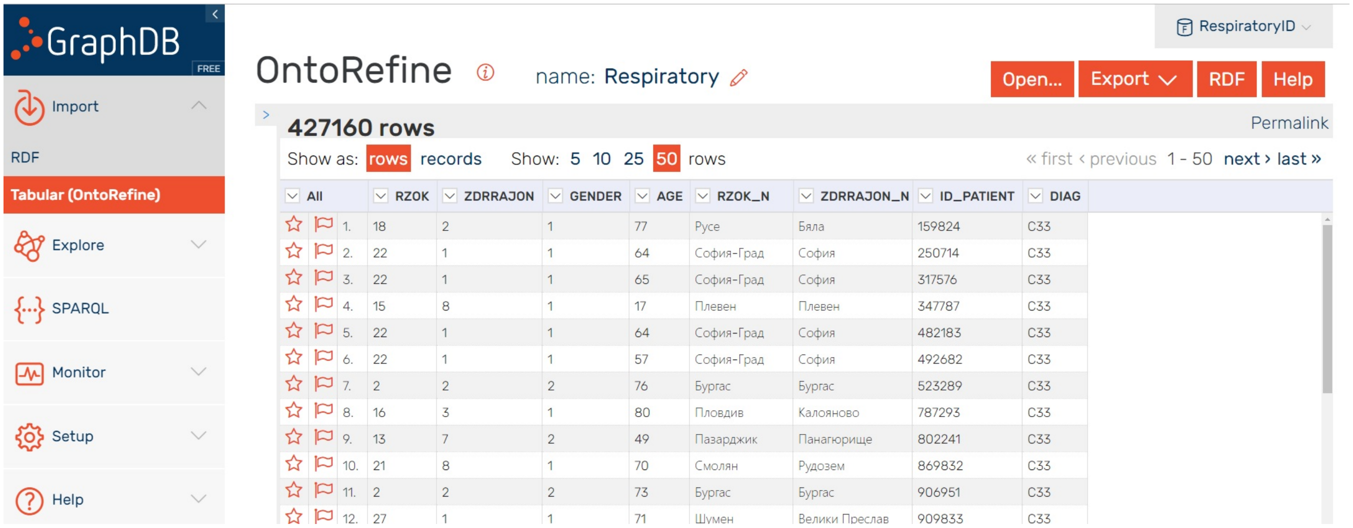
Figure 3: Geonames data about the Bulgarian city Pleven

Data Selection – in the initial setup diagnoses of interest are chosen for further deep analyses. This is necessary in order to speed up the risk factors search because ICD-10 encodes more than 14,000 diagnoses. After that the ORs of all patients, suffering from the specified diseases, are extracted from the Repository as structured records in comma separated format (CSV), which is the intermediate format for data exchange between different modules of the system.

For this study only some attributes of interest are extracted from the ORs (Fig. 4): RZOK (ID of the regional NHIF branch), ZDRRAJON (code for the location of the doctor's practice), GENDER (1–male, 2–female), AGE (age of the patient), RZOK_N (the name of the regional NHIF branch), ZDRRAJON_N (the name of the location - city, town or village of the doctor's practice), ID_PATIENT (unique patient ID generated by the pseudonymization algorithm), DIAG (a 3–sign ICD–10 code of the disease). Names of locations are given in Bulgarian language with Cyrillic alphabet. Nevertheless this is not a problem, because the Geonames LOD supports multilingual literals for names. For example, there are 39 alternate names in different languages for the city of Pleven in Bulgaria (Fig. 3).

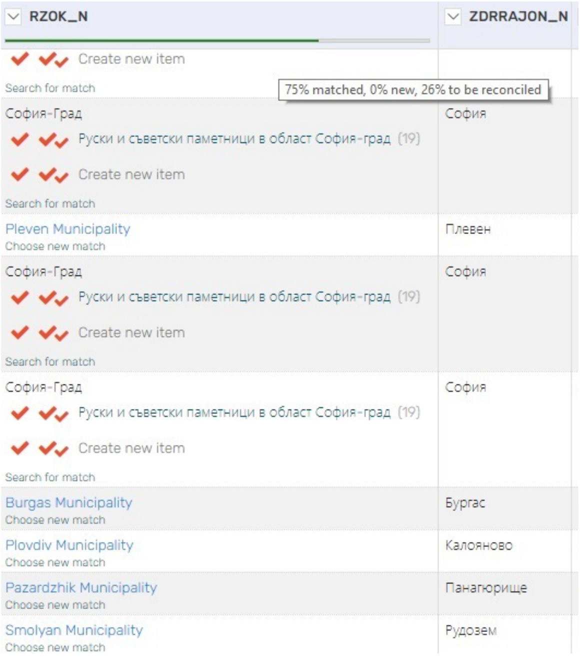
Preprocessing – the main goal of this step is to resolve the problems after mapping ontologies over the ORs. The process includes data cleaning and reconciliation. Fig. 5 illustrates automatic reconciliation of RZOK_N matched to the multilingual WikiData[9] about municipalities in Bulgaria. About 75% of the data items are matched successfully using the best match score. There are about 190 rows without data for RZOK_N and the remaining values that require manual resolution are only about Sofia city, Sofia Province, Kardzhali and Veliko Tarnovo, that were resolved adding four additional rules. The automatic match of ZDRAJON_N over WikiData cities resolves about 80% of the cases. The main problems are caused by multiple cities with the same name in different Bulgarian provinces, which requires some manual disambiguation (Fig. 6).

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



All	RZOK	ZDRRAJON	GENDER	AGE	RZOK_N	ZDRRAJON_N	ID_PATIENT	DIAG
1.	18	2	1	77	Русе	Бяла	159824	C33
2.	22	1	1	64	София-Град	София	250714	C33
3.	22	1	1	65	София-Град	София	317576	C33
4.	15	8	1	17	Плевен	Плевен	347787	C33
5.	22	1	1	64	София-Град	София	482183	C33
6.	22	1	1	57	София-Град	София	492682	C33
7.	2	2	2	76	Бургас	Бургас	523289	C33
8.	16	3	1	80	Пловдив	Калояново	787293	C33
9.	13	7	2	49	Пазарджик	Панагюрище	802241	C33
10.	21	8	1	70	Смолян	Рудозем	869832	C33
11.	2	2	2	73	Бургас	Бургас	906951	C33
12.	27	1	1	71	Шумен	Велики Преслав	909833	C33

Figure 4: Ontorefine - CSV file load

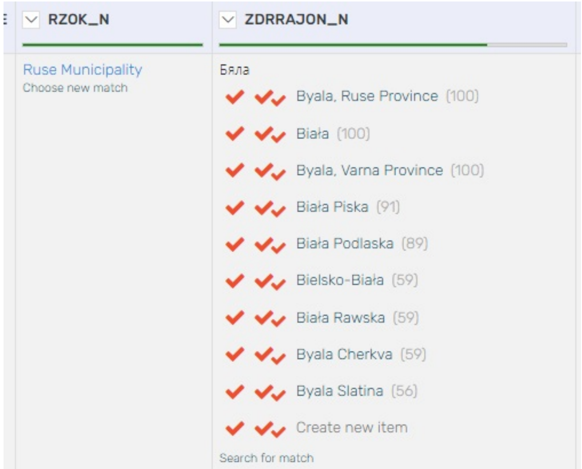


RZOK_N	ZDRRAJON_N
Create new item Search for match	
София-Град Руски и съветски паметници в област София-град (19) Create new item Search for match	София
Pleven Municipality Choose new match	Плевен
София-Град Руски и съветски паметници в област София-град (19) Create new item Search for match	София
София-Град Руски и съветски паметници в област София-град (19) Create new item Search for match	София
Burgas Municipality Choose new match	Бургас
Plovdiv Municipality Choose new match	Калояново
Pazardzhik Municipality Choose new match	Панагюрище
Smolyan Municipality Choose new match	Рудозем

Figure 5: Ontorefine - RZOK_N reconciliation

Another problem is due to the different granularity of the ZDRAJON_N values – some of them are cities, other are towns, villages or even smaller municipalities. The values of DIAG are matched to the ICD-10-CM classification in WikiData. Again this step is performed using the Ontorefine tool.

Knowledge Graph transformation – a free version of GraphDB is used for the KG creation and semantic processing, locally hosted and managed for our experiments. In this way sensitive patient ORs are kept on a local machine and extended by additional information



RZOK_N	ZDRAJON_N
Ruse Municipality Choose new match	Бяла Byala, Ruse Province (100) Biała (100) Byala, Varna Province (100) Biała Piska (91) Biała Podlaska (89) Bielsko-Biała (59) Biała Rawska (59) Byala Cherkva (59) Byala Slatina (56) Create new item Search for match

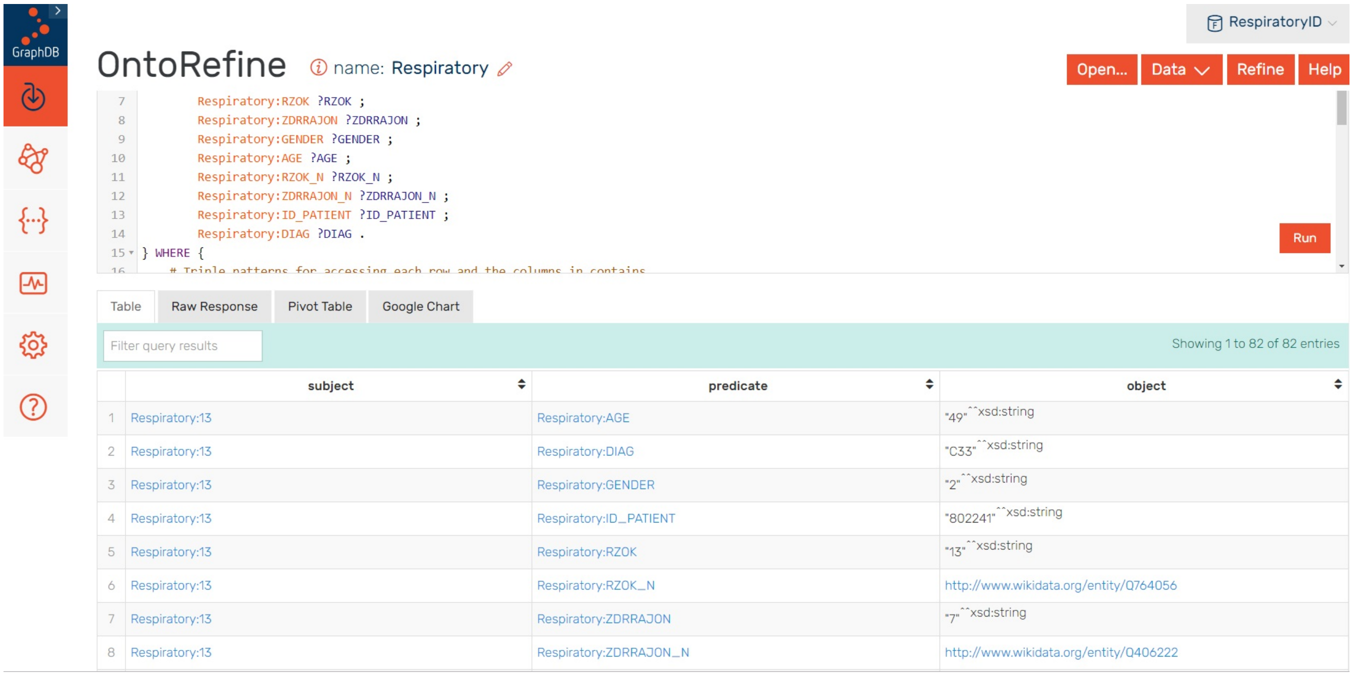
Figure 6: Ontorefine - ZDRAJON_N reconciliation

available in the cloud without sharing or exporting in-house data. Initially data is imported in CSV format and then is transformed into RDF (Fig. 7). RDF triples (subject-predicate-object) are created for regions (RZOK), locations (ZDRAJON) and diseases (DIAG), based on the W3C RDF standard[10].

Linking datasets – The RDF triples generated at the previous step are bounded with the Geonames and ICD–10–CM ontology and inserted into knowledge graph. This task is performed using SPARQL[11] query language for RDF.

Data Mining – at this step the search is expanded to all links directly or indirectly related to the KG nodes. For instance, not only the city names are searched in Geonames, but also nearby places – whether there are mountains, rivers, hot springs, hills,

[10] https://www.w3.org/RDF/
[11] https://www.w3.org/TR/rdf-sparql-query/



```
        Respiratory:RZOK ?RZOK ;
        Respiratory:ZDRRAJON ?ZDRRAJON ;
        Respiratory:GENDER ?GENDER ;
        Respiratory:AGE ?AGE ;
        Respiratory:RZOK_N ?RZOK_N ;
        Respiratory:ZDRRAJON_N ?ZDRRAJON_N ;
        Respiratory:ID_PATIENT ?ID_PATIENT ;
        Respiratory:DIAG ?DIAG .
} WHERE {
    # Triple patterns for accessing each row and the columns in contains
```

Showing 1 to 82 of 82 entries

	subject	predicate	object
1	Respiratory:13	Respiratory:AGE	"49"^^xsd:string
2	Respiratory:13	Respiratory:DIAG	"C33"^^xsd:string
3	Respiratory:13	Respiratory:GENDER	"2"^^xsd:string
4	Respiratory:13	Respiratory:ID_PATIENT	"802241"^^xsd:string
5	Respiratory:13	Respiratory:RZOK	"13"^^xsd:string
6	Respiratory:13	Respiratory:RZOK_N	http://www.wikidata.org/entity/Q764056
7	Respiratory:13	Respiratory:ZDRRAJON	"7"^^xsd:string
8	Respiratory:13	Respiratory:ZDRRAJON_N	http://www.wikidata.org/entity/Q406222

Figure 7: Ontorefine - RDF triples in format <subject,predicate,object> generated automatically by SPARQL query

etc. The most frequent shared links are used as a similarity measure between two nodes in the graph. The upper terminology for natural places is provided by DBPedia's classification. Ontologies and background knowledge added by Linked Open Data allow to enrich information in ORs and to search for some exogenous risk factors. *Frequent Patterns mining* – finally, generalization is applied for different frequent patterns of similarities between nodes in the KG. The main idea of ranking is based on the trajectories of diseases in the population graph proposed in [10]. A predefined threshold is selected for minimal frequency (minimal support - *minsup*). Initially the frequency is counted for triples linked to a patient in the database. Our system filters only those edges/links in the graph that are incident with at least a minsup number of pairs of the KG.

Interpretation – in this step we try to identify some hidden patterns in the data. Usually this task is performed by human experts only. Additional resources like ontologies and background knowledge from LOD allows semi-automatically to interpret identified patterns.

Risk Factors Set Delivery – final decision about risk factors is made. In addition a human expert is included in the analyses and interpretation of the automatically extracted frequent patterns and generated hypothesis for risk factors.

5 EXPERIMENTS AND RESULTS

Table 1 lists the diseases of interest selected for this study. The ICD-10 hierarchy of Chronic lower respiratory diseases (J40-J45) is shown on Fig. 1. Risk factors for respiratory diseases are closely related to the geolocations and specific climate. Nine experiments are run, one per each disease dataset separately.

For all experiments minsup=3 is used as a threshold value. This value was selected taking into account the prevalence of the selected set of diseases and the distribution of patients in different geolocations in our ORs repository. The frequent patterns found in the graph consist of locations and common relations between them with frequency above minsup (shown in the rightmost column of Table 1). Although all items in the rightmost column are frequent, the results for C33 are considered as insignificant for further investigation due to the lower number of patients in the training dataset with high dispersion in the country. For all other sets there are sufficient evidences which present the strong relation between locations.

Example 1: It was found that in six regions of Bulgaria - Montana, Vratza, Gabrovo, Kyustendil, Pleven and Silistra, there is a higher prevalence of the C34 diagnose than the average for the Bulgarian population. These six geographic areas have similar features: location in North Bulgaria, with humid subtropical climate and close to streams in the region. All these locations are also on relatively similar altitude.

Example 2: For J44 it was found that its prevalence is higher than the average in the following regions: Vratza, Vidin, Shumen, Haskovo, Sliven and Gabrovo. These locations share common links to relatively higher number of mountains and hills in the respective areas. Moreover, all the cities are in the foot of some mountain or higher hills, which prevents free circulation of fresh air and thus might imply higher pollution.

Table 1: Experimental datasets used in the study

ICD–10 code	Patients	RDF triples generated	Frequent Patterns
C33 Malignant neoplasm of trachea	53	278	3 locations / 4 relations
C34 Malignant neoplasm of bronchus and lung	9251	11203	6 locations / 15 relations
J31 Chronic rhinitis, nasopharyngitis and pharyngitis	47938	50951	2 locations / 4 relations
J40 Bronchitis, not specified as acute or chronic	38562	41450	2 locations / 5 relations
J41 Simple and mucopurulent chronic bronchitis	17504	20076	3 locations /5 relations
J42 Unspecified chronic bronchitis	23744	26529	2 locations /4 relations
J43 Emphysema	3030	4608	3 locations / 4 relations
J44 Other chronic obstructive pulmonary disease	167296	169007	6 locations /9 relations
J45 Asthma	119782	122305	6 locations /8 relations

6 CONCLUSION AND FUTURE WORK

Data linking is an investment in a cumulative store of knowledge [6]. As this task is time-consuming and uneasy, specific recommendations are provided for researchers, who plan to undertake a data linkage project in the health domain [5]. Obviously development of big linked data is difficult for languages other than English due to lack of critical mass of investments (via research or industrial project). But in this paper we show how available LOD resources in English can be employed for mining attributes in a multilingual context, and this is a promising scenario for automatic analysis of patient records in small languages. The results suggest that the method has a potential for identification of complex relations between diseases and geolocations thus allowing to augment the structured information from the ORs with external data that are not explicitly available in the clinical narratives.

In eHealth the human-in-the-loop solution is crucial for final decision making and fine tuning of the automatic process [8]. In the present experiment the final decision about the importance and feasibility of the extracted potential risk factors was made by human experts.

Our plans for future work include integration of other ontologies from the Bioportal like the human phenotype ontology [12]. Then we can mine more complex correlations using available temporal data in the ORs concerning the frequency of patient visit to doctors.

7 ACKNOWLEDGMENTS

The research presented in this paper 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

[1] Jans Aasman and Parsa Mirhaji. 2018. Knowledge Graph Solutions in Healthcare for Improved Clinical Outcomes. *Posters and Demonstrations, Industry and Blue Sky Ideas Tracks of the 17th Int. Semantic Web Conference, CEUR Workshop Proceedings* 2180 (2018).

[2] Marcia Barros and Francisco M Couto. 2016. Knowledge representation and management: a linked data perspective. *Yearbook of medical informatics* 25, 01 (2016), 178–183.

[3] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.

[4] Svetla Boytcheva, Galia Angelova, Zhivko Angelov, and Dimitar Tcharaktchiev. 2017. Integrating Data Analysis Tools for Better Treatment of Diabetic Patients. *CEUR Workshop Proceedings* 2022 (2017), 229–236.

[5] Stacie B Dusetzina, Seth Tyree, Anne-Marie Meyer, Adrian Meyer, Laura Green, and William R Carpenter. 2014. Linking data for health services research: a framework and instructional guide.

[6] Elizabeth Green, Felix Ritchie, Julie Mytton, Don J Webber, Toity Deave, Alex Montgomery, Lynn Woolfrey, Salim Chowdhury, et al. 2015. Enabling data linkage to maximise the value of public health research data.

[7] Jiawei Han, Jian Pai, and Yiwen Yim. 2000. Mining Frequent Patterns without Candidate Generation. *Proc. Conference on the Management of Data SIGMOD'00, Dallas, TX, ACM Press, New York, USA* (2000), 1–12.

[8] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[9] Nicolas Jay and Mathieu d'Aquin. 2013. Linked data and online classifications to organise mined patterns in patient data. In *AMIA Annual Symposium Proceedings*, Vol. 2013. American Medical Informatics Association, 681.

[10] Anders Boeck Jensen, Pope L Moseley, Tudor I Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. 2014. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications* 5 (2014), 4022.

[11] Milos Jovanovik, Aleksandra Bogojeska, Dimitar Trajanov, and Ljupco Kocarev. 2015. Inferring cuisine-drug interactions using the linked data approach. *Scientific reports* 5 (2015), 9346.

[12] Maulik R Kamdar and Mark A Musen. 2017. Mechanism-based Pharmacovigilance over the Life Sciences Linked Open Data Cloud. In *AMIA Annual Symposium Proceedings*, Vol. 2017. American Medical Informatics Association, 1014.

[13] Kouji Kozaki, Yuki Yamagata, Riichiro Mizoguchi, Takeshi Imai, and Kazuhiko Ohe. 2017. Disease Compass–a navigation system for disease knowledge based on ontology and linked data techniques. *Journal of biomedical semantics* 8, 1 (2017), 22.

[14] David J Odgers and Michel Dumontier. 2015. Mining electronic health records using linked data. *AMIA Summits on Translational Science Proceedings* 2015 (2015), 217.

[15] 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 (2013), 682.

[16] Petar Ristoski and Heiko Paulheim. 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web* 36 (2016), 1–22.

[17] Matthias Samwald, Anja Jentzsch, Christopher Bouton, Claus Stie Kallesøe, Egon Willighagen, Janos Hajagos, M Scott Marshall, Eric Prud'hommeaux, Oktie Hassanzadeh, Elgar Pichler, et al. 2011. Linked open drug data for pharmaceutical research and development. *Journal of cheminformatics* 3, 1 (2011), 19.

[18] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. 2014. Adoption of the linked data best practices in different topical domains. In *International Semantic Web Conference*. Springer, 245–260.

[19] 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 (2017).

[20] Amit Singhal. 2012. Introducing the Knowledge Graph: things, not strings. *Official Google Blog* 16 May (2012).

[12] http://bioportal.bioontology.org/ontologies/47321