Integrating Heterogeneous Knowledge Sources to Acquire Executable Drug-Related Knowledge

Xiaoyan Wang MPhil, Herbert S. Chase MD, Jianhua Li MD, George Hripcsak MD MS, Carol Friedman PhD

Department of Biomedical Informatics, Columbia University, New York, NY

ABSTRACT

Knowledge of medical entities, such as drug-related information is critical for many automated biomedical applications, such as decision support and pharmacovigilance. In this work, heterogeneous information sources were integrated automatically to obtain drug-related knowledge. We focus on one type of knowledge, drug-treats-condition, in the study and propose a framework for integrating disparate knowledge sources. Evaluation based on a random sample of drug-condition pairs indicated an overall coverage of 96%, recall of 98% and a precision of 87%. In conclusion, the preliminary study demonstrated that the knowledge generated from this study was comparable to the manually curated gold standard and that this method of automatically integrating knowledge sources is effective. The automated method should also be applicable to integrate other clinical knowledge, such as drug-related knowledge with omics information.

INTRODUCTION

Medical knowledge, such as drug specific information, is a major concern from bench to bedside. Such knowledge is essential for automated clinical applications, such as pharmacovigilance, quality of care and decision support. Multiple disparate knowledge sources, such as MicroMedex and Drugs.com, provide complementary drug related knowledge.[1, 2] There are, however, some major problems in using these knowledge sources. One of key issues is that biomedical drug-related knowledge sources are not always up-to-date or complete because this information is constantly evolving as new discoveries are made and practices change over time, requiring that the knowledge bases be constantly updated. On the other hand, knowledge concerning drugs is complex and voluminous. As the number of knowledge source increases, it becomes more difficult to locate sources appropriate to a given purpose (i.e. what are the indications of a drug/drug classes or what are the possible side effects of a drug). Therefore it would be beneficial to automatically create and update executable knowledge bases for a given specific purpose.

Traditional creation of knowledge bases typically starts with manual curation. Some of these knowledge bases, such as Micromedex, ensure accuracy and quality and serve as a gold standard for many clinical applications. Manual construction, however, is laborious, and time-consuming, and also costly but critical to keep up-to-date. In addition, the proprietary nature of some manually curated knowledge bases limits their usage for automated clinical applications, particularly for research uses.

There has been some research concerned with generation of automatic drug-related knowledge bases, such as indications for drugs. Some of this information occurs but is often hidden in either the biomedical literature or narrative clinical reports. One problem is that textual information is difficult to access reliability. Another is that these relations are frequently not explicitly stated in the text. Various NLP systems have been applied to the biomedical literature and narrative clinical reports for the purpose of extracting and establishing knowledge base for drugs. Rindflesch and colleagues extracted drug and disease entities from the Mayo Clinic notes using SemRep and constructed a repository of drug-disease co-occurrences to validate inferences produced by SemRep about drug treatments for diseases.[3, 4] In another study, these researchers proposed a methodology based on automatic summarization to identify drug information in Medline citations and then present the results to users in a convenient form. Their results indicated that automatic summarization from information in the literature can provide a valuable adjunct to curated knowledge databases for drugs and their corresponding indications.[5, 6] Similarly, another group used an NLP system Biomedlee and semantic processing to extract knowledge from Cochrane reviews knowledge to assess drug, therapy and disease concepts.[7] Chen et al in our group acquired knowledge of disease-drug associations automatically from both the literature and from clinical documents using text mining and statistical approaches. [8]

With the increase and advance of drug-related knowledge, along with the increasing availability of information sources, selecting the most relevant sources for a given purpose is becoming challenging. Sharp et al. examined 23 sources that provided drug information and characterized drug information using 39 dimensions. The group also proposed a framework for characterizing information sources.[9] The framework has been proved to be useful for

comparing drug resources and locating sources appropriate to a given need.

In this study, we experiment with an approach involving the integration of knowledge collected from disparate sources to automatically generate an executable and publically available drug-indication knowledge base. We integrated three complementary publically-available drug-related information sources consisting of the NDF-RT resource in NLM's Unified Medical Language System (UMLS)[10], FDA's Adverse Event Reporting System (AERS)[11] and SemMed[4], focusing on the 'drug-indication' relation, and we evaluated the coverage and precision obtained using the individual resources as well as the combined one.

MATERIALS & METHODS

Materials

1. NDF-RT in UMLS

The NLM's Unified Medical Language System (UMLS) Knowledge Sources consists of the Metathesaurus (Meta), Semantic Network, and SPECIALIST Lexicon. The Meta is a collection of biomedical-related concepts with unique concept identifiers (CUI) assigned to each of them. The Meta is based on numerous source vocabularies (i.e. 153 sources' in 2009AB release) and contains rich information about concepts, including variations and various relationships between them. Among the sources which used in this study is a set of relations from The VHA National Drug File Reference Terminology (NDF-RT), such as 'a condition may be treated by a drug'.

2. The FDA Adverse Event Reporting System (AERS)

AERS is a database supported by the FDA's post-marketing safety surveillance program for all approved drug and therapeutic biologic products. Reporting of adverse events is voluntary in the United States for health care professionals and consumers but mandatory for manufacturers. In addition to reporting adverse events associated with drugs, other information, such as 'indications for the drugs', are also reported. Reports to the system could be heterogeneous and mixed with correct and incorrect information. For example, '*Diabetes*', '*Hypertensive disease*', '*Chronic Obstructive Airway Disease*', '*Vomiting*' and '*Bradycardia*' were all be reported as indications for drug *carvedilol*. Some statistical techniques, such as conditional probability, may therefore be necessary to use in order to determine the indication occurrences that are likely to be correct based on frequency. In this study, the drug indications mentioned in the AERS database from 2004 through 2008 were used.

3. SemMED

SemMED is a database generated from the automatic summarization of Medline citations using SemREP, a natural language processing system that extracts semantic predications from the medical literature.[3, 4] It contains predications such as "a drug 'treats' a condition". Data processed from the literature could be questionable as well. For example, '*Hot flushes*', '*Hematologic Neoplasms*', '*Alzheimer's Disease*' were described as treated by '*Paroxetine*'. Similar to data from AERS databases, conditional probability should be computed to reduce the noisy data.

4. Drugs of interest

We focused on drugs used in inpatient settings. The data warehouse of NewYork-Presbyterian Hospital (NYPH) collects and maintains a variety of structured and unstructured data for patient records. Textual discharge summaries for patients admitted to NYPH in 2004 were used in this study to determine the drugs that were covered by the knowledge sources.

Methods

1. Generating knowledge bases

In this study, we focus on generating knowledge for drug-treats-condition. The framework for generating knowledge bases from each source consists of six main phases: (1) collecting the sets of data from each knowledge source. Data from a) the UMLS 2009AB, b) AERS (consisting of reports during 2004-2008), and c) SemMed (consisting of articles published during 2006-2007) were collected in this study; (2) selecting drug-condition pairs. Relations of 'may_be_treated_by' were selected from the MRREL table of UMLS, 'indications of drug' were selected from AERS, and predicates of 'treats' from SemMed; (3) semantic mapping and filtering. Drug and condition entities obtained from each resource were mapped to UMLS codes. The semantic classes of the UMLS codes were used to select the appropriate information types for this study. For example, the UMLS codes that were extracted and that corresponded to the semantic classes *Pharmacologic Substance* (T121), *Antibiotic* (T195), and *Clinical Drug* (T200) were used to select the drug entities. The UMLS codes that corresponded to the UMLS semantic classes *Disease or Syndrome* [T047], *Mental or Behavioral Dysfunction* [T048], *Neoplastic Process* [T191], *Sign or Symptom* [T184] and *Finding* [T033] were used for condition entities; (4) Drug normalization. The RxNorm (RxNorm vocabulary maintained by the National Library of Medicine) defines several types of relationships between concepts. For example, it relates generic classes and the trade names of drugs, by relations such as *tradename_of* and *has_tradename.* This was

used to map all trade names to their generic names[12]; (5) applying conditional probability cutoffs; Our initial experiments indicated that a large volume of indications from SemMed and AERs with relatively sparse occurrences were false positives (FP) and to reduce the FP rate we used. conditional probabilities P(condition|drug) for each drug-condition pair from SemMed and AERS. Different cutoffs of the probabilities were explored and different knowledge bases created; (6) integrating individual resources. Data from the previous steps were combined to form knowledge bases for further analysis.

2. Evaluation:

We evaluated the individual and integrated knowledge bases using 20 drugs selected from the drugs of interest. Metrics of recall and precision were calculated by comparing the results to the reference standard, which is described below, and qualitative analysis was performed to further understand the errors.

Evaluation Dataset The drugs extracted from the NYPH discharge summaries were ranked and stratified according to frequency of occurrence. The strata of top 1-50, 51-150, 151-300 and >300 were considered to represent the "most common", "common", "less common" and "rare" drugs. Five drugs were randomly selected from each stratum for evaluation.

Reference standard To evaluate the accuracy of the drug indications in each knowledge base, a reference standard was used that consisted of two experts who were presented with the drugs in the evaluation dataset. The experts classified the drug-treat-condition information for each drug based on their medical knowledge and Micromedex, a well-respected, evidence-based and reliable reference. The indications were classification into two categories 1). FDA approved indications; and 2) non-FDA approved indications. The classifications from each of the experts were then combined to create a reference standard as follows: (1) if a drug-condition pair was agreed upon by the two experts, the pair was chosen as the reference standard; (2) if a pair was not agreed upon, a random response was chosen to be the reference standard.

Quantitative evaluation Two metrics were used to assess the performance of each experiment. Recall was calculated as the ratio of the number of distinct drug-condition pairs that were identified by an experiment over the total number of the corresponding pairs in the reference standards (i.e. TP/(TP+FN)). Precision was measured as the ratio of the number of distinct drug-condition pairs returned by an experiment that were correct according to the reference standards divided by the total number of pairs found by the experiment (i.e. TP/(TP+FP)).

Qualitative evaluation To understand the types of errors affecting recall and precision for acquiring drug-indication relations, a qualitative analysis was performed. Incorrect drug-condition pairs were manually reviewed for further analysis. False pairs were categorized into the following groups: 1) 'broad indication' in which a condition is related to an indication for the drug but at a broader granularity, such as *mood disorders* for *paroxetine* which is used to treat *depression* (*paroxetine-(depression)-mood disorders)* 2) 'symptom of an indication' in which a condition is a symptom of one the indications for a drug (*carvedilol-(myocardial infarction)-shortness of breath*) 3) 'related to a comorbidity of an indication' (*lisinopril-(hypertension)-diabetes*) 4) 'adverse drug event' in which a condition is an adverse event that the drug causes (*carvedilol –bradycardia*). 5) 'no known association' in which there is no association between the drug and the condition according to current knowledge (*carvedilol- Tardive dyskinesia)*

RESULTS

1. Data Statistics

There were 1997 unique drug concepts among 2004 discharge summaries and the total number of occurrences of the drug concepts was 143,828. The coverage of drugs used for inpatients in 2004 for each of the three individual sources and the combined source is summarized in Table 1. The coverage of the individual sources ranged from 60% (SemMed) to 95% (AERS) of the unique drug concepts. The occurrence of drugs that were covered in SemMed and the NDF-RT in the UMLS was around 70% whereas it was 94% in AERS. The combined coverage was 96% for both the unique drug concepts and total occurrences.

Table 1 Drug Coverage in Knowledge Sources

Item	Total	SemMed (%)	AERS (%)	UMLS (NDF-RT) (%)	Combined (%)
Unique drug con-cepts	1997	1196 (60%)	1893 95%)	1418 (72%)	1908 (96%)
Total drug occur-ence	143,828	107,987 (75%)	135,247 (94%)	100,679 (70%)	138,636 (96%))

2. Results of Evaluation

Quantitative evaluation

A total of 1643 unique indications were determined to be associated with the 20 drugs. Validity result indicated a kappa of 0.67 between the two experts. Recall and precision were measured for FDA labeled

Table 2 Quantitative Evaluation

Metrics	Recall		Precision	
Database	FDA labeled	Both FDA and non FDA	FDA labeled	Both FDA and non FDA
SemMed cutoff (0.1%)	0.82	0.72	0.34	0.37
SemMed cutoff (2%)	0.67	0.59	0.51	0.51
AERS cutoff (0.2%)	0.94	0.83	0.60	0.48
AERS cutoff (2%)	0.78	0.67	0.84	0.73
UMLS (NDF-RT)	0.87	0.52	0.92	0.87
*Combined	0.98	0.74	0.87	0.83

*dataset combined with SemMed 2% cutoff, AERS 2% cutoff and UMLS NDF-RT.

indications, and also for combined FDA and non FDA labeled indications respectively are shown in Table 2. For sources obtained using a conditional probability, the relations obtained using a higher cutoff resulted in a higher precision (i.e. for AERS, the precision was 0.60 with a cutoff of 0.2% versus. precision of 0.84 with a cutoff of 2%) and lower recall (0.94 with cutoff of 0.2% versus 0.78 with a cutoff of 2%). For FDA labeled indications, recall of individual sources with high cutoff resulted in recalls of 0.67, 0.78 and 0.87 respectively, whereas the combined knowledge base resulted in a recall of 0.98. Precision of the individual sources were 0.51, 0.84, 0.92 respectively, whereas prevision of the combined knowledge base resulted in a precision of 0.87. Recall and precision for FDA and non FDA combined followed a similar pattern as that of the FDA labeled indication.

Qualitative evaluation

Overall, we determined that 35% of errors were caused by broad indications, 8% of errors were symptoms of indications, 29% of errors were comorbidities, 3% of errors were drug-ADE associations, and 25% of errors were no associations. Examples of the qualitative analysis are shown in Table 3 for several drugs.

For individual sources, AERS tended to result in more of the broad indication (65% of the total amount of broad indications) category of errors, whereas SemMed tended to result in more errors concerning comorbidities (72% of the total amount of comorbidity category). Errors in the UMLS NDF-RT resource tended to be the broad indications (i.e. *Epstein--Barr Virus Infections,* which is a class of viruses instead of the more specific *herpes simplex infection,* which is an indication for *Acyclovir*).

DISCUSSION

The aim of the present study was to automatically generate knowledge of drug-indications by integrating heterogeneous and complementary sources. Our results indicate that use of this automated approach is feasible and effective for acquiring or updating drug-indications.

In this work, we intentionally chose three complementary sources which were publically accessible, with SemMed as knowledge summarized from the literature, AERS as knowledge reported by clinicians and consumers, and the UMLS NDF-RT as manually curated knowledge. The NDF-RT offers the highest precision but the lowest recall and coverage, whereas AERS and SemMed had higher recalls and coverage and moderate precision, and the errors they caused were of different types. The integration of the three sources improved the acquisition of drug-indications with high coverage (0.96), high recall (0.98) and reasonable precision (0.87). Different cutoffs based on conditional probability were explored for AERS and SemMED to improve the results. Not surprisingly, higher cutoffs resulted in higher precision but lower recall. However, we observed that lower cutoffs (i.e. AERS with 0.2% cutoff) tended to yield more symptoms which were related to the indications. For some applications such as pharmacovigilance, understanding symptoms of indications is important since symptoms of indications could confound the detection of ADEs. Therefore, different cutoffs should be chosen for different applications, depending on the requirements for sensitivity versus precision. One of the interesting phenomena we observed in this work involves the different granularities for expressing diseases and symptoms associated with drug indications. For example, compared to the gold standard for '*carvedilol* –(treats)-*hypertension',* we obtained a range of granularities and types of hypertension: *hypertensive disease, renovascular hypertension*, *essential hypertension,* and *hypertensive crisis*. Such knowledge concerning different granularities would be interesting to investigate further to help develop or expand an ontology of diseases and symptoms.

One of the limitations in this study is that for coverage we used inpatient reports, which may not reflect usage in the general clinical setting. In the future, we will combine inpatient and outpatient data to explore more comprehensive drug related usage. A second limitation of this investigation is that, although the evaluation involved a total of 1643 drug-condition relations, they corresponded to a set of 20 drugs. Also, the reference standard was obtained

Table 3 Qualitative Analysis

Drug	Gold Standard		Error Analysis					
	FDA labeled indication	Non FDA labeled indication	Source	Broad Indication	Symptom of indication	Comorbidities	ADE	No known assoc
Carve-dilol	'Heart failure' 'Hypertension' 'Impaired left Ventricular function – Myocardial infarction'	'Angina pectoris Chronic' 'Atrial arrhythmia' 'Cardiac dysrhythmia', 'Congestive cardiomyopathy' 'Congestive heart failure, Nitrate tolerance' 'Disease of liver' 'Gastroesophageal varices; Prophylaxis' 'Surgical procedure'	**SemMed**	'Endothelial dysfunction' 'Postinfarction' 'Cardiovascular diseases'	'Vomiting' 'Cachexia'	'Diabetes' 'Diabetes, Mellitus-Non-Insulin-Dependent' 'Neoplasm Metastasis' 'Metabolic syndrome'		'Tardive dyskinesia'
			AERS	'Heart Diseases' 'Heart irregular' 'Vascular diseases' 'Tachycardia'			'Bradyca -rdia'	
			UMLS NDF-RT					
Paro-xetine	'Generalized anxiety disorder' 'Major depressive disorder' 'Obsessive-compulsive disorder' 'Panic disorder' 'Posttraumatic stress disorder' 'Premenstrual dysphoric disorder' 'Social phobia'	'Compulsive gambling' 'Drug-induced depressive state' 'Fibromyalgia' 'Hot sweats' 'Premature ejaculation'	**SemMed**	'Primary Insomnia' 'physical disorders'	'Feeling tense'	'Hematologic' 'Neoplasms' 'Alzheimer's Disease' 'Somatic pain'	'Dizzines' 'Pruritus' 'Neuroleptic malignant syndrome'	
			AERS	'Personality Disorders' 'Mood Disorders'	'Sleeplessness' 'Panic' 'Premenstrual syndrome'	'Migraine Disorders' 'Schizophrenia' 'Sleep Disorders'		
			UMLS NDF-RT	'Sexual Dysfunction'				

using only two experts. The experts generally agreed on the FDA-labeled indications but agreed less on the non-FDA labeled indications. There are potential problems with this strategy, one of which is inter-rater agreements among experts. Researchers have shown that inter-rater reliability is an issue[13]. Our results indicate a kappa of 0.67 between the two experts, which is acceptable. There are more sophisticated techniques to evaluate and improve the reliability of the reference standard, as discussed by Hripcsak and Heitjan.[14] A more comprehensive evaluation involving a larger sample size and a more reliable reference standard will be undertaken in future work

CONCLUSION

Creation and updating of medical knowledge is challenging. Therefore it is important to automatically create and update executable drug-related knowledge bases so that they can be used for automated applications. Our results suggest that the drug-indication knowledge generated by integrating complementary databases was comparable to the manually curated gold standard. Knowledge automatically acquired from these disparate sources could be applied for many clinical applications such as pharmacovigilance and document summarization. In the future, the methodology could be extended to integrating data collected in research with patient clinical data and linking omics sciences.[15]

Acknowledgments

The authors thank Lyudmila Shagina for assistance with MedLEE. This work is supported in part by grants T15-LM007079 (XW), R01 LM010016 (CF), R01 LM010016-0S1 (CF), R01 LM010016-0S2 (CF), R01 LM008635 (CF), and R01 LM06910 (GH) from the National Library of Medicine.

References

1. http://www.micromedex.com/.
2. http://www.drugs.com/.
3. Fiszman, M.T.C.R. and H.K. *Summarizing drug information in Medline citations.* Proceedings of the AMIA Annual Symposium, , 2006. **254-8**.
4. Fiszman, M., T.C. Rindflesch;, and H. Kilicoglu, *Abstraction summarization for managing the biomedical research literature.* Proceedings of the Workshop on Computational Lexical Semantics, 2004. **pp. 76-83. HLT-NAACL**.
5. Bray BE, et al., *Using semantic predications to characterize the clinical cardiovascular literature.* AMIA Annu Symp Proc, 2008 **Nov 6:887.**
6. Fiszman M, et al., *Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation.* J Biomed Inform, 2009.**Oct;42(5):801-13.**
7. Borlawsky T, Friedman C, and L. YA, *Generating executable knowledge for evidence-based medicine using natural language and semantic processing.*AMIA Annu Symp Proc, 2006. **56-60**.
8. Chen, E.S., et al., *Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study.* J Am Med Inform Assoc, 2008. **15**(1): p. 87-98.
9. Sharp M, Bodenreider O, and W. N., *A framework for characterizing drug information sources.* AMIA Annu Symp Proc. , 2008 **Nov 6:662-6**.
10. http://www.nlm.nih.gov/research/umls/.
11. http://www.fda.gov/Drugs/GuidanceComplianceRegulatory Information/Surveilance/AdverseDrugEffects/default.htm.
12. Liu S, M.W., Moore R, Ganesan V, Nelson S., *RxNorm: prescription for electronic drug information exchange.* IT Professional, 2005. **;7(5):17-2**.
13. Radev D, T.S., Saigon H, et al, *Evaluation challenges in large-scale document summarization.* Proceedings of the 41st annual meeting on association for computational linguistics, 2003. **375-382.**
14. Hripcsak, G. and D.F. Heitjan, *Measuring agreement in medical informatics reliability studies.* J Biomed Inform, 2002. **35**(2): p. 99-110.
15. Burgun A and O. Bodenreider, *Accessing and integrating data and knowledge for biomedical research.* Yearb Med Inform,2008.**91-101.**