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.

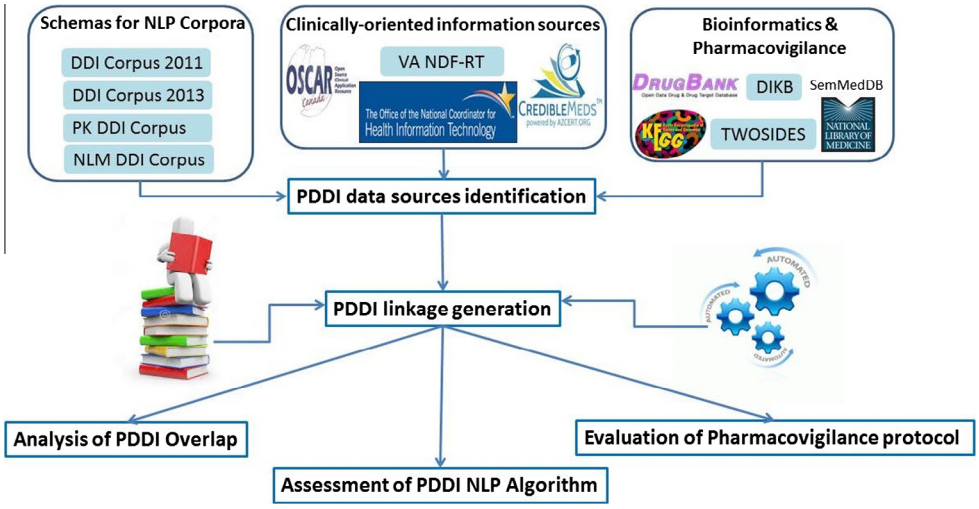

Fig. 1. Potential drug–drug interaction information resources included in the study and an overview of the study framework.

Drug entities in each dataset were mapped to DrugBank identifiers [17] wherever possible to enable cross-dataset comparisons. DrugBank was chosen for this purpose because of its broad inclusion of drugs, including drugs marketed in different countries. The resource also provides a variety of identifiers for drugs including string names, codes from various terminologies, and chemical structure identifiers.

Once the drug entities used by each source was mapped to DrugBank we use simple counts and percentages to compare PDDI overlap. We created a tabular representation of the results so that the overlap of the datasets, including those supporting NLP experiments, could be easily compared with each other. We also queried the merged dataset to identify specific PDDI instances that were in common across sources.

2.4. Testing if the PDDI dataset could improve the performance of a PDDI NLP algorithm

We tested if a merged PDDI dataset could improve the performance of the existing PDDI NLP pipeline created by Bui et al [18]. The Bui pipeline was chosen because (1) its performance is well characterized for three different NLP corpora, and (2) the code implementing the pipeline is available as an open source project. The system uses three steps to classify sentences for the mention of a PDDI. First, it pre-processes the input text to convert input sentences into a structured representation. During this phase, the sentence text is scanned to identify "trigger" words from a list of more than 200 words similar to "alteration", "blocked", "caused", and "potentiated." Sentences for which no trigger words are found are dropped from further processing. Sentences that mention trigger words are mapped into a suitable syntactic structure and then used to generate feature vectors. The third step in the process is to use the obtained feature vectors to train a support vector machine (SVM) classifier.

We added one additional step during pre-processing for those sentences that lacked a trigger word. Rather than simply exclude the sentence, a version of the merged PDDI dataset was queried for each drug pair mentioned in the sentence. If the query returned a result, the associated sentence was passed to the remaining steps of the NLP pipeline. Our reasoning for this approach is that the list of trigger words compiled by the Bui et al. might be an imperfect filter for PDDI NLP because it arose from their research on NLP of protein–protein interactions. We expected that there might be cases where a PDDI mention uses terms not present in the trigger list. In those cases, the presence in the merged PDDI dataset of a drug pair from the sentence might be sufficient "domain knowledge" to justify retaining the sentence for further NLP.

We altered the code written by Bui et al. [18] to use the test method when flagged. We then tested the Bui pipeline using the NLP corpora that they used in their original evaluation, both with and without the use of the merged PDDI dataset during pre-processing. The merged PDDI dataset used for testing the approach was slightly different from the one used to compare PDDI overlap (Section 2.3). Specifically, datasets that included the exact same set of PDDIs as the two NLP corpora were excluded. Also excluded were PDDI datasets generated using NLP or by pharmacovigilance signal generation. These were excluded because the number of false positives in these datasets is not known. All tests were run on a 64-bit Dell XPS configured with Ubuntu Linux 14.04 and using Java 1.7.0_72. The source code for the modified NLP pipeline was made available for download to interested researchers.[1]

2.5. Augmenting PDDI information in Drug Product Labeling

The drug product label, also known as a "package insert" in the United States (US) and the summary of product characteristics in Europe, is a document required by law, written for clinicians and patients containing information on the drug by the drug manufacturers. In this respect, it differs from an *NLP corpus*, which is a large collection of texts, written or spoken material upon which a linguistic analysis is based and used in the development of NLP tools, and a *knowledge base*, which is a technology used to store complex structured and unstructured information used by a computer system. US Drug product labeling information is provided to the public for free by the National Library of Medicine DailyMed web site.

Once all datasets were interlinked by DrugBank identifiers, we investigated the potential value of the interlinked dataset for augmenting PDDI information provided in drug product labeling. Studies have shown product labeling to be incomplete [19, 20] and one motivation for this activity is to extend previous pilot work on methods to address known limitations of the information source [21]. To do this, we examined cases where there existed an overlap between the PDDI datasets designed for NLP research with product labeling and other data sources.

[1] https://github.com/dbmi-pitt/pk-ddi-role-identifier/tree/master/nlp-ddi-role-identifier.

2.6. Applying the merged PDDI sources to predict new potential interactions

We tested the potential value of the interlinked PDDI dataset for improving the performance of a pharmacovigilance protocol for making PDDI predictions [14]. The protocol generates new PDDI candidates through the integration of chemical, structure, and drug interaction similarity measures via linear algebra techniques. The protocol specifies that drug interaction similarity measures should be integrated into a reference standard of well-established PDDIs. To apply the protocol, the drugs described in the different sources were mapped to 928 DrugBank drugs used in previous studies for which the chemical and structure similarity information had already been calculated.

A key component of the protocol is a PDDI reference standard that is representative of the many different mechanisms by which drug–drug interactions might occur. We experimented with four different methods for creating such a reference standard: (1) combining all clinically oriented applications (except one left out for validation, see below), (2) combining all text corpora developed for NLP research, (3) combining all knowledge bases used for bioinformatics/pharmacovigilance, and (4) combining all datasets into a single dataset.

The performance of the protocol using each dataset as the reference PDDI knowledge base was evaluated by comparing the list of predicted PDDIs with a list of PDDIs from the VA NDF- RT [22] – a dataset that, until September of 2014, covered all drugs used in United States Veteran's Administration (VA) formulary. Plots of area under the receiver operating characteristic (ROC) curves were used for performance comparisons. We selected the VA NDF-RT as the comparison information source because it was designed to be a comprehensive and clinically-oriented PDDI information source. Some drug–drug interactions data sources, such as DrugBank, contain sets of theoretical DDIs with limited implications at clinical level.

3. Results

3.1. Survey of DDI Data Sources

Our systematic search identified 14 publicly available sources of PDDIs (Table 1).

- Five of the sources were developed for clinical application: CredibleMeds [23] – a list of PDDIs thought to be clinically relevant and supported by strong scientific evidence; VA NDF- RT [22] – a list of PDDIs formerly maintained by the United States Veteran's Administration (VA) for use in VA care settings; ONC High Priority [24] – a list of PDDIs suggested as a high priority to alert clinicians in any care environment; ONC Non-interruptive [25] – a list of PDDIs not requiring interruptive alerting in any care environment; and OSCAR – a list of PDDIs derived by expert consensus in the late 1990s [26] that were more recently used in an Open Source Electronic Health Records system called OSCAR [27].
- Four of the sources were developed to support NLP algorithm development: DDI Corpus 2011 [28] – the reference standard for the 2011 DDIExtraction Challenge on drug–drug interaction NLP; DDI Corpus 2013 [29] the reference standard for the 2013 SemEval DDIExtraction Challenge that followed the 2011 challenge; PK DDI Corpus [30] – a reference standard used to develop NLP to identify pharmacokinetic PDDIs mentioned in product labeling; and NLM CV DDI Corpus [31] – a reference standard used to develop NLP to identify PDDIs mentioned in drug product labeling affecting cardiovascular drugs.
- Five other sources were developed to support either pharmacovigilance or bioinformatics applications. KEGG DDI [32] – a list provided by the Kyoto Encyclopedia of Genes and Genomes (KEGG) resource [33] of PDDIs extracted from the interaction tables of Japanese product labels; TWOSIDES [15] – a list of PDDI pharmacovigilance signals derived by data mining a database of spontaneously reported adverse events; DrugBank – PDDIs listed in v4.0 of DrugBank [17]; DIKB – observed and predicted pharmacokinetic interactions listed in the Drug Interaction Knowledge Base [34]; and SemMedDB – a database of subject, predicate, object relationships extracted from MEDLINE abstracts by the NLP program SemRep [35]. For this project, all "INTERACTS_WITH" relationships between two drugs were treated as PDDIs.

Four non-NLP sources were comprehensive in their identification of PDDIs across all drugs (VA NDF-RT, DrugBank, TWOSIDES, KEGG DDI and SemMedDB). Of these, only the VA NDF-RT was designed for use in a clinical setting. TWOSIDES was the only pharmacovigilance signal source. Both KEGG DDI and SemMedDB were generated by automated text extraction methods.

One source developed for NLP (DDI Corpus 2011) was derived primarily from data available in DrugBank at the time it was created while another source (DDI Corpus 2013) included DrugBank PDDIs and PDDIs identified in a sample of a few hundred abstracts present in MEDLINE. Another NLP corpus (PK DDI Corpus) was intended to be a representative sample of statements in product labels describing pharmacokinetic PDDIs. The fourth NLP corpus (NLM CV DDI Corpus) focused on all types of PDDIs involving cardiovascular drugs.

The remaining sources focused on PDDIs of a certain severity (CredibleMeds), clinical importance (ONC High Priority and ONC Non-interruptive), mechanism of interaction (DIKB), or frequency of prescription (OSCAR). To elaborate the meaning of severity concept, it can be described as a label qualifying the relative importance of a PDDI. For example, the VA NDF-RT system assigned a severity level of "Critical" for PDDIs that were thought by the system's developers to be of generally greater concern than those labeled as "Significant".

3.2. Data element surveys

Table 2 shows a combined view of the data elements provided from each PDDI source. The data elements provided varied considerably depending on the source. CredibleMeds was the most comprehensive in terms of variety of data elements and was the only non-NLP source to provide management options. Two of the NLP sources provided a tag that could be used to determine if a PDDI mention described management options. In DDI Corpus 2013 the tag was "advise" while the tag was "Caution Interaction" in the NLM DDI Corpus.

All resources that provided a description of the PDDI (including NLP sources) did so as unstructured text. Four non-NLP sources (CredibleMeds, KEGG, DrugBank, and DIKB) provided some mention of PDDI mechanism (e.g., enzyme inhibition) but only two sources (KEGG and DIKB) provided this information as a term from an ontology (thus, making it interpretable by a computer). Two of the NLP sources provided tags that could be used to determine if a PDDI mention explained the mechanism of the interaction. In DDI Corpus 2013 the tag was "mechanism", while the tags "Increase_Interaction" and "Decrease_Interaction" were used in the NLM DDI Corpus. No such tag was provided by the PK DDI Corpus but all PDDIs provided by this source occurred by the mechanism of enzyme inhibition.

The clinical effect fields were computable in only one of the four sources that provided such data (TWOSIDES). DDI Corpus 2013

Table 1
List of PDDI Data Sources – A short description of publicly available PDDI sources including the number of PDDIs in the original data extraction (denominator) and the number of PDDIs that we could map to DrugBank (numerator). Counts for mapped PDDIs are on unique interacting pairs without consideration of other data elements and with no distinction of the precipitant or object of the interaction.

Source	Description	Mapped/original	Category	Data owner/maintainer	Frequency of updates
Crediblemeds.org	A list of clinically important drug–drug interactions	82/83	Clinically-oriented	Crediblemeds.org	As needed
VA NDF-RT	PDDIs used until 2014 by the Veteran's Administration health care system	2606/5265	Clinically-oriented	Veterans Health Administration	No future updates. Discontinued
ONC High Priority	A consensus list of PDDIs that are recommended by the Office of the National Coordinator as high priority for inclusion in alerting systems	1150/1150	Clinically-oriented	ONC	One-time
ONC Non-interruptive	A consensus list of PDDIs that are recommended by the Office of the National Coordinator for use in non-interruptive alerts	2101/2101	Clinically-oriented	ONC	One-time
OSCAR	PDDIs used on an open source electronic health records system	7969/7969	Clinically-oriented	Oscar McMaster EMR	One-time
DDI Corpus 2011	Training corpus for the 2011 SemEval PDDI text extraction challenge	586/3160	NLP Corpora	Isabel Segura-Bedmar	One-time
DDI Corpus 2013	Training corpus for the 2013 SemEval PDDI text extraction challenge	1287/5021	NLP Corpora	Isabel Segura-Bedmar	One-time
PK DDI Corpus	Training corpus for NLP to extract pharmacokinetic PDDIs from all drug product labels	166/298	NLP Corpora	Richard D. Boyce	One-time
NLM CV DDI Corpus	Training corpus for NLP to extract all PDDIs from cardiovascular drug product labels.	247/2963	NLP Corpora	National Library of Medicine	One-time
KEGG DDI	PDDIs extracted from the interaction tables of Japanese product labels	26,664/298,337	Bioinformatics–Pharmacovigilance	Kanehisa Laboratories	As needed
TWOSIDES	Pharmacovigilance signals indicative of possible associations between drug combinations and adverse events	9921/63,473	Bioinformatics–Pharmacovigilance	Nicholas Tatonetti	One-time
DrugBank	Comprehensive drug information resource	12,113	Bioinformatics–Pharmacovigilance	DrugBank .ca	Roughly bi-annual
DIKB	An evidence-focused knowledge base of pharmacokinetic PDDIs	561/561	Bioinformatics–Pharmacovigilance	Richard D. Boyce	Periodic
SemMedDB	PDDIs extracted by NLP from the titles and abstracts in PubMed	3952/190,219	Bioinformatics–Pharmacovigilance	National Library of Medicine	As needed

Table 2
Data elements provided by publicly available PDDI sources.

Data element	Credible Meds	NDF-RT	ONC High Priority	ONC Non-interruptive	OSCAR	DDI Corpus 2011	DDI Corpus 2013	PK DDI Corpus	NLM DDI Corpus	KEGG	TWO-SIDES	Drug-Bank	DIKB	Sem MedDB
Confidence value											x			
Description	x					x	x	x	x			x	x	
Clinical effect	x				x		x				x[a]			
Citation of evidence					x								x[a]	x
Management options	x						x		x					
Mechanism	x						x		x	x[a]		x[d]	x[a]	
Modality								x						
Precipitant/object distinction[e]	x[a]		x	x	x		x	x	x				x[a]	
Related drugs	x													
severity concept		x[a,b]	x	x	x					x[c]				

[a] Data element is computable rather than in unstructured text.
[b] Critical or severe.
[c] Precaution or contraindicated.
[d] Available on the public website but not explicitly in the downloadable data.
[e] The individual drugs involved in PDDI were tagged as having either the precipitant or object role.

used the tag "effect" to indicate that a PDDI mention described a clinical effect. However, records from the source provided no other semantic relationships such as a concept from a biomedical terminology indicating the clinical effect.

Considering data elements that were unique to a single source, TWOSIDES was the only source to provide a *confidence value* for its PDDI-adverse event associations. A confidence value is a real-valued number representing the level of certainty that a pharmacovigilance drug safety signal is real. We note that it is also the only pharmacovigilance signal source present in this analysis. PK DDI Corpus was the only source to include drug pairs that are known to *not* interact by using the "modality" tag to distinguish positive from negative PDDIs. Finally, CredibleMeds was the only source to list other drugs associated with the anticipated effect of the PDDI. This information is important in cases where a clinician must assess if an ADE occurring in a patient is associated with a PDDI [36].

3.3. Analysis of the overlap between the data sources

We analyzed overlap between the subset of the PDDIs in the public data sources shown in Table 1 for which both drugs involved in an interaction could be mapped to DrugBank. Here we provide a brief summary of the mapping procedure and then summarize the results.

For the ONC High Priority PDDIs and ONC Non-Interruptive PDDIs, we manually extracted the PDDIs from the publications [24], [25]. The extracted list often mentioned drug class interactions rather than individual drugs. Our pharmacy experts helped identify the drugs belonging to each class. Another recent study [37] has also identified the same individual drugs within each drug class of ONC High Priority PDDIs, short of QT Prolonging agents. For QT Prolonging agents in ONC High Priority list, we obtained the list of agents with a known risk of torsades de pointes (TdP) provided by Crediblemeds.org with the directions of our pharmacy experts. We then mapped drug names from the two sources to DrugBank entities using an exact string match on DrugBank names and synonyms. PDDIs from CredibleMeds were manually extracted and mapped to DrugBank entries in the same way.

DrugBank provides KEGG drug identifiers for many drugs that we used as a mapping. Similarly, TWOSIDES uses PubChem drug identifiers that are included within many DrugBank drug records. Drugs in the DIKB were already identified by DrugBank identifiers so no mapping was required.

The remaining mappings were done by identifying intermediate mappings that could serve as a bridge between the source and DrugBank. In particular, we use the Unique Ingredient Identifiers (UNII) list provided by the FDA that contains a list of preferred substance names, synonyms, chemical structure strings, and UNII codes for drugs. The first step was to match active ingredient strings from RxNorm [38] with the UNII preferred terms using a case insensitive string match because this was formerly reported as a viable method in RxNorm documentation [39]. With this robust mapping, we then sought to perform an automated mapping of FDA Unique Ingredient Identifiers to DrugBank based on our previous work interlinking these sources [40]. The final FDA UNII to DrugBank mappings were manually reviewed for accuracy.

We developed two different approaches that were based on fact that a large number of drugs in both the UNII list and DrugBank contain IUPAC International Chemical Identifiers (InChI) [41]. One approach was more conservative and limited matches between a DrugBank and UNII record to those cases where an exact case-insensitive match was identified for *both* an InChI identifier and either the drug preferred term or synonym. A less conservative approach involves a match on InChI key *or* an exact case-insensitive match of preferred term or a synonym. The latter approach resulted in a greater number of mappings (1613 and 2139 mappings respectively). We report the overlap analysis based on the latter mapping approach.

The OSCAR list of PDDIs used ATC codes for drug identifiers. We utilized a mapping from ATC codes to RxNorm drug identifiers available in the OMOP/IMEDS Standard Vocabulary (version 4) [42]. We then used our mapping of RxNorm to DrugBank via the UNII list. In cases where the ATC codes provided by Oscar were more general than just a single drug, our drug experts helped identify the drug specific identifiers to link back to DrugBank. Similarly, the NDF-RT interactions were mapped by combining a mapping of the NDF-RT to RxNorm present in the Unified Medical Language System(UMLS) [43], and then mapping from RxNorm to DrugBank via the UNII list. PK DDI Corpus mappings were also done using the same method described above; identifying intermediate mappings and linking drugs with RxNorm identifiers to DrugBank.

In the case of SemMedDB, we extracted all "INTERACTS_WITH" predictions from SemMedDB, which provides UMLS Concept Unique Identifiers (CUI) for both concepts in the interaction prediction. As UMLS CUIs are not limited to drugs, we only exploited the CUIs where both interacting concepts could be mapped to drugs with RxNorm CUIs. After mapping to RxNorm, we followed the same aforementioned methodology for linking drugs with RxNorm identifiers to DrugBank.

The "mapped/original" columns of Table 1 show both the original number of PDDIs for each data source and the number of PDDIs that remained after mapping to DrugBank. While most mappings are complete, there are differences between mapped/original counts in some resources (e.g., VA NDF-RT, KEGG, and TWOSIDES). The main reason behind the discrepancy between mapped and unmapped PDDIs is the PDDI duplication in the original sources. Under "mapped" PDDIs, we only consider the unique pair of drugs in the PDDI regardless of their position as object or precipitant. For instance, PDDI between Acetylcholine and Morphine was reported 20 times in SemMedDB by different studies. However, we consider it as a single mapped PDDI. Similarly, the PDDI pair of Carnitine and Bupropion was reported in association with 47 different adverse events in TWOSIDES. In KEGG, there were duplications due to directionality of the PDDI drug pairs. Also, the KEGG PDDI mapping relies on the cross reference links between DrugBank ID to KEGG ID provided in DrugBank. Since we currently have no other direct way of mapping KEGG drugs to DrugBank, we could only map the KEGG PDDIs, where both drugs involved in PDDIs with DrugBank cross reference links. In the future, we might explore alternative methods for improving mapping between KEGG and DrugBank such as text parsing and string matching techniques.

The DDI Corpora, PK DDI Corpus, NLM Corpus are designed for NLP text mining purposes. Hence, they only have drug names as drug identifier that can be utilized for mapping process. For mapping from these data sources, we used exact string matching on the drug names, synonyms and brand names (see above). However, some of the PDDIs from the NLP resources were not mappable as they contained PDDIs that were specified as metabolites, groups of chemicals, or drug classes rather than individual drugs. The number of drug groups or classes that were listed in one or both of the drugs in the PDDIs was 288 in the DDI 2011 Corpus, 445 in DDI 2013 Corpus and 493 in the NLM PDDI Corpus. Out of the drugs groupings in these resources, there were many instances of atypical, ambiguous and error-prone drug group mentions such as "drugs that are commonly taken by the elderly", "drugs that have a narrow therapeutic range", "centrally acting drugs", "drugs that are actively secreted by the kidney has not been investigated in humans". Therefore, we limited the mapping process to the individual drugs from the aforementioned resources.

Table 3 shows the pairwise overlap between the fourteen sources. The largest overlap in terms of PDDI count was between DrugBank and KEGG (2143 PDDIs). In terms of average percentage overlap, DrugBank and KEGG covered the most drug pairs across other sources (28.6% and 25.6% respectively). The greatest percentage of PDDIs covered by another source was DDI Corpus 2013's coverage of DDI Corpus 2011 (535 PDDIs, 91.3%) followed by the DrugBank's coverage of the CredibleMeds (57 PDDIs, 69.1%). Three data sources had no overlap with CredibleMeds (PK DDI Corpus, TWOSIDES, and SemMedDB) and ONC-Non-interruptive had no overlap with DIKB.

Each of the NLP datasets had very little overlap with the clinically-oriented datasets. The greatest overlap was between DDI Corpus 2013 and the VA NDF-RT with 295 PDDIs in common (11.4% of the NDF-RT and 22.9% of the DDI Corpus 2013). The overlap between the NLP datasets and the two clinical datasets that focused on PDDIs of high clinical importance was much less.

Table 3
Overlap between publicly available PDDI datasets for those drug pairs that could be mapped to DrugBank in both dataset.

	SemMedDB												
Credible Meds	0 (0.0%, 0.0%)	Credible Meds											
NDF-RT	69 (2.7%, 1.7%)	16 (0.6%, 19.5%)	NDF-RT										
ONC High Priority	12 (1.0%, 0.3%)	8 (0.7%, 9.8%)	225 (19.6%, 8.7%)	ONC High Priority									
ONC Non-interruptive	8 (0.4%, 0.2%)	4 (0.2%, 4.9%)	27 (1.3%, 1.0%)	2 (0.1%, 0.2%)	ONC Non-interruptive								
OSCAR	124 (1.6%, 3.1%)	23 (0.3%, 28.0%)	201 (2.5%, 7.7%)	44 (0.6%, 3.8%)	861 (10.8%, 41.0%)	OSCAR							
DDI Corpus 2011	68 (11.6%, 1.7%)	4 (0.7%, 4.9%)	162 (27.6%, 6.2%)	13 (2.2%, 1.1%)	4 (0.7%, 0.2%)	67 (11.4%, 0.8%)	DDI Corpus 2011						
DDI Corpus 2013	114 (8.9%, 2.9%)	5 (0.4%, 6.1%)	295 (22.9%, 11.4%)	23 (1.8%, 2.0%)	5 (0.4%, 0.2%)	112 (8.7%, 1.4%)	535 (41.6%,91.3%)	DDI Corpus 2013					
PK DDI Corpus	12 (7.2%, 0.3%)	0 (0.0%, 0.0%)	50 (30.1%, 1.9%)	1 (0.6%, 0.1%)	1 (0.6%, 0.0%)	22 (13.3%, 0.3%)	28 (16.9%, 4.8%)	51 (30.7%, 4.0%)	PK DDI Corpus				
NLM Corpus	35 (14.2%, 0.9%)	3 (1.2%, 3.7%)	50 (20.2%, 1.9%)	9 (3.6%, 0.8%)	2 (0.8%, 0.1%)	33 (13.4%, 0.4%)	48 (19.4%, 8.2%)	80 (32.4%, 6.2%)	26 (10.5%,15.7%)	NLM Corpus			
KEGG	403 (1.5%, 10.2%)	27 (0.1%, 32.9%)	777 (2.9%, 29.9%)	159 (0.6%, 13.8%)	511 (1.9%, 24.3%)	844 (3.2%, 10.6%)	218 (0.8%, 37.2%)	419 (1.6%, 32.6%)	77 (0.3%, 46.4%)	104 (0.4%, 42.1%)	KEGG		
TWOSIDES	51 (0.5%, 1.3%)	0 (0.0%, 0.0%)	82 (0.8%, 3.2%)	25 (0.3%, 2.2%)	40 (0.4%, 1.9%)	101 (1.0%, 1.3%)	14 (0.1%, 2.4%)	25 (0.3%, 1.9%)	11 (0.1%, 6.6%)	6 (0.1%, 2.4%)	724 (7.3%, 2.7%)	TWOSIDES	
DRUGBANK	150 (1.2%, 3.8%)	57 (0.5%, 69.5%)	1296 (10.7%, 49.9%)	319 (2.6%, 27.7%)	180 (1.5%, 8.6%)	490 (4.0%, 6.1%)	213 (1.8%, 36.3%)	448 (3.7%, 34.8%)	75 (0.6%, 45.2%)	111 (0.9%, 44.9%)	2143 (17.7%, 8.0%)	289 (2.4%, 2.9%)	DRUG BANK
DIKB	2 (0.4%, 0.1%)	21 (3.7%, 25.6%)	85 (15.2%, 3.3%)	33 (5.9%, 2.9%)	0 (0.0%, 0.0%)	7 (1.2%, 0.1%)	25 (4.5%, 4.3%)	36 (6.4%, 2.8%)	16 (2.9%, 9.6%)	8 (1.4%, 3.2%)	152 (27.1%, 0.6%)	69 (12.3%, 0.7%)	189 (33.7%, 1.6%)

Two percentages are shown, the first representing the percentage of PDDIs in the row-mention that overlapped with the column-mention, the second vice-versa. Because most datasets did not distinguish precipitant and object drugs, PDDI drug pairs were compared without consideration of the directionality of the interaction.

Table 4
A comparison of the performance of a previously published NLP pipeline for PDDIs with and without using a version of the merged PDDI dataset during the pre-processing phase of the pipeline.

	DrugBank 2011 (DDI Corpus 2011)			DrugBank 2013 (DDI Corpus 2013)			Medline 2013		
	Precision (%)	Recall (%)	F_1 (%)	Precision (%)	Recall (%)	F_1 (%)	Precision (%)	Recall (%)	F_1 (%)
Original NLP pipeline	68.56	71.92	70.20	85.87	81.11	83.42	67.57	52.63	59.17
Modified PDDI pipeline	68.59	72.58	70.53	85.65	81.67	83.61	65.79	52.63	58.48

Here again, the DDI Corpus 2013 had the greatest overlap but it represented only 2% of the PDDIs in the ONC High Priority source and 6% of the PDDIs in the CredibleMeds source. As was mentioned above, the overlap between the DDI Corpus 2011 and DDI Corpus 2013 was very high. However, the DDI Corpus 2013 represented slightly less than 1/3rd of the PDDIs present in the PK DDI Corpus, and roughly the same proportion of the PDDI present in the NLM Corpus.

Examining overlap across multiple sources, we found no PDDIs common to all 14 sources, only one PDDI (Rifampin/Bisoprolol) common to all four NLP sources, and only four PDDIs (Haloperidol/Clozapine, Triazolam/Voriconazole, Triazolam/Fluconazole, Midazolam/Fluconazole) common to the Bioinformatics/Pharmacovigilance sources. The Clinically-oriented information sources had PDDIs common to at most three sources, with the combination NDF-RT, ONC-HighPriority and OSCAR resulting in the greatest overlap (24) in this category. The number of PDDIs common to any three sources varied depending on which sources were selected. CredibleMeds, NDF-RT, ONC-Non Interruptive List had no common PDDIs, while DDI Corpus 2011, DDI Corpus 2013, and NLM CV Corpus had 38.

3.4. Testing if the PDDI dataset could improve the performance of a PDDI NLP algorithm

The version of the merged PDDI dataset that was used for NLP testing retained all of the original datasets except DDI Corpus 2011, DDI Corpus 2013, DrugBank, SemMedDB, KeGG, and TWOSIDES. The first two were excluded because they were exactly the same datasets as two of the NLP corpora used by Bui et al. for evaluation of their pipeline. DrugBank was excluded because it was used to create the two NLP corpora. The last three datasets were excluded because the number of false positives in these datasets is not known.

Table 4 shows a comparison of the performance of a previously published NLP pipeline for PDDIs, with and without using a version of the merged PDDI dataset during the pre-processing phase of the pipeline. As measured by the balanced F measure, using a the merged PDDI dataset in the pre-processing phase led to a slight performance improvement for the two DrugBank NLP datasets but a slight decrease for the MedLine dataset. The improved performance for the DrugBank datasets was primarily associated with an increased in recall of true PDDI mentions. No increase or decrease in recall was observed for the MedLine dataset.

3.5. Augmenting PDDI information in drug product labeling

The number of cases where PDDI information extracted from drug product labeling might be automatically augmented by other sources is shown in Table 3 in the cells where either of the two NLP datasets that focus on product labeling (PK DDI Corpus and NLM Corpus) intersects other information sources. In general, the percentage overlap was less than 50%. The greatest percentage of overlap was between the PK DDI Corpus and KEGG (46.4%), followed by PK DDI Corpus and DrugBank (45.2%), and NLM Corpus and DrugBank (44.9%).

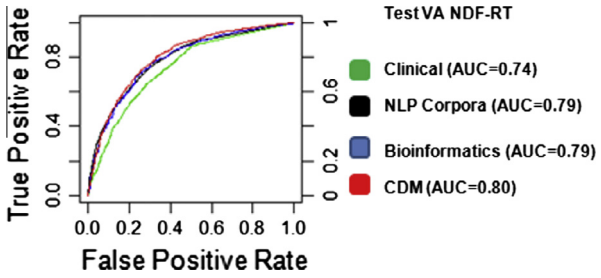

Fig. 2. ROC curves for VA NDF-RT dataset evaluated with the DDI models (Clinical, NLP Corpora and Bioinformatics models). The data related to the previous 3 models were combined to form the Common Data Model (CDM).

3.6. Applying the merged PDDI sources to predict new potential interactions

Fig. 2 shows that the PDDI pharmacovigilance protocol performs better using the combined CDM list of PDDIs as a reference standard than with just the clinical, NLP, or bioinformatics/pharmacovigilance datasets (AUROC = 0.80 vs 0.74, 0.79, and 0.79 respectively).

4. Discussion

To the best of our knowledge, while there has been related work comparing PDDI information sources (see below), this is the first study to attempt to bring together all publically available sources of PDDI information into a single dataset. A systematic search identified 14 public data sources shown in Table 1. The merged dataset[2,3], and code[4] used to create it, is available for download for research purposes.

An analysis of the overlap of PDDIs from the datasets with drugs that could be mapped to DrugBank identifiers found that there is very little overlap between or across publicly available PDDI resources. For many pairwise comparisons, the percentage of overlap was limited by the relatively small size of one of the datasets. For example, three data sources had no overlap with CredibleMeds which listed only 83 interactions. However, 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 clinically-oriented sets that focus on PDDIs of interest in most clinical settings (The two ONC lists [24,25] and CredibleMeds [23]).

Our finding of low agreement across public sources providing PDDI information is concordant with other studies, including the two mentioned in the introduction [7,8]. Most relevant to the current study, Peters, Bodenreider, and Bahr mapped drugs in the VA NDF-RT, DrugBank, and ONC High-priority list to RxNorm and then

[2] http://purl.org/net/drug-interaction-knowledge-base/PDDI-data-merged-conservative.
[3] http://purl.org/net/drug-interaction-knowledge-base/PDDI-data-merged-non-conservative.
[4] http://purl.org/net/drug-interaction-knowledge-base/PDDI-data-merging-project.

examined PDDI overlap [37]. Their study found only 24–30% overlap between the PDDIs listed in the VA NDF-RT and DrugBank. This same study found that roughly 60% of the ONC PDDIs were present in the VA NDF-RT, with roughly the same proportion present in DrugBank. We note that the level of overlap they found, while still low, was considerably higher than what was found in the current study. We think that the main reason for this is that we expanded the ONC High-priority list to include all agents with a known risk of torsades de pointes (TdP) as provided by Crediblemeds.org (see Section 3.3). Thus, the ONC High-priority list used in the current study had more than three times the number of PDDIs than that used in the Peters study (1150 vs 360).

Disagreement among sources of PDDI information has been known for more than a decade [44] and much work has been done in recent years to better understand the underlying reasons [1]. Many reasons have been identified by drug interaction experts including the need for a more standard way to assess the evidence that a drug combination can actually result in an interaction, agreement about how to assess if an interaction applies to a single drug or all drugs in its class, and guidance on how a drug information source should handle PDDIs listed in product labeling [2]. We would suggest adding to these reasons that there is currently no interoperable standard for representing PDDIs and associated evidence in a computable form (i.e., as assertions linked to evidence). Since evidence for PDDIs is distributed across several resources (e.g., product labeling, the scientific literature, case reports, social media), editors of drug information resources (public or proprietary) must resort to *ad hoc* information retrieval methods that can yield different sets of evidence to assess. Based on this information, we think that systems that provide access to the comprehensive lists, such as APIs into RxNorm [37], should provide results using a interoperable common data model for PDDIs, while also being careful to inform users that the lists may be incomplete with respect to PDDIs that drug experts suggest clinicians be aware of.

We tested a very basic approach to integrating the PDDI dataset into an existing NLP pipeline and found that it slightly improved the pipeline's performance on two of the three NLP datasets. We take this as evidence that the merged dataset has potential to improve PDDI NLP, but that more research is necessary to determine the optimal integration method. In our simple test, the PDDI dataset was used as a second check before filtering out sentences that were unlikely to be informative for training a machine learning classifier because they lacked "trigger" words. For example, without the method, sentences such as the following would be excluded from use in training the NLP classifier for the DrugBank 2013 NLP corpora (DDI Corpus 2013 in the current study):

> "If at all possible guanethidine should be discontinued well before minoxidil is begun."

> "Theophylline: Grepafloxacin is a competitive inhibitor of the metabolism of theophylline."

> "Dose adjustment of Sensipar may be required and PTH and serum calcium concentrations should be closely monitored if a patient initiates or discontinues therapy with a strong CYP3A4 inhibitor (e.g., ketoconazole, erythromycin, itraconazole."

The inclusion of these sentences might have helped improved the recall of a classifier. However, the method also led to the inclusion of other sentences that could potentially reduce a classifier's precision such as:

> "Less potent inhibitors include saquinavir, nefazodone, fluconazole, grapefruit juice, fluoxetine, fluvoxamine, zileuton, and clotrimazole."

> "Example inducers include aminoglutethimide, carbamazepine, nafcillin, nevirapine, phenobarbital, phenytoin, and rifamycins."

While the use of the PDDI dataset was beneficial for the Bui pipeline with two PDDI datasets, it harmed its performance with the MedLine dataset. We think that future NLP research should explore methods to improve this simple approach to integrating the merged PDDI dataset so that performance improvements on any dataset will result.

The merged dataset makes it possible for the pharmacovigilance text mining community to compare the representativeness of NLP corpora for PDDI text extraction tasks. The low level of overlap identified between current NLP corpora and clinically oriented datasets might be indicative that current corpora are not representative of all PDDIs. The merged dataset also specifies elements that future PDDI extraction tasks might want to include. For example, none of the NLP sources included in this study provide severity information or related drugs for the PDDIs.

While each PDDI source was developed for different purposes, we found some evidence that making the sources interoperable would indeed enable a better synthesis of PDDI knowledge. For example, there were more than a dozen interactions from drug product labeling sources that were also present in the ONC High-priority or Non-interruptive lists. This means that, for these PDDIs, an information retrieval system could use the merged dataset to provide consensus recommendations on alert prioritization to a reader of the electronic product label (e.g., a drug information compendia editor). Similarly, three interactions from drug product labeling were also present in the CredibleMeds dataset. Two of these interactions involved the cardiovascular drug digoxin and macrolide antibiotics erythromycin and clarithromycin. In both cases, CredibleMeds provided information on management options that were not present in the interactions extracted from the label. To be specific, the following two quotes show all of the information on the PDDI provided by the NLM Corpus and then the management recommendations provided by CredibleMeds:

> (NLM Corpus)
> "Erythromycin and clarithromycin (and possibly other macrolide antibiotics) and tetracycline may increase digoxin absorption in patients who inactivate digoxin by bacterial metabolism in the lower intestine, so that digitalis intoxication may result"
> (CredibleMeds)
> "Take Precautions
> Consider Alternatives: Consider alternative antimicrobials that do not inhibit P-glycoprotein (e.g., 2nd/3rd generation cephalosporins, penicillin, quinolones)
> Monitor: If alternatives are not appropriate, evaluate patients for evidence of digoxin toxicity (e.g., nausea, malaise, fatigue, visual changes, headache, arrhythmias), with downward digoxin dosage adjustments as needed. Also monitor for altered digoxin effect when clarithromycin or erythromycin are changed in dosage or discontinued. If digoxin is started in the presence of one of these agents, consider conservative initial digoxin dosing."

There are other examples where combining the information available across the multiple sources into the simple PDDI data model provided much richer description of these interactions. One example is the PDDI between clozapine and fluoxetine – DrugBank provided an unstructured text explanation of the PDDI between clozapine and fluoxetine. Additionally, DIKB enhanced the description of this PDDI by providing both a URI to the PRO ontology, indicating that the metabolic enzyme CYP2D6 is involved in the interaction, and also citations to evidence in the scientific literature (including product labeling) supporting the

mechanism. Finally, KEGG noted that the PDDI is labeled as a precaution in Japanese product labels. This case is interesting because, in theory, the mechanism of the PDDI could be used by a computer to suggest alternate treatment strategy such as substituting fluoxetine with an antidepressant (e.g., citalopram) that is not a potent CYP2D6 inhibitor.

The previous example of inferring a plausible management option from the merged information on the clozapine and fluoxetine PDDI suggests that there is great potential benefit from linking clinically useful PDDI information (e.g., effects, severity, and management options) with the chemical and pharmacological properties (e.g., chemical structure, function, pharmacokinetic and pharmacodynamic properties) of not only the participating drugs, but all related drugs. We think that such a representation should also make an explicit and logically sound connection between the drugs involved in a PDDI (usually pharmacologically active molecules) and the pharmaceutical products that are prescribed and administered to patients. For example, at the time of this writing, DrugBank currently lists 164 PDDIs involving ketoconazole.[5] Unfortunately, DrugBank provides no way to infer that most, if not all, of the PDDIs are relevant for only a handful of ketoconazole drug products that are systemic rather than topical. Thus, while drug-focused databases like DrugBank and KEGG provide a one-stop shop for information regarding chemical structure, function, interactions, and clinical application, their data models tend to conflate attributes of chemical entities with properties of complex pharmaceutical products. Indeed, recent work on the Drug Ontology (DrOn) suggests a compelling mechanism by which accurate descriptions of drug composition, route of administration, and function can be achieved using orthogonal ontologies such as ChEBI and the Protein Ontology [45]. We recently proposed a new approach to building an ontology for representing PDDI knowledge and evidence that we think might improve the utility of PDDI listings for clinical purposes [46]. In future work, we plan to apply the merged PDDI dataset from the current study to test and improve the ontology which is currently in an early stage of development.

There was a significant proportion of PDDIs from KEGG, TWOSIDES, and the NDF-RT that we were unable to map to DrugBank because we could find no URI for one or both of the involved drugs (Table 1). A query of the number of DrugBank drug records that provide cross reference URIs to PubChem or KEGG shows that the linkage between the data sources is currently incomplete.[6] Yet other cases involved drugs that were not in the RxNorm to DrugBank mapping [40] perhaps because it was not possible to create the mapping when the dataset was created (fall 2014). Regarding SemMedDB, we used "INTERACTS_WITH" relationships for drug interactions in this study but there exist alternative predictions that could potentially be used for inferring some drug interactions and we might explore additional types of predictions in SemMedDB for PDDIs in the future.

We also tested if the combined dataset would improve the performance of a recently published PDDI pharmacovigilance protocol. The protocol performed best when using the combined CDM model as a reference standard for PDDIs. We think that this is because the integrated PDDI model was more representative of the universe of all potential PDDIs and therefore led to a more generally predictive model. These preliminary results indicate that merging PDDIs from all available public sources may be important for PDDI pharmacovigilance.

Another interesting direction of future work is the improvement of drug interaction information in Wikipedia by adding expert-vetted PDDI data to Wikidata [47]. Wikipedia is easily accessible, routinely used by clinicians for finding medical information [48] and could play a significant role in global health promotion [49]. However, the quality of information on Wikipedia was found to be lower than in some proprietary resources [50]. Wikidata is a recent addition to the Wikipedia infrastructure, providing an open, efficient database for serving content to Wikipedia in different languages. We are working on integrating high-quality PDDI information to Wikidata and currently finishing the evaluation of a prototype for making these data available in Wikipedia.

4.1. Potential limitations

It is possible that our search missed some relevant sources, especially sources that might be available in non-English speaking countries. Also, it is possible that new sources have become available since the time we conducted the search. Our plans for future work include continuing to update the merged dataset with newly identified sources and making the data available as part of the Drug Interaction Knowledge Base project [51].

We focused on the non-proprietary information sources rather than commercially available PDDI sources. Both use PDDI content obtained from a handful of public sources included drug product labeling and the indexed scientific literature. While access to PDDI data is similar between non-proprietary and commercially available sources, they do differ in their efforts to assess the data and put it into perspective for users. For example, some sources use specific criteria for inclusion of PDDIs while others strive to be all inclusive. Facilitating the gathering of PDDI data is an important first step, but must be followed by knowledgeable assessment of the PDDI's risk to patients to avoid the current problem of excessive, inappropriate PDDI alerts. Improving the interoperability of these sources might simplify drug interaction experts' or compendia editors' task of acquiring all information known about a PDDI and making better decisions regarding its potential to cause patient harm. Moreover, the number of public PDDI sources has grown in recent years and, with proper evaluation, the PDDIs that they provide might enhance other widely used public information systems such as Wikidata [47].

The NLP resources contained many instances of atypical and error-prone drug group mentions in the PDDIs i.e. "drugs that are actively secreted by the kidney has not been investigated in humans", "drugs that do not require a similar titration", "live-attenuated vaccines", "drugs that can interfere with sinus node function". As the definitions for all possible drugs in each atypical drug group can be various and subjective to each reviewer, the manual review was not feasible and beyond the scope of this study. These challenges rise from the nature of the original data sources independent of our data model. Use of atypical drug mentions in the NLP PDDIs makes it very difficult if not impossible to come up with a reasonable and widely agreed upon comprehensive mapping to active ingredients, either manually or using automation. This is a critical issue that we think the NLP community needs to be aware of and address if future tools are to be of real use to the pharmacovigilance community. In the future, we will also investigate new approaches for overcoming the mapping challenges.

5. Conclusion

In this study, we combined all the publicly available sources of PDDI information using a common data model after conducting a comprehensive and broad search. The merged dataset consists of the synthesis of 14 different sources including 5 sources obtained from clinically-oriented information sources, 4 sources from Natural Language Processing Corpora, and 5 sources from Bioinformatics/Pharmacovigilance information sources.

[5] http://www.DrugBank.ca/drugs/DB01026.

[6] Of 6825 DrugBank drugs, 6100 contained PubChem `xref` URIs and 2535 contained KEGG drug or compound `xref` URIs.

We examined the overlap between and across the data sources and our analysis found out that there was little overlap and that there is heterogeneity between the information provided by each source. Another interesting finding was that most of the ONC High Priority PDDIs were not included in comprehensive drug interaction data sources such as NDF-RT, KEGG, and DrugBank. Despite this, the combined dataset provided a more complete overview of the PDDIs and a richer definition of the PDDIs than the original sources in cases where the PDDIs were noted in multiple resources.

Additionally, we investigated the cases where PDDI information extracted from drug product labeling might be automatically augmented by other sources. We also experimented with the use of the combined dataset in a recently published PDDI pharmacovigilance protocol. The results demonstrated an improvement on the performance of the protocol.

As future work, we plan to improve the mapping mechanisms and integrate high-quality PDDI information in the merged dataset to Wikidata and make it available for a broader community of stakeholders. Another future work is to make the combined dataset accessible to humans and computer programs as Semantic Web Linked Data.

Conflict of interest

None declared.

Acknowledgments

This work was supported primarily by National Library of Medicine Grant 1R01LM011838-01 and National Institute on Aging Grant K01AG044433. Additional support was provided by the NIGMS R01 GM107145, the Center for Expanded Data Annotation and Retrieval (NIH U54I117925), the Agency for Healthcare Research and Quality (K12HS019461), the NLM Research Participation Program (administered by the Oak Ridge Institute for Science and Education), and the Austrian Science Fund (FWF): [PP 25608-N15]. The content is solely the responsibility of the authors and does not represent the official views of the Agency for Healthcare Research and Quality or any of the other funding sources.

A special note of thanks to Lisa Hines, PharmD, Center for Health Outcomes & PharmacoEconomic Research, University of Arizona, for providing the initial list of individual drugs associated with the drug classes in the ONC High Priority drug interactions publication. Also, thanks to programmer Yifan Ning for his assistance throughout the project.

References

[1] L.E. Hines, D.C. Malone, J.E. Murphy, Recommendations for generating, evaluating, and implementing drug–drug interaction evidence, Pharmacother. J. Hum. Pharmacol. Drug Ther. 32 (4) (2012) 304–313.
[2] R. Scheife, L.E. Hines, R. Boyce, S. Chung, J. Momper, C. Sommer, D. Abernethy, J. Horn, S. Sklar, S. Wong, G. Jones, M. Brown, A. Grizzle, S. Comes, T. Wilkins, T. Borst, M. Wittie, A. Rich, D. Malone, Consensus recommendations for systematic evaluation of drug–drug interaction evidence for clinical decision support, Drug Saf., 2015.
[3] J.R. Nebeker, P. Barach, M.H. Samore, Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting, Ann. Intern. Med. 140 (10) (2004) 795–801.
[4] L.E. Hines, J.E. Murphy, Potentially harmful drug–drug interactions in the elderly: a review, Am. J. Geriatr. Pharmacother. 9 (6) (Dec. 2011) 364–377.
[5] FASTSTATS – Emergency Department Visits, 23-Sep-2013. <http://www.cdc.gov/nchs/fastats/ervisits.htm> (accessed 24.09.13).
[6] L. Magro, U. Moretti, R. Leone, Epidemiology and characteristics of adverse drug reactions caused by drug–drug interactions, Expert Opin. Drug Saf. 11 (1) (Jan. 2012) 83–94.
[7] L.M. Wang, M. Wong, J.M. Lightwood, C.M. Cheng, Black box warning contraindicated comedications: concordance among three major drug interaction screening programs, Ann. Pharmacother. 44 (1) (2010) 28–34.
[8] R.D. Boyce, C. Collins, M. Clayton, J. Kloke, J.R. Horn, Inhibitory metabolic drug interactions with newer psychotropic drugs: inclusion in package inserts and influences of concurrence in drug interaction screening software, Ann. Pharmacother. 46 (10) (2012) 1287–1298.
[9] CMS, Eligible Professional Meaningful Use Core Measures Measure 2 of 15, Centers for Medicare and Medicaid Services, 2010.
[10] M.S. Ridgely, M.D. Greenberg, Too many alerts, too much liability: sorting through the malpractice implications of drug–drug interaction clinical decision support, St. Louis Univ. J. Health Law Policy 5 (2) (2012) 257–296.
[11] A. Cami, S. Manzi, A. Arnold, B.Y. Reis, Pharmacointeraction network models predict unknown drug–drug interactions, PLoS One 8 (4) (2013) e61468.
[12] F. Azuaje, Drug interaction networks: an introduction to translational and clinical applications, Cardiovasc. Res. 97 (4) (Mar. 2013) 631–641.
[13] B. Vandervalk, E.L. McCarthy, J. Cruz-Toledo, A. Klein, C.J.O. Baker, M. Dumontier, et al., The SADI personal health lens: a web browser-based system for identifying personally relevant drug interactions, JMIR Res. Protoc. 2 (1) (2013) e14.
[14] S. Vilar, E. Uriarte, L. Santana, T. Lorberbaum, G. Hripcsak, C. Friedman, et al., Similarity-based modeling in large-scale prediction of drug–drug interactions, Nat. Protoc. 9 (9) (2014) 2147–2163.
[15] N.P. Tatonetti, P.P. Ye, R. Daneshjou, R.B. Altman, Data-driven prediction of drug effects and interactions, Sci. Transl. Med. 4 (125) (2012) 125ra31.
[16] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, J. Morissette, Bio2RDF: towards a mashup to build bioinformatics knowledge systems, J. Biomed. Inform. 41 (5) (2008) 706–716.
[17] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, et al., DrugBank 4.0: shedding new light on drug metabolism, Nucleic Acids Res. 42 (January) (2014) D1091–1097 (Database issue).
[18] Q.-C. Bui, P.M.A. Sloot, E.M. van Mulligen, J.A. Kors, A novel feature-based approach to extract drug–drug interactions from biomedical text, Bioinf. Oxf. Engl. 30 (23) (2014) 3365–3371.
[19] L.E. Hines, D. Ceron-Cabrera, K. Romero, M. Anthony, R.L. Woosley, E.P. Armstrong, et al., Evaluation of warfarin drug interaction listings in US product information for warfarin and interacting drugs, Clin. Ther. 33 (1) (Jan. 2011) 36–45.
[20] B. Pfistermeister, A. Saß, M. Criegee-Rieck, T. Bürkle, M.F. Fromm, R. Maas, Inconsistencies and misleading information in officially approved prescribing information from three major drug markets, Clin. Pharmacol. Ther. 96 (5) (2014) 616–624.
[21] R.D. Boyce, J.R. Horn, O. Hassanzadeh, A. de Waard, J. Schneider, J.S. Luciano, et al., Dynamic enhancement of drug product labels to support drug safety, efficacy, and effectiveness, J. Biomed. Semant. 4 (1) (2013) 5.
[22] E.L. Olvey, S. Clauschee, D.C. Malone, Comparison of critical drug–drug interaction listings: the department of Veterans Affairs medical system and standard reference compendia, Clin. Pharmacol. Ther. 87 (1) (2010) 48–51.
[23] Crediblemeds.org, 05-Oct-2013. <http://www.crediblemeds.org/> (accessed 05.10.13).
[24] S. Phansalkar, A.A. Desai, D. Bell, E. Yoshida, J. Doole, M. Czochanski, et al., High-priority drug–drug interactions for use in electronic health records, J. Am. Med. Inform. Assoc. JAMIA 19 (5) (2012) 735–743.
[25] S. Phansalkar, H. van der Sijs, A.D. Tucker, A.A. Desai, D.S. Bell, J.M. Teich, et al., Drug–drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records, J. Am. Med. Inform. Assoc. JAMIA 20 (3) (2013) 489–493.
[26] N.R. Crowther, A.M. Holbrook, R. Kenwright, M. Kenwright, Drug interactions among commonly used medications. Chart simplifies data from critical literature review, Can. Fam. Physician Médecin Fam. Can. 43 (November) (1997) 1972–1976 (1979–1981).
[27] Oscar-McMaster, OSCAR Electronic Medical Record,OSCAREMR, 2014. <http://oscar-emr.com/> (accessed 13.10.14).
[28] I. Segura-Bedmar, P. Martınez, D. Sánchez-Cisneros, The 1st DDIExtraction-2011 challenge task: Extraction of Drug–Drug Interactions from biomedical texts, 2011.
[29] I. Segura-Bedmar, P. Martınez, and M. Herrero-Zazo, Semeval-2013 task 9: extraction of drug–drug interactions from biomedical texts, in: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), 2013.
[30] R. Boyce, G. Gardner, and H. Harkema, Using natural language processing to extract drug–drug interaction information from package inserts, in: BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Montréal, Canada, 2012, pp. 206–213.
[31] Johann Stan, A Machine-Learning Approach for Drug–Drug Interaction Extraction from FDA Structured Product Labels, Presented at the 2014 National Library of Medicine Training Conference, Pittsburgh PA, USA, 17-Jun-2014.
[32] M. Takarabe, D. Shigemizu, M. Kotera, S. Goto, M. Kanehisa, Network-based analysis and characterization of adverse drug–drug interactions, J. Chem. Inf. Model. 51 (11) (2011) 2977–2985.
[33] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, et al., Data, information, knowledge and principle: back to metabolism in KEGG, Nucleic Acids Res. 42 (January) (2014) D199–205 (Database issue).
[34] R. Boyce, C. Collins, J. Horn, I. Kalet, Computing with evidence Part II: An evidential approach to predicting metabolic drug–drug interactions, J. Biomed. Inform. 42 (6) (2009) 990–1003.
[35] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T.C. Rindflesch, SemMedDB: a PubMed-scale repository of biomedical semantic predications, Bioinformatics 28 (23) (2012) 3158–3160.

[36] J.R. Horn, P.D. Hansten, L.-N. Chan, Proposal for a new tool to evaluate drug interaction cases, Ann. Pharmacother. 41 (4) (Apr. 2007) 674–680.
[37] L. Peters, O. Bodenreider, N. Bahr, Evaluating drug–drug interaction information in NDF-RT and DrugBank, in: Proceedings of the Workshop on Vaccines and Drug Ontology Studies (VDOS-2014), Houston, Texas, 2014.
[38] RxNorm, 2014. <http://www.nlm.nih.gov/research/umls/rxnorm/> (accessed 16.10.14).
[39] RxNorm Documentation, 2014. <http://www.nlm.nih.gov/research/umls/rxnorm/docs/2012/rxnorm_doco_full_2012-3.html#s8_0> (accessed 16.10.14).
[40] O. Hassanzadeh, Q. Zhu, R. Freimuth, R. Boyce, Extending the 'Web of Drug Identity' with knowledge extracted from United States product labels, in: Proc. 2013 AMIA Summit Transl. Bioinforma., Mar. 2013.
[41] InChI, 2014. <http://www.iupac.org/home/publications/e-resources/inchi.html> (accessed 16.10.14).
[42] F.J. Defalco, P.B. Ryan, M. Soledad Cepeda, Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure, Heal Serv. Outcomes Res. Methodol. 13 (1) (2013) 58–67.
[43] Q. Zhu, G. Jiang, C.G. Chute, Profiling structured product labeling with NDF-RT and RxNorm, J. Biomed. Semant. 3 (1) (2012) 16.
[44] T.K. Hazlet, T.A. Lee, P.D. Hansten, J.R. Horn, Performance of community pharmacy drug interaction software, J. Am. Pharm. Assoc. 41 (2) (2001) 200–204 (Washington, DC 1996).
[45] W.R. Hogan, J. Hanna, E. Joseph, M. Brochhausen, Towards a consistent and scientifically accurate drug ontology, in: ICBO 2013 Conference Proceedings, 2013.
[46] M. Brochhausen, J. Schneider, D. Malone, P. Empey, W.R. Hogan, R.D. Boyce, Towards a foundational representation of potential drugdrug interaction knowledge, in: Drug Interaction Knowledge Representation (DIKR 2014), Houston, Texas, 2014.
[47] Wikidata, <http://www.wikidata.org/wiki/Wikidata:Main_Page> (accessed 05.10.13).
[48] M. Kritz, M. Gschwandtner, V. Stefanov, A. Hanbury, M. Samwald, Utilization and perceived problems of online medical resources and search tools among different groups of European physicians, J. Med. Internet Res. 15 (6) (Jun. 2013) e122.
[49] J.M. Heilman, E. Kemmann, M. Bonert, A. Chatterjee, B. Ragar, G.M. Beards, et al., Wikipedia: a key tool for global public health promotion, J. Med. Internet Res. 13 (1) (Jan. 2011) e14.
[50] K.A. Clauson, H.H. Polen, M.N.K. Boulos, J.H. Dzenowagis, Scope, completeness, and accuracy of drug information in Wikipedia, Ann. Pharmacother. 42 (12) (Dec. 2008) 1814–1821.
[51] Richard D. Boyce, The Drug Interaction Knowledge Base, 2014. <http://purl.org/net/drug-interaction-knowledge-base/> (accessed 14.10.14).