Journal of Biomedical Informatics 55 (2015) 206–217

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Toward a complete dataset of drug–drug interaction information from publicly available sources

Serkan Ayvaz [a,*], John Horn [b], Oktie Hassanzadeh [c], Qian Zhu [d], Johann Stan [e], Nicholas P. Tatonetti [f], Santiago Vilar [f], Mathias Brochhausen [g], Matthias Samwald [h], Majid Rastegar-Mojarad [i], Michel Dumontier [j], Richard D. Boyce [k]

[a] Department of Computer Science, Kent State University, 241 Math and Computer Science Building, Kent, OH 44242, USA
[b] Department of Pharmacy, School of Pharmacy and University of Washington Medicine, Pharmacy Services, University of Washington, H375V Health Sciences Bldg, Box 357630, Seattle, WA 98195, USA
[c] IBM T.J. Watson Research Center, 1101 Kitchawan Rd Route 134, P.O. Box 218, Yorktown Heights, NY 10598, USA
[d] Department of Information Systems, University of Maryland Baltimore County, Baltimore, MD 21250, USA
[e] Lister Hill National Center for Biomedical Communications, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA
[f] Departments of Biomedical Informatics, Systems Biology, and Medicine, Columbia University, 622 West 168th St VC5, New York, NY 10032, USA
[g] Division of Biomedical Informatics, University of Arkansas for Medical Sciences, 4301 W. Markham St, #782, Little Rock, AR 72205-7199, USA
[h] Section for Medical Expert and Knowledge-Based Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria
[i] Biomedical Statistics & Informatics, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA
[j] Stanford Center for Biomedical Informatics Research, Stanford, CA 94305, USA
[k] Department of Biomedical Informatics, Suite 419, 5607 Baum Blvd, Pittsburgh, PA 15206-3701, USA

ARTICLE INFO

Article history:
Received 22 October 2014
30 March 2015
Accepted 15 April 2015
Available online 24 April 2015

Keywords:
Drug–drug interaction
Record linkage
Natural language processing
Pharmacovigilance

ABSTRACT

Although potential drug–drug interactions (PDDIs) are a significant source of preventable drug-related harm, there is currently no single complete source of PDDI information. In the current study, all publically available sources of PDDI information that could be identified using a comprehensive and broad search were combined into a single dataset. The combined dataset merged fourteen different sources including 5 clinically-oriented information sources, 4 Natural Language Processing (NLP) Corpora, and 5 Bioinformatics/Pharmacovigilance information sources. As a comprehensive PDDI source, the merged dataset might benefit the pharmacovigilance text mining community by making it possible to compare the representativeness of NLP corpora for PDDI text extraction tasks, and specifying elements that can be useful for future PDDI extraction purposes.

An analysis of the overlap between and across the data sources showed that there was little overlap. Even comprehensive PDDI lists such as DrugBank, KEGG, and the NDF-RT had less than 50% overlap with each other. Moreover, all of the comprehensive lists had incomplete coverage of two data sources that focus on PDDIs of interest in most clinical settings. Based on this information, we think that systems that provide access to the comprehensive lists, such as APIs into RxNorm, should be careful to inform users that the lists may be incomplete with respect to PDDIs that drug experts suggest clinicians be aware of. In spite of the low degree of overlap, several dozen cases were identified where PDDI information provided in drug product labeling might be augmented by the merged dataset. Moreover, the combined dataset was also shown to improve the performance of an existing PDDI NLP pipeline and a recently published PDDI pharmacovigilance protocol. Future work will focus on improvement of the methods for mapping between PDDI information sources, identifying methods to improve the use of the merged dataset in PDDI NLP algorithms, integrating high-quality PDDI information from the merged dataset into Wikidata, and making the combined dataset accessible as Semantic Web Linked Data.

* Corresponding author at: 2432 Echo Valley Dr., Stow, OH 44224, USA. Tel.: +1 (330) 766 5310.
E-mail addresses: sayvaz1@kent.edu (S. Ayvaz), jrhorn@uw.edu (J. Horn), hassanzadeh@us.ibm.com (O. Hassanzadeh), qianzhu@umbc.edu (Q. Zhu), johann.stan.phd@gmail.com (J. Stan), nick.tatonetti@columbia.edu (N.P. Tatonetti), sav7003@dbmi.columbia.edu (S. Vilar), mbrochhausen@uams.edu (M. Brochhausen), matthias.samwald@meduniwien.ac.at (M. Samwald), Mojarad.Majid@mayo.edu (M. Rastegar-Mojarad), michel.dumontier@stanford.edu (M. Dumontier), rdb20@pitt.edu (R.D. Boyce).

http://dx.doi.org/10.1016/j.jbi.2015.04.006
1532-0464/

1. Introduction

Exposure to a potential drug–drug interaction (PDDI) occurs when a patient is prescribed or administered two or more drugs that can interact, even if no harm ensues [1]. "Known" interactions involve drug combinations for which (a) physiological data exists from clinical studies pointing to a potential interaction, (b) mechanistic assertions point toward a potential interaction, or (c) a potential interaction can be inferred based on reasonable extrapolation [2]. While exposure to a known PDDI does not always result in an adverse drug event [3], such events are a significant source of preventable drug-related harm. Sixteen cohort and case-control studies reported an elevated risk of hospitalization in patients who were exposed to PDDIs [4]. Clinically important events attributable to PDDI exposure are estimated to occur in 5.3–14.3% of inpatients, and are responsible for 0.02–0.17% of the nearly 130 million emergency department visits that occur each year in the United States [5,6].

At the time of this writing, there is no single complete source of PDDI information. While several proprietary and public PDDI information sources exist to help improve prescriber knowledge, they differ substantially in their coverage and agreement in the inclusion of PDDIs. One recent study found that only one quarter of 59 contraindicated drug pairs were listed in three proprietary PDDI information sources [7]. Another recent study comparing drug product labeling to the published literature for information on pharmacokinetic DDIs found that 40% of the 44 pharmacokinetic drug–drug interactions affecting 25 psychotropic drugs were located exclusively in product labeling [8]. These findings suggest that there is a pressing need for informatics research on how to best organize both existing and emerging PDDI information for search and retrieval.

Several groups would benefit from a more effective synthesis of existing available PDDI knowledge. For those individuals researching text mining of the pharmacovigilance literature, one possible benefit would be to enable a better understanding of the representativeness of a given natural language processing (NLP) *corpus* relative to all known PDDIs. A merged PDDI dataset might help improve existing text mining algorithms by providing computable domain knowledge. Text mining researchers might also find the PDDI synthesis useful for identifying gaps in PDDI information sources that text mining might be able to address. The development of a common PDDI framework could also benefit United States healthcare organizations who are currently striving to incorporate PDDI screening along with other strategies to achieve meaningful use of electronic medical records [9,10]; drug-safety scientists who monitor post-market data related to drug use for new concerns [11]; researchers in drug development who build *in silico* models to help identify new drug candidates or drugs that can be 'repositioned' for new uses [12]; those who create and maintain drug information resources that help clinicians guide patients to safe and effective medication therapies [1]; and patients seeking information on the safety of the medicines they take [13].

The objective of the project described here was to assess the feasibility and potential value to different stakeholders of interlinking all publicly available PDDI data sources using a common data model. We first conducted a comprehensive and broad search of public PDDI knowledge sources. We then established links between the PDDI sources and evaluated their information coverage. This resulted in single integrated PDDI dataset, and list of the specific data elements provided by each source. Finally, we conducted some preliminary analyses of the potential value of the merged dataset. These included (1) examining the overlap between the data sources including existing NLP corpora relative to other PDDI datasets, (2) testing if the PDDI dataset could improve the performance of a PDDI NLP algorithm, (3) examining cases where PDDI information provided in drug product labeling might be augmented by the merged dataset, and (4) testing if the combined dataset would improve the performance of a recently published pharmacovigilance protocol [14].

2. Materials and methods

2.1. Survey of DDI data sources

The scope of the PDDI source search included drug interaction lists designed for use in clinically oriented applications, annotated text corpora used for NLP research, knowledge bases used for clinical and translational research, and suspected PDDI associations (i.e., pharmacovigilance signals) [15]. We searched for all potentially relevant resources by querying bibliographic databases (PubMed and Google Scholar), reviewing the tertiary literature, and scanning conference proceedings for papers describing drug-related resources. This search was augmented by requests for input from members of various pharmacoinformatics and chemoinformatics interest groups and maintainers of major meta-repositories for RDF data such as Bio2RDF [16]. We then manually inspected each potentially relevant resource to determine if it (1) supported NLP experiments, (2) provided information for use by clinicians, or (3) supported bioinformatics or pharmacovigilance research. These three categories were chosen because we think that they cover the three primary use cases for PDDI knowledge. We considered the resources that are non-proprietary and represented as structured data or require minimal efforts to structure. Fig. 1 demonstrates the resources within each category and an overview of the study framework.

2.2. Data element survey

We acquired all publicly available PDDI datasets identified by the aforementioned search and then designed a simple PDDI data model (i.e., an associative array or "dictionary") to combine the data elements provided from each source. We then developed custom scripts to translate the PDDIs listed in each source to the model. This activity and all analyses described below were conducted between June and September 2014 using the versions of the data sets current at that time.

2.3. Analysis of the overlap between the data sources

With the goal of integrating publicly available PDDI datasets, we first performed an analysis of the overlap between drug entities found across the sources. The first step in this analysis involved identifying attributes across the sources that could be used to match records that refer to the same drug entity (i.e., linkage points). Because our goal was to facilitate drug mapping across different drug resources while avoiding erroneous mappings, we restricted linkage points to:

- Existing mappings where one source provided an unambiguous drug identifier from another source (e.g., Source A provides the exact unique identifier for drug X in Source B).
- An exact case insensitive match of the string name or synonym for the drug as provided in two sources.
- An intermediate source provided a data item (e.g., a chemical structure string) that could be used to create an unambiguous mapping between a drug entity to other sources.