Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Using semantic predications to uncover drug–drug interactions in clinical data

Rui Zhang [a], Michael J. Cairelli [b,*], Marcelo Fiszman [b], Graciela Rosemblat [b], Halil Kilicoglu [b], Thomas C. Rindflesch [b], Serguei V. Pakhomov [a,c], Genevieve B. Melton [a,d]

[a] Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA
[b] Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[c] College of Pharmacy, University of Minnesota, Minneapolis, MN, USA
[d] Department of Surgery, University of Minnesota, Minneapolis, MN, USA

ARTICLE INFO

Article history:
Received 25 April 2013
Accepted 13 January 2014
Available online xxxx

Keywords:
Drug–drug interactions
MEDLINE
SemRep
SemMedDB
Natural language processing
Clinical data
Semantic predication

ABSTRACT

In this study we report on potential drug–drug interactions between drugs occurring in patient clinical data. Results are based on relationships in SemMedDB, a database of structured knowledge extracted from all MEDLINE citations (titles and abstracts) using SemRep. The core of our methodology is to construct two potential drug–drug interaction schemas, based on relationships extracted from SemMedDB. In the first schema, Drug1 and Drug2 interact through Drug1's effect on some gene, which in turn affects Drug2. In the second, Drug1 affects Gene1, while Drug2 affects Gene2. Gene1 and Gene2, together, then have an effect on some biological function. After checking each drug pair from the medication lists of each of 22 patients, we found 19 known and 62 unknown drug–drug interactions using both schemas. For example, our results suggest that the interaction of Lisinopril, an ACE inhibitor commonly prescribed for hypertension, and the antidepressant sertraline can potentially increase the likelihood and possibly the severity of psoriasis. We also assessed the relationships extracted by SemRep from a linguistic perspective and found that the precision of SemRep was 0.58 for 300 randomly selected sentences from MEDLINE. Our study demonstrates that the use of structured knowledge in the form of relationships from the biomedical literature can support the discovery of potential drug–drug interactions occurring in patient clinical data. Moreover, SemMedDB provides a good knowledge resource for expanding the range of drugs, genes, and biological functions considered as elements in various drug–drug interaction pathways.

1. Introduction

Translational informatics is a relatively new field that emerged to bridge the gap between biomedical research and clinical practice. This gap is exacerbated by the rapid growth of knowledge contained in the biomedical literature and the relatively slow manual access to this information due to its unstructured nature. The growing gap between scientific knowledge and clinical practice makes the tasks of translational informatics all the more important and urgent.

Electronic health records (EHR) are being used by clinicians as primary tools for documentation and communication in clinical practice, and this trend can be expected to continue. Clinical data contain highly personalized patient information that has the potential to be explored for clinical research and especially the complex care of patients with multiple and chronic disorders. Many of these more complex patients have a long list of medications with new drugs added, existing drugs removed, or medication doses adjusted frequently due to the nature of their conditions and the need for disease management (e.g., medication titration or changes for a poorly controlled hypertensive patient). Even a single drug can have a diverse effect profile in individual patients, so the combination of multiple drugs increases the possibility of unexpected effects. One possible reason for unexpected medication effects are potential interactions between drugs within a patient's medication list. Such interactions can make the therapeutic effect of one or more prescribed medications weaker (or stronger) than intended or make side effects more pronounced.

* Corresponding author. Address: BG 38A Rm 9S912A, 8600 Rockville Pike, Bethesda, MD 20892, USA. Fax: +1 301 496 0673.
E-mail addresses: zhan1386@umn.edu (R. Zhang), cairellimj@mail.nih.gov (M.J. Cairelli), fiszmanm@mail.nih.gov (M. Fiszman), grosemblat@mail.nih.gov (G. Rosemblat), kilicogluh@mail.nih.gov (H. Kilicoglu), trindflesch@mail.nih.gov (T.C. Rindflesch), pakh0002@umn.edu (S.V. Pakhomov), gmelton@umn.edu (G.B. Melton).

1532-0464/$ - see front matter © 2014 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jbi.2014.01.004

Drug–drug interactions (DDIs) are a serious concern in clinical practice, as physicians strive to provide the highest quality and safety in patient care. While DDI lists are commonly used in clinical practice to alert clinicians during prescribing, some DDIs result from combinations or various mechanistic pathways that are not widely known. The traditional model for DDIs consists of considering the effect of one drug on a protein or other targets that are involved in the metabolism or transport of a second. The effect of the second drug may have the same target as the first drug or a different target [1]. This can be considered a direct interaction between the two drugs and many examples of this type of interaction affect cytochrome P450 metabolism [2–5]. Significant interactions may result beyond this traditional schema and can be extended to different genes being affected within the same biological pathway [6]. These interactions can also extend from pathway to biological processes for a particular clinical application. In other words, when two drugs are not linked through a specific gene network but target the same biological function, they can produce an effect at the clinical level that is not evident at the level of gene expression or protein interaction, especially by compounding an effect that can be induced through distinct pathways. For example, dehydration can be caused by failure of the colon to reabsorb water leading to diarrhea. Dehydration can also be caused by increased water output in the urine, or diuresis. If one drug in a patient's medication list induces diarrhea while another is a diuretic, these effects would be compounded, increasing the risk of dehydration and its complications. To our knowledge, no previous attempt at identifying DDIs includes these clinical-level physiological effects in searching for DDI interactions.

DDIs can be identified through several approaches, including *in vitro* pharmacology experiments [7,8], *in vivo* clinical pharmacology studies [8,9], and pharmacoepidemiology studies [10]. However, these methods are limited by the need to focus on a small set of target proteins and drugs and are, therefore, slow to elucidate an exhaustive set of DDIs while new drugs are continually added into the pharmacopeia. Because they depend on these methods of DDI discovery and anecdotal clinical reporting, current DDI databases do not include all of the potential DDIs. However, some of these interactions may be indirectly derived from the scientific literature [11] or drug-related documents [12] through informatics methods. Thus, a powerful literature-based discovery (LBD) tool that can extract DDI information from the biomedical literature has the potential to significantly enhance patient care.

In this paper, we propose a system rooted in natural language processing (NLP) that can find potential DDIs existing in the clinical data of an individual patient based on the knowledge transferred from the biomedical literature. Specifically, our system extracts patients' medication lists from clinical data, extracts all relevant semantic predications related to these medications from SemMedDB [13] (i.e., a database of semantic predications generated by SemRep [14]), and, thereby, suggests potential DDIs based on our DDI pathway schemas (i.e., *drug1* → *gene* → *drug2*, and *drug1* → *gene1* → *biological function* ← *gene2* ← *drug2*) and physician selection. Our methodology identifies potential patient-specific DDIs that are supported by evidence in the literature but are not contained in standard databases.

2. Related work

We propose using schemas describing relationships between drugs, genes, and physiological conditions to extract relationships from the literature that reflect DDIs. Previous investigation has explored methods of identifying these types of relationships. Weeber et al. developed a tool to systematically analyze online literature, which uses concept co-occurrence frequency coupled with expert review to identify promising "pathways" (or schemas) between a drug and a potential new target disease [15]. Wren et al. created three-concept drug–gene–disease schemas of co-occurring concept pairs in MEDLINE records with overlapping gene terms that serve as intermediates in an implicit relationship between drug and disease [16]. Frijters et al. have also reported CoPub, which couples relationships determined by co-occurrence of biomedical keywords in literature, to predict new relationships between genes, drugs, pathways and diseases. They validated several predicted relationships by using either independent literature sources or biological experiments [17].

Several investigators have extracted DDIs using NLP [11,18–21]. Most have focused on a specific set of genes, especially cytochrome P450s [21,22], or a focused set of drugs [21]. Percha et al. similarly predicted novel DDIs through the drug–gene–drug relationships by using text mining techniques on MEDLINE abstracts, though these were not applied to clinical data as in our method [11]. Their effort is shown to be effective on a limited set of 731 genes, heavily enriched in P450s known to be involved in drug metabolism. In the process of relationship extraction, their consideration of verbs and nominalized verbs as the sole relationship candidates for drug and gene entities misses relationships that can be reported using less explicit language. They make note of the limitation of using a constrained set of genes and not capturing gene–gene intermediate interactions or biological functions such as those that we have incorporated [11].

Similarly, Duke et al. combined cytochrome P450-based potential DDIs from the biomedical literature with EHR data to identify DDIs that might increase the risk of myopathy [18]. Their approach was even more focused since the literature mining was restricted to a group of P450s. The investigators did, however, combine their results with clinical data and were able to find 5 drug pairs with significant relative risks, thereby demonstrating some of the potential impact of their methodology. See Uzuner et al. [23] for a discussion of the detection of semantic relations between medical concepts within the context of the i2b2-2010 Challenges.

Our methodology builds on prior approaches by integrating literature-derived interactions not only between drugs and genes but also between genes and biological functions and by using actual medication combinations occurring in clinical data. This together specifies the potential interactions to individual patients while allowing for more complex interactions through multiple genes and pathways involved in biological functions.

3. Background

This study relied on several publicly available NLP tools that have been developed at the National Library of Medicine (NLM) including MetaMap and SemRep.

3.1. MetaMap

MetaMap [24] is an NLP system that maps biomedical text to concepts in the Unified Medical Language System (UMLS) Metathesaurus [25]. MetaMap processes input text using a series of lexical/syntactic analyses, followed by variant generation, candidate identification, mapping construction, and word sense disambiguation. It provides multiple processing options that allow users to choose vocabularies and the data model, control the algorithms, and select the output formats. MetaMap has been widely used for many applications including information retrieval [26], relation extraction [27], text mining [22], question answering [28], and knowledge discovery [29]. In this study, we use MetaMap to map medication lists extracted from clinical data to UMLS concepts for further information extraction.

Please cite this article in press as: Zhang R et al. Using semantic predications to uncover drug–drug interactions in clinical data. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.004