PDD Graph: Bridging Electronic Medical Records and Biomedical Knowledge Graphs via Entity Linking

Meng Wang[1], Jiaheng Zhang[1], Jun Liu [1(✉)], Wei Hu[2], Sen Wang[3], Xue Li[4], and Wenqiang Liu[1]

1. MOEKLINNS lab, Xi'an Jiaotong University, Xi'an, China
2. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
3. Griffith Universtiy, Gold Coast Campus, Australia
4. The Universtiy of Queensland, Brisbane, Australia

wangmengsd@stu.xjtu.edu.cn,liukeen@xjtu.edu.cn

Abstract. Electronic medical records contain multi-format electronic medical data that consist of an abundance of medical knowledge. Facing with patients symptoms, experienced caregivers make right medical decisions based on their professional knowledge that accurately grasps relationships between symptoms, diagnosis and corresponding treatments. In this paper, we aim to capture these relationships by constructing a large and high-quality heterogenous graph linking patients, diseases, and drugs (PDD) in EMRs. Specifically, we propose a novel framework to extract important medical entities from MIMIC-III (Medical Information Mart for Intensive Care III) and automatically link them with the existing biomedical knowledge graphs, including ICD-9 ontology and DrugBank. The PDD graph presented in this paper is accessible on the Web via the SPARQL endpoint, and provides a pathway for medical discovery and applications, such as effective treatment recommendations.

Keywords: Linked Data · MIMIC-III · EMR · Drug · Disease

Resource type: Dataset
Permanent URL: http://kmap.xjtudlc.com/pdd

1 Introduction

Big data vendors collect and store large number of electronic medical records (EMRs) in hospital, with the goal of instantly accessing to comprehensive medical patient histories for caregivers at a lower cost. Public availability of EMRs collections has attracted much attention for different research purposes, including clinical research [14], mortality risk prediction [7], disease diagnosis [15], etc. An EMR database is normally a rich source of multi-format electronic data but remains limitations in scope and content. For example, MIMIC-III (Medical Information Mart for Intensive Care III) [8] collected bedside monitor trends,

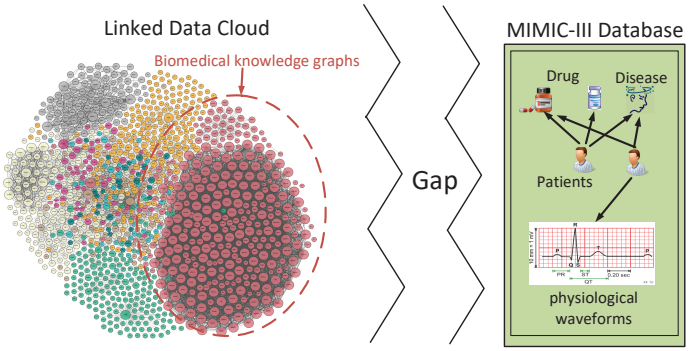

Fig. 1. Left part is the Linked Data Cloud[1], which contains interlinked biomedical knowledge graphs. Right part is the MIMIC-III database.

electronic medical notes, laboratory test results and waveforms from the ICUs (Intensive Care Units) of Beth Israel Deaconess Medical Center between 2001 and 2012. Abundant medical entities (symptoms, drugs and diseases) can be extracted from EMRs (clinical notes, prescriptions, and disease diagnoses). Most of the existing studies only focus on a specific entity, ignoring the relationship between entities. Given clinical data in MIMIC-III, discovering relationship between extracted entities (e.g. sepsis symptoms, pneumonia diagnosis, glucocorticoid drug and aspirin medicine) in wider scope can empower caregivers to make better decisions. Obviously, only focusing on EMR data is far from adequate to fully unveil entity relationships due to the limited scope of EMRs.

Meanwhile, many biomedical knowledge graphs (KGs) are published as Linked Data [1] on the Web using the Resource Description Framework (RDF) [4], such as DrugBank [9] and ICD-9 ontology [13]. Linked Data is about using the Web to set RDF links between entities in different KGs, thereby forming a large heterogeneous graph[1], where the nodes are entities (drugs, diseases, protein targets, side effects, pathways, etc.), and the edges (or links) represent various relations between entities such as drug-drug interactions. Unfortunately, such biomedical KGs only cover the basic medical facts, and contain little information about clinical outcomes. For instance, there is a relationship "adverse interaction" between glucocorticoid and aspirin in DrugBank, but no further information about how the adverse interaction affect the treatment of the patient who took both of the drugs in the same period. Clinical data can practically offer an opportunity to provide the missing relationship between KGs and clinical outcomes.

As mentioned above, biomedical KGs focus on the medical facts, whereas MIMIC-III only provides clinical data and physiological waveforms. There exists a gap between clinical data and biomedical KGs prohibiting further exploring medical entity relationship on ether side (see Fig.1). To solve this problem, we proposed a novel framework to construct a patient-drug-disease graph dataset (called PDD) in this paper. We summarize contributions of this paper as follows:

[1] Linking Open Data cloud diagram 2017. http://lod-cloud.net/

- To our best knowledge, we are the first to bridge EMRs and biomedical KGs together. The result is a big and high-quality PDD graph dataset, which provides a salient opportunity to uncover associations of biomedical interest in wider scope.
- We propose a novel framework to construct the PDD graph. The process starts by extracting medical entities from prescriptions, clinical notes and diagnoses respectively. RDF links are then set between the extracted medical entities and the corresponding entities in DrugBank and ICD-9 ontology.
- We publish the PDD graph as an open resource[2], and provide a SPARQL query endpoint using Apache Jena Fuseki[3]. Researchers can retrieve data distributed over biomedical KGs and MIMIC-III, ranging from drug-drug interactions, to the outcomes of drugs in clinical trials.

It is necessary to mention that MIMIC-III contains clinical information of patients. Although the protected health information was de-identified, researchers who seek to use more clinical data should complete an on-line training course and then apply for the permission to download the complete MIMIC-III dataset[4].

The rest of this paper is organized as follows. Section 2 describes the proposed framework and details. The statistics and evaluation is reported in Section 3. Section 4 describes related work and finally, Section 5 concludes the paper and identifies topics for further work.

2 PDD Construction

We first follow the RDF model [4] and introduce the PDD definition.

PDD Definition: PDD is an RDF graph consisting of PDD facts, where a PDD fact is represented by an RDF triple to indicate that a patient takes a drug or a patient is diagnosed with a disease. For instance,

$$\langle pdd{:}274671,\ pdd{:}\text{diagnosed},\ \text{sepsis}\rangle^{5}.$$

Fig.2 illustrates the general process of the PDD dataset generation, mainly includes two steps: PDD facts generation (described in Section 2.1), and linking PDD to biomedical KGs (described in Section 2.2).

2.1 PDD Facts Generation

According to the PDD definition, we need to extract three types of entities from MIMIC-III (patients, drugs, and diseases), and generate RDF triples of the prescription/diagnosis facts.

Patients IRI Creation: MIMIC-III contains 46,520 distinct patients, and each patient is attached with an unique ID. We add IRI prefix to each patient ID to form a patient entity in PDD.

Prescription Triple Generation: In MIMIC-III, the prescriptions table contains all the prescribed drugs for the treatments of patients. Each prescription

[2] http://kmap.xjtudlc.com/pdd
[3] https://jena.apache.org/documentation/fuseki2/index.html
[4] https://mimic.physionet.org/
[5] *pdd* is the IRI prefix http://kmap.xjtudlc.com/pdd_data/

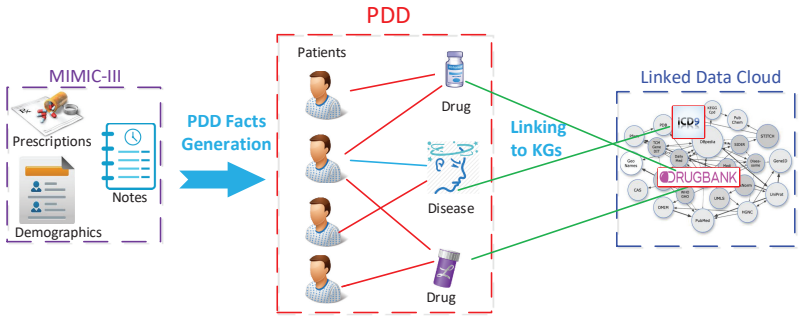

Fig. 2. Overview of PDD bridging MIMIC-III and biomedical knowledge graphs.

record contains the patient's unique ID, the drug's name, the duration, and the dosage. We extracted all distinct drug names as the drug entities in PDD. Then we added a prescription triple in to PDD. An example is

$$\langle pdd{:}18740,\ pdd{:}\text{prescribed},\ \text{aspirin}\rangle,$$

where *pdd*:18740 is a patient entity, and aspirin is the drug's name.

Diagnosis Triple Generation: MIMIC-III provides a diagnosed table that contains ICD-9 diagnosis codes for patients. There is an average of 13.9 ICD-9 codes per patient, but with a highly skewed distribution, as shown in Fig. 3. Beyond that, each patient has a set of clinical notes. These notes contain

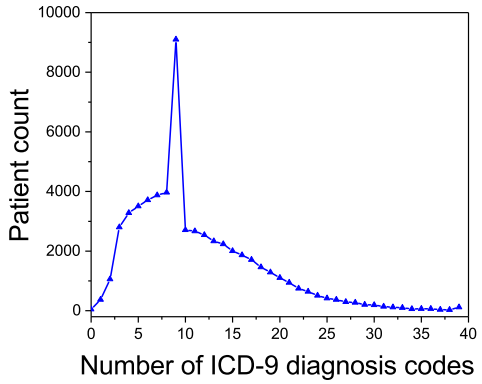

Fig. 3. The distribution of assigned ICD-9 codes per patient.

the diagnosis information. We use the named entity recognition (NER) tool C-TAKES [12] to extract diseases from clinical notes. C-TAKES is the most commonly used NER tool in the clinical domain. Then we use the model [15] (our previous work) to assign ICD-9 codes for extracted diseases. We extracted all ICD-9 diagnosis codes as the disease entities in PDD. Then we added a diagnosis triple into PDD. An example is

$$\langle pdd{:}18740,\ pdd{:}\text{diagnosed},\ \text{icd99592}\rangle,$$

where *pdd*:18740 is a patient entity, and icd99592 is the ICD-9 code of sepsis.

2.2 Linking PDD to Biomedical Knowledge Graphs

After extracting entities, we need to tackle the task of finding *sameAs* links [5] between the entities in PDD and other biomedical KGs. For drugs, we focused on linking drugs of PDD to the DrugBank of Bio2RDF [6] version, as the project Bio2RDF provides a gateway to other biomedical KGs. Following the analogous reason, we interlinked diseases of PDD with the ICD-9 ontology in Bio2RDF.

Drug Entity Linking: In MIMIC-III, drug names are various and often contain some insignificant words (10%, 200mg, glass bottle, etc.), which challenges the drug entity linking if the label matching method is directly used. In order to overcome this problem, we proposed an entity name model (ENM) based on [2] to link MIMIC-III drugs to DrugBank. The ENM is a statistical translation model which can capture the variations of an drug's name.

Fig. 4. The translation from *Glucose* to *Dextrose 5%*.

Given a drugs name m in MIMIC-III, the ENM model assumes that it is a translation of the drugs name d in DrugBank, and each word of the drug name could be translated through three ways:

1) Retained (translated into itself);
2) Omitted (translated into the word NULL);
3) Converted (translated into its alias).

Fig. 4 shows how the drug name *Glucose* in DrugBank translated into *Dextrose 5%* in MIMIC-III.

Based on the above three ways of translations, we define the probability of drug name d being translated to m as follows:

$$P(m|d) = \frac{\varepsilon}{(1_d + 1)^{l_m}} \prod_{j=1}^{l_m} \sum_{i=0}^{l_d} t(m_i|d_j) \tag{1}$$

where ε is a normalization factor, l_m is the length of m, l_d is the length of d, m_i is the i_{th} word of m, d_j is the j_{th} word of d, and $t(m_i|d_j)$ is the lexical translation probability which indicates the probability of a word d_j in DrugBank being written as m_i in MIMIC-III. DrugBank contains a large amount of drug aliases information, which can be used as training sets to compute the translation probability $t(m_i|d_j)$. After training the ENM from sample data, a drug name in MIMIC-III will be more likely to be translated to itself or aliases in DrugBank, whereas the insignificant words tend to be translated to NULL. Hence, our ENM can reduce the effects of insignificant words for drugs entity linking.

In addition, we propose two constraint rules when selecting candidate drugs for m, and discard those at odds with the rules.

Rule 1: One of the drug indications in DrugBank must be in accordance with one of the diagnoses of the patients who took the corresponding drug in MIMIC-III at least .

Rule 2: The dosage of a drug that patients took in MIMIC-III must be in accordance with one of the standard dosages listed in DrugBank.

Finally, we will choose the drug name d in DrugBank for the given drug m in MIMIC-III with maximal $P(m|d)$, and d satisfies the two constraint rules.

Disease IRI Resolution: In our previous work [15], we have assigned ICD-9 disease codes for extracted disease entities. Since the ICD-9 code is the international standard classification of diseases, and each code is unique. We can directly link the ICD-9 codes of PDD to ICD-9 ontology by string matching.

3 Statistics and Evaluation

In this section, we report the statistics of PDD and make the evaluation on its accuracy . At present PDD includes 58,030 entities and 2.3 million RDF triples.

Table 1. Statistics of Entities

	#Overall	#Drug/disease linked to KG
Patient	46,520	
Drug	4,525	3,449
Disease	6,985	6,983

Table 2. Statistics of RDF triples

	#Overall	#Drug/disease linked to KG
Demographics	165,526	
Patients-Drugs	1,517,702	1,259,702
Patients-Diseases	650,987	650,939

Table 1 shows the result of entities linked to the DrugBank and ICD-9 ontology. For drugs in PDD, 3,449 drugs are linked to 972 distinct drugs in DrugBank. For diseases in PDD, 6,983 diseases are connected to ICD-9 ontology. The only two failures of matching ICD-9 codes in MIMIC-III are '71970' and 'NULL', which are not included in ICD-9 ontology. Table 2 shows the result of RDF triples in PDD. In particular, 1,259,702 RDF triples contain drugs that have *sameAs* links to DrugBank, and 650,939 RDF triples have ICD-9 diseases codes. It indicates 83.4% drug-taken records in MIMIC-III can find corresponding entity in DrugBank, and 99.9% diagnosed information can link to ICD-9 ontology. A subgraph of PDD is illustrated in Fig. 5 to better understand the PDD graph.

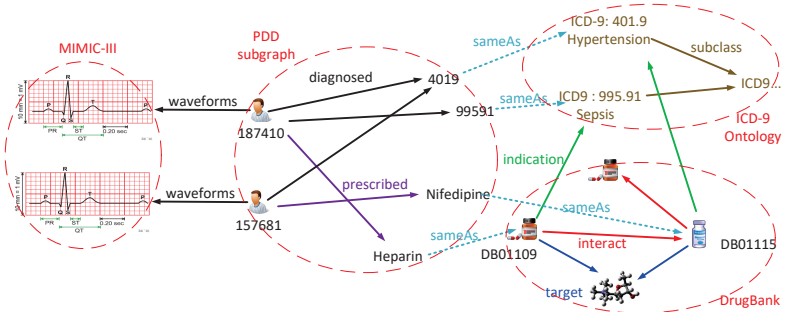

Fig. 5. An annotated subgraph of PDD.

To evaluate the ENM model, 500 samples are randomly selected, manually verified and adjusted. The ratio of positive samples to negative samples is 4:1, where positive means the entity can be linked to DrugBank. The precision is 94% and the recall is 85%. For linked entities in PDD we randomly chose 200 of them and manually evaluated the correctness of them, and the precision of entity links is 93% which is in an accordance with the result of our examples.

The overall accuracy of entity linking will be affected by the performance of the entity recognition tool. No entity recognition tools so far can achieve 100% accuracy. The average accuracy of C-TAKES (we used in this paper) is 94%. Therefore, the overall precision and recall may be lower.

In order to find out why those 1,076 drugs have not been linked to DrugBank yet, we extract 100 of them that hold the highest usage frequency. The observation shows that most of them are not just contained in DrugBank. For instance, DrugBank does not consider NS (normal saline) as a drug, but PDD contains several expressions of NS (NS, 1/2 NS, NS (Mini Bag Plus), NS (Glass Bottle), etc.). For drugs wrongly linked to DrugBank, the names of those drugs are too short, e.g. N i.e nitrogen. These short names provide little information and affect the performance of ENM directly. Also, the training data from DrugBank does not include the usage frequency of each drug name. That might lead to some inconsistence with applications in MIMIC-III and cause linking errors.

4 Related Work

In order to bring the advantages of Semantic Web to the life science community, a number of biomedical KGs have been constructed over the last years, such as Bio2RDF [6] and Chem2Bio2RDF [3]. These datasets make the interconnection and exploration of different biomedical data sources possible. However, there is little patients clinical information within these biomedical KGs. STRIDE2RDF [10] and MCLSS2RDF [11] apply Linked Data Principles to represent patients electronic health records, but the interlinks from clinical data to existing biomedical KGs are still very limited. Hence, none of the existing linked datasets are bridging the gap between clinical and biomedical data.

5 Conclusion and Future Work

This paper presents the process to construct a high-quality patient-drug-disease (PDD) graph linking entities in MIMIC-III to Linked Data Cloud, which satisfies the demand to provide information of clinical outcomes in biomedical KGs, when previous no relationship exists between the medical entities in MIMIC-III. With abundant clinical data of over forty thousand patients linked to open datasets, our work provides more convenient data access for further researches based on clinical outcomes, such as personalized medication and disease correlation analysis. The PDD dataset is currently accessible on the Web via the SPARQL endpoint. In future work, our plan is to improve the linking accuracy of ENM model by feeding more data into its training system.

Acknowledgment

This work is sponsored by The Fundamental Theory and Applications of Big Data with Knowledge Engineering under the National Key Research and Development Program of China with grant number 2016YFB1000903; National Science Foundation of China under Grant Nos.61672419, 61370019, 61532004, 61672420, and 61532015; MOE Research Center for Online Education Funds under Grant No.2016YB165; Ministry of Education Innovation Research Team No.IRT17R86.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Semantic services, interoperability and web applications: emerging concepts pp. 205–227 (2009)
2. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational linguistics 19(2), 263–311 (1993)
3. Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.J.: Chem2bio2rdf: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC bioinformatics 11(1), 255 (2010)
4. Consortium, W.W.W., et al.: Rdf 1.1 concepts and abstract syntax (2014)
5. Ding, L., Shinavier, J., Shangguan, Z., McGuinness, D.: Sameas networks and beyond: analyzing deployment status and implications of owl: sameas in linked data. The Semantic Web–ISWC 2010 pp. 145–160 (2010)
6. Dumontier, M., Callahan, A., Cruz-Toledo, J., Ansell, P., Emonet, V., Belleau, F., Droit, A.: Bio2rdf release 3: a larger connected network of linked data for the life sciences. In: Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272. pp. 401–404. CEUR-WS. org (2014)
7. Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., Szolovits, P.: Unfolding physiological state: Mortality modelling in intensive care units. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 75–84. ACM (2014)
8. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data 3 (2016)
9. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., et al.: Drugbank 4.0: shedding new light on drug metabolism. Nucleic acids research 42(D1), D1091–D1097 (2014)
10. Odgers, D.J., Dumontier, M.: Mining electronic health records using linked data. AMIA Summits on Translational Science Proceedings 2015, 217 (2015)
11. Pathak, J., Kiefer, R.C., Chute, C.G.: Applying linked data principles to represent patient's electronic health records at mayo clinic: a case report. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. pp. 455–464. ACM (2012)
12. 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)
13. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease ontology: a backbone for disease semantic integration. Nucleic acids research 40(D1), D940–D946 (2012)
14. Wang, S.J., Middleton, B., Prosser, L.A., Bardon, C.G., Spurr, C.D., Carchidi, P.J., Kittler, A.F., Goldszer, R.C., Fairchild, D.G., Sussman, A.J., et al.: A cost-benefit analysis of electronic medical records in primary care. The American journal of medicine 114(5), 397–403 (2003)
15. Wang, S., Chang, X., Li, X., Long, G., Yao, L., Sheng, Q.Z.: Diagnosis code assignment using sparsity-based disease correlation embedding. IEEE Transactions on Knowledge and Data Engineering 28(12), 3191–3202 (2016)