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

Rui Zhang, Michael J. Cairelli, Marcelo Fiszman, Graciela Rosemblat, Halil Kilicoglu, Thomas C. Rindflesch, Serguei V. Pakhomov, Genevieve B. Melton

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.

© 2014 Published by Elsevier Inc.

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 is 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.

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
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 pharmacopeiidiom 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 relationships 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 key-words 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
3.2. SemRep and SemMedDB

SemRep is a semantic interpreter also developed at NLM. SemRep relies on the UMLS SPECIALIST Lexicon [30] and MedPost part-of-speech tagger [31] for lexical and underspecified syntactic analysis and MetaMap to access domain knowledge in the UMLS Metathesaurus and extracts semantic relationships from the biomedical literature in the form of semantic predications. To enhance SemRep, work has been done to link to other structured resources, such as Entrez Gene [32], to recognize and normalize protein/gene names.

Each semantic predication (i.e., a subject-PREDICATE-object triple) consists of UMLS Metathesaurus concepts as arguments (i.e., subject and object) and a semantic relationship from an extended version of the UMLS Semantic Network as a predicate. The predications cover clinical medicine (e.g., TREATS, DIAGNOSES), substance interaction (e.g., INTERACTS_WITH, STIMULATES), genetic etiology of disease (e.g., ASSOCIATED_WITH, CAUSES), and pharmacogenomics (e.g., AFFECTS, AUGMENTS) [33]. For example, SemRep interprets the biomedical text in (1) as semantic predications in (2):

(1) Treatment with fenofibrate for 24 h significantly increased the expression of leptin and TSHr genes. (PMID: 15291748).

(2) Fenofibrate STIMULATES Leptin, Fenofibrate STIMULATES Thyrotropin Receptor.

Note that TSHr is normalized to the UMLS Metathesaurus concept Thyrotropin Receptor, and the verb “increased” to the predication STIMULATES.

The SemMedDB database [13] used for this study was generated by SemRep from all MEDLINE citations published as of June 30, 2012. The database contains over 21 million citations and 119 million sentences. About 57.6 million predication instances (12.9 million unique predications involving 58 predicates) were extracted from 37.2 million MEDLINE sentences. The database maintains links from each predication to its original sentence and PubMed entry through its PMID. It also contains the positional information regarding arguments and predications in a given sentence.

4. Methods

Our approach (Fig. 1A) included six basic components: (1) extracting the personal medication list from clinical data and mapping to UMLS using MetaMap, (2) extracting all semantic predications relevant to these medications and the genes and biological functions that they affect from SemMedDB, (3) normalizing gene names to approved gene symbols, (4) discovering all possible DDIs based on combinations of semantic predications according to pathway schemas, (5) providing potential unknown DDIs after human review and exclusion of known DDIs, (6) evaluating semantic predications. These components are achieved through a series of steps detailed below.

Step 1: Medication list extraction. We randomly picked patient records from the Epic™ EHR system from 22 patients (13 females and 9 males) with outpatient clinical visits at Fairview Health Services (the integrated health system affiliated with the University of Minnesota Medical Center). The average age of patients was 63 with men slightly older than women, as shown in Fig. 2. Most patients had chronic disorders (e.g., diabetes). Aspirin was the most commonly used medication in this cohort. Additional disorders and medication are depicted and ranked in Fig. 3. The list of drugs was extracted from the EHR clinical data of each individual patient. Each drug was mapped to the appropriate concept in the UMLS Metathesaurus using MetaMap. The list of generic drug names for each office visit of an individual patient was used as the input to the DDI discovery system. A total of 224 unique drugs for all 22 patients were included in this study. Only interactions between two drugs with co-occurrence in the same patient record were examined.

Step 2: Predication extraction from SemMedDB (Fig. 1C). We extracted four types of predications from SemMedDB: drug–gene (i.e., predications with a drug as the subject and a gene as the object), gene–drug, gene–function, and drug–function. We extracted all predications describing an influence between a drug from the extracted medication list (Step 1) and a gene. Specifically, predications having a drug as the subject, a gene (including gene product or protein) as the object and three restricted predicate types (INHIBITS, INTERACTS_WITH, and STIMULATES) were extracted as drug–gene predications. The inverse of these, representing genes having an influence on drugs, with gene as the subject, drug as the object, and the same three predicate types were extracted as gene–drug predications. To identify genes influencing biological functions, we extracted gene–function predications with the genes as subjects, functions as objects, and predicate types including AFFECTS, AUGMENT, CAUSES, DISRUPTS, INHIBITS, and PREDIPOSES. Finally, to capture medications with an effect on biological functions, we generated drug–function predications having a drug as the subject and a function as the object. By function we are specifically referring to all descendants of the Biological Function semantic type in the UMLS semantic type hierarchy, as shown in Fig. 4.

Step 3: Gene name normalization. We used the downloadable Gene Nomenclature Committee dataset from HUGO [34] (the organization providing the international standard for gene names and symbols) as a dictionary to translate all gene and protein names to approved gene symbols (e.g., Sex Hormone-Binding Globulin to SHBG). 139 exact match genes were found in the gene–function predications from Step 2. Non-matches were usually non-specific terms such as classes of genes or proteins.

Step 4: DDI discovery pathways (Fig. 1B).

(i) Drug1 → Gene → Drug2 (DGD) pathway (Fig. 1B-1). We identified the potential DDIs between each pair of drugs in the medication list from clinical data for an individual patient using the drug–gene and gene–drug predications previously extracted in Step 2. Potential DDI candidates were generated by matching the object gene in a drug–gene predication with the subject gene in a gene–drug predication, requiring a different drug in each of the predications.

(ii) Drug1 → Biological Function → Gene2 → Drug2 (DGFGD) pathway (Fig. 1B-2). We also identified the potential DDIs between each pair of drugs in the drug list using the drug–gene, gene–function, and drug–function predications. Potential DDI candidates were generated when the following constraints were satisfied:

(1) There exists a set of predications satisfying Drug1 → Gene1 → Biological Function → Gene2 → Drug2.

(2) There exists no direct Drug1 → Biological Function or Drug2 → Biological Function predications for this function.

This ensures that both drugs have an independent effect on the function and the function is not an established effect of either drug.

Step 5: Physician selection of salient predications and exclusion of known DDIs. We first retrieved the MEDLINE sentences that produced DDI candidate pairs based on DGD and DGFGD pathways from SemMedDB and contained drug pairs found in the patient medication lists. One of the authors (MJC, a physician) then selected the most promising candidates from the chains of predications which matched each of the schemas, considering the validity of the component predications relative to their source sen-
sentence, method of administration, location of administration for drugs with low systemic distribution, and similar concerns. The initial set consisted of 1,029,271 chains from which 2839 were randomly selected and shown to the physician for initial review. From these, 111 chains were selected as interesting. Characteristics of interesting chains included common usage, systemic penetration, and specificity (e.g., Ascorbic Acid over Vitamins). Gene products with wide distribution systemically or in multiple pathways were preferred (e.g., hormones over immune receptors). For biological functions, specificity was preferred over general groups of disorders (e.g., Mood Disorders, Cardiomyopathies). The predications in the 111 most interesting chains were provided to the physician with the source sentence for each and evaluated again in the context of the sentence. At this stage the review was focused on the accuracy of the predications. Additionally, whether the context was exceptional and unlikely to be relevant to real patients (versus cell culture or model organism) was also considered. This resulted in a selection of 54 predications and we then identified all of the

Fig. 1. (A) Overview of DDI discovery system. (B) DDI discovery methodology for two drugs (details depicted in Step 5). Blue lines indicate allowed interactions. Red lines are prohibited. Purple double-arrowed line indicates the potential DDI. (C) Semantic predication extraction process from SemMedDB. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2. (A) Histogram and (B) boxplot of age ranges for female and male patients.
chains matching the schemas that were composed solely of these selected predications. Each candidate DDI was also checked against Drugs.com \[36\] to identify known DDIs and those found in their database were discarded.

**Step 6: Evaluation of predications.** Although the physician’s selection in Step 5 included validation of the predications against the source sentences as part of the filtering process, we also performed a separate linguistic evaluation of the predication set. For this evaluation, we selected 300 sentences that assert substance interactions – including drug–gene and gene–drug interactions – (200 sentences) or drug–biological function relations (100 sentences) and manually annotated them with relevant semantic predications. The annotation process consisted of two steps.

In the first step, three authors of this paper (GR, HK, MF) annotated 100 sentences (60 substance interactions, 40 drug–function relations) randomly selected from the 300-sentence set for predications that involve relevant predicates. The relevant predicates are STIMULATES, INHIBITS, INTERACTS_WITH for substance interactions and AUGMENTS, DISRUPTS, AFFECTS, CAUSES, and PREDISPOSES for drug–function relations. The annotation process followed the guidelines developed in a previous annotation study \[35\]. The main difference was that the annotators were asked to limit themselves to 2006AA UMLS concepts, since this UMLS release underlies the SemMedDB database. The gene/gene product equivalence criterion from that study was also followed. This first step allowed us to compute interannotator agreement to determine the reliability of the annotations. We adopted the F-measure.
In the second step, based on the same guidelines and the discussion among pairs of annotators as the interannotator agreement measure. In a previous study [35], we had found that agreement among annotators was relatively low for the aforementioned predicates and we also assessed whether annotation experience gained in that study helped with interannotator agreement.

In the second step, based on the same guidelines and the discussion among pairs of annotators, one of the authors (GR) annotated the remaining 200 sentences and adjudicated 100 sentences from the previous step. This set of annotations (all 300 sentences) was taken as the gold standard for the linguistic evaluation.

In assessing SemRep performance, we used standard information extraction evaluation metrics of precision, recall, and F-measure for this purpose. We also calculated precision/recall curves by using argument distance of predications as the inclusion criterion. We limited the precision/recall curve calculation to verbal predicates only, since argument distance is most meaningful with such predicates due to SemRep’s underspecified argument identification algorithm.

**5. Results**

### 5.1. DDIs discovered through Drug1 → Gene → Drug2 (DGD) pathway schema

Using the DGD pathway schema (Step 4i), we found 14 unknown pairs of potentially interacting drugs (Table 1) in medication list of clinical data after physician selection of interesting predications (Step 5). Three DDI examples with corresponding citations are provided in Fig. 5. The sentences that were used to generate three chains are listed in Table 2. The underlined words in sentences are extracted as subjects and objects in the predications. Highlighted words in the sentences indicate the relationships (predicates) between two biomedical concepts. In the first example, Lisinopril STIMULATES Vasoactive Intestinal Peptide (VIP), which in turn STIMULATES Thyroxine (both an endogenous hormone and given as a supplement for hypothyroidism). This chain indicates a potential DDI between lisinopril and thyroxine supplementation, suggesting that co-administration could lead to iatrogenic hyperthyroidism.

### 5.2. DDIs discovered through Drug1 → Gene1 → Biological Function — Gene2 → Drug2 (DGFGD) pathway schema

Applying the DGFGD pathway schema (Step 4ii) to our predications set and the subsequent physician selection of 300 semantic predications (Step 5) yielded 48 unknown DDIs (Table 3). A selection of three pathways is detailed in Fig. 6. When several drugs are the subjects of an otherwise consistent interaction, they are grouped together for clarity. In the left wing of the first pathway, Metformin STIMULATES gene Sex Hormone-Binding Globulin (SHBG), which CAUSES the function Benign Prostatic Hypertrophy (BPH). In the right wing of the chain, Cholecalciferol, Clonidine, and Sertraline each STIMULATES gene PRL which CAUSES the same biological function BPH. No interactions between Metformin and Cholecalciferol, Clonidine, Thyroxine, or Sertraline were reported in the DDI database, indicating four potential pairs of drug interactions as well as a compound effect by combining more than two of these medications. We also found examples of potential DDIs occurring through different pathways. For example, (Table 3, No. 33–38), Thyroxine can either INTERACTS_WITH Epidermal Growth Factor (EGF) which AUGMENTS Adipogenesis, or STIMULATES Prolactin (PRL) which CAUSES Adipogenesis. Each pair of drugs appeared in the same patient medication list. Some examples of the sentences and their extracted predications that generated these pathways are shown in Table 4. As in Table 2, the underlined words in sentences are extracted as subjects and objects in the predications and highlighted words in the sentences indicate the relationships (predicates) between two biomedical concepts.

### 5.3. DDIs discovered through different pathway schema

By browsing the generated DDI results, we found five drug interactions that were included in both pathway schemas. For example, the interaction between Lisinopril and Thyroxine was
found through DGD pathway (Table 1, No. 6) and DGF pathway (Table 3, No. 36). Others include Aspirin and Metformin (Table 1, No. 9; Table 3, No. 15). Aspirin and Thyroxine (Table 1, No. 4; Table 3, No. 33), Clonidine and Thyroxine (Table 1, No. 14; Table 3, No. 35), and Misoprostol and Thyroxine (Table 1, No. 11; Table 3, No. 37). The DGF pathway DDIs either provided further detail to the DGD pathway version (i.e., they contain the same gene in both wings) or they provided additional mechanisms of interaction (i.e., different genes in each wing).

5.4. Discovery of known DDIs

We checked all resulting drug interactions against the Drugs.com DD database [36]. Three of the drug interactions discovered by DGD pathway schema were found in this database and 16 of the DDIs discovered by DGF pathway schema matched interactions in the database. All of the drug pairs identified as ‘known’ and their corresponding degree of severity are listed in Table 5.

5.5. Annotation and evaluation of semantic predications

The number of predications annotated in the first phase of the annotation study and the agreement among annotators are provided in Table 6. The results show that an average of 2.22 predications were annotated per sentence, with a higher average (2.39) for substance interactions than that for drug–function relations (1.98). Agreement among annotators ranges from 0.650 to 0.753, with higher agreement for substance interactions than for drug–function relations (see Table 7). The distribution of the number of predications in the gold standard reference (by theme and by predicate type) is given in Tables 8 and 9, respectively.

Evaluation was conducted on 300 sentences, from which SemRep extracted 524 relevant predications. Overall, 304 of these predications were deemed true positives, 220 false positives, and 384 false negatives, yielding precision of 0.580, recall of 0.442, and F-measure of 0.502. The evaluation metrics varied between the predicate types: substance interactions (precision 0.585, recall 0.451, F-measure 0.509); drug–function interactions (precision 0.646, recall 0.420, F-measure 0.509). Additionally, variance in performance was noted between nominal and verbal predications. For nominal predications, SemRep precision was 0.574, recall 0.483, making F-measure 0.524. The F-measure for verbal predications was lower at 0.473 with a slightly higher precision of 0.592 but much lower recall of 0.393.

A variance in performance of verbal predications by argument distance was also observed. A plot of precision, recall, and F-measure metrics for substance interactions and drug–function relationships are shown in Figs. 7 and 9, respectively. Precision-recall curves for substance interactions and drug–function relationships are provided in Figs. 8 and 10, respectively.

6. Discussion

Our method of identifying drug–drug interactions from the biomedical literature is novel in two ways: (a) it makes use of the knowledge from the entire MEDLINE database (via semantic predications extracted by SemRep); and (b) it uses biological functions, as defined by the UMLS, in the definition of drug effects to allow for the detection of interactions that would not be noticed when considering only those linked through a single gene or protein. The use of SemMedDB predications expands the range of genes and drugs considered to all drugs contained in the UMLS and all genes and proteins within either the UMLS or Entrez Gene. The predications in SemMedDB are extracted from the complete set of biomedical citations contained in MEDLINE, and they are based on actual assertions in the text, not co-occurrence or similarity functions. The wide-open approach to drug list, gene list and drug interaction target offers more flexibility in finding potential interactions, while in its current incarnation our system can produce significant results even using a small clinical data set as done in this particular study.

6.1. Significance of results

An interesting example from the DGD pathways in Fig. 5 is “Lisinopril STIMULATES VIP/VIP STIMULATES Thyroxine”. The source citation for the first predication (PMID: 2822521) asserts that lisinopril is found to induce increased plasma levels of vasoactive intestinal peptide (VIP). Although the stimulation of thyroxine in the second predication is referring to increased endogenous production (PMID: 2829144), clearly an increase in endogenous production could precipitate unwanted increases in plasma levels of thyroid hormone and repercussions of hyperthyroidism without adjustment of dosage administration.

The results shown in Fig. 6 for the DGF pathway include “Lisinopril SIMULATES VIP/VIP CAUSES Psoriasis” coupled with “[Clofibrate SIMULATES Prolactin, Cholecalciferol SIMULATES Prolactin, Sertraline SIMULATES Prolactin, Thyroxine SIMULATES Prolactin]/Prolactin CAUSES Psoriasis”. This demonstrates multiple drug–gene–function combinations having an effect on psoriasis. Their respective source citations (PMID: 19350575, 1372339) suggest that both VIP and prolactin are implicated in the pathogenesis of psoriasis. As noted above, lisinopril is noted to increase VIP (PMID: 2829144).
6.2. DDI network visualization

The dramatic reduction in network density that resulted from physician selection of significant relationships/predications is demonstrated in Fig. 11. We used Cytoscape [37] to visualize a network of all DDIs obtained through the DGFGD pathway (box A). We also produced a focused DDI network after human review and known DDIs exclusion (box B).

Before human review, the network of potential DDIs through the DGFGD pathway contains 620 nodes (including 192 drugs, 137 genes, and 291 biological functions) and 5293 edges. The refined network, with 47 nodes (including 25 drugs, 11 genes and 11 biological functions) and 54 edges, is composed from the set of salient predications resulting from the physician selection process (Step 6). Details of this network provided in Fig. 11 (box C) highlight some relationships provided as examples in Fig. 6.

6.3. Use of curated databases to validate drug interactions

It is common practice to use curated databases such as Drugs.com [36], DrugBank [38,39], or PharmGKB [40] as a gold standard.
6.4. Advantages of SemMedDB predications to find unknown DDIs

While interaction databases contain detailed information on gene and drugs based on published data, because these data are curated, the number of reviewed documents is subject to human limitations. SemMedDB contains not only interactions from the original documents reviewed in the generation of the databases but also millions of other documents not included in their formal review. This provides access to a wealth of knowledge that goes beyond the curated databases. The only limitation on which genes are included is whether they are contained in either the UMLS or Entrez Gene, which is used by SemRep as a supplementary resource for gene and protein terms. SemMedDB also contains predications regarding biological functions specifically linked to corresponding genes, thereby providing an additional framework for interactions beyond the traditional paradigm. Finally, SemRep considers a wider range of linguistic structures in identifying relationships; we take into account not only verbs and nominalizations, but also a wider range of linguistic structures in identifying relationships; we take into account not only verbs and nominalizations, but also a wider range of linguistic structures in identifying relationships.

6.5. Clinical usage: providing clinical information rather than giving alerts

Due to the theoretical nature of the extracted DDIs, this system is not intended to replace current approaches to DDI alert systems.
There is a significant potential of inducing alert fatigue if the approach was to suggest any possible effect that may occur from all of the drug combinations in a patient’s medication list. Instead, we envision this system to suggest potential causes for unexplained patient symptoms or unusual responses to treatment. That is, by matching reported symptoms in the note or a search query to potential drug interactions in the same clinical record, we would be able to provide patient-specific, physician-sought information as opposed to an intrusive alert system.

Another potential application for this methodology is to facilitate curated database development. This system provides theoretical DDIs that can be investigated clinically to ascertain incidence and severity. Also, this mechanism could be incorporated into the development of new drugs, providing additional predictive capability for potential interactions.

Last, but not least, this approach may be used as a hypothesis generating step in the broader context of identifying potentially harmful drug–drug interactions. Hypotheses generated with this approach can then be tested either in vivo (prospective patient observation) or in silico (prospective or retrospective examination of medical records) clinical studies.

### Table 5
Known drug–drug interactions found by our method.

<table>
<thead>
<tr>
<th>No.</th>
<th>Drug1</th>
<th>Drug2</th>
<th>Degree of interactions (provided by Drugs.com)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Drug1 → Gene1 → Drug2 (DG) pathway</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Metformin</td>
<td>Carvediol</td>
<td>Moderate</td>
</tr>
<tr>
<td>2</td>
<td>Exenatide</td>
<td>Aspirin</td>
<td>Moderate</td>
</tr>
<tr>
<td>3</td>
<td>Glimepiride</td>
<td>Aspirin</td>
<td>Moderate</td>
</tr>
<tr>
<td>II. Drug1 → Gene1 → Biological function → Gene2 → Drug2 (DGFGD) pathway</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Metoprolol</td>
<td>Aspirin</td>
<td>Minor</td>
</tr>
<tr>
<td>2</td>
<td>Thyroxine</td>
<td>Lovastatin</td>
<td>Minor</td>
</tr>
<tr>
<td>3</td>
<td>Thyroxine</td>
<td>Metoprolol</td>
<td>Minor</td>
</tr>
<tr>
<td>4</td>
<td>Thyroxine</td>
<td>Simvastatin</td>
<td>Minor</td>
</tr>
<tr>
<td>5</td>
<td>Vitamin D</td>
<td>Misoprostol</td>
<td>Minor</td>
</tr>
<tr>
<td>6</td>
<td>Insulin</td>
<td>Aspirin</td>
<td>Moderate</td>
</tr>
<tr>
<td>7</td>
<td>Insulin</td>
<td>Citralex</td>
<td>Moderate</td>
</tr>
<tr>
<td>8</td>
<td>Lisinopril</td>
<td>Glycerol</td>
<td>Moderate</td>
</tr>
<tr>
<td>9</td>
<td>Lisinopril</td>
<td>Insulin</td>
<td>Moderate</td>
</tr>
<tr>
<td>10</td>
<td>Lisinopril</td>
<td>Exenatide</td>
<td>Moderate</td>
</tr>
<tr>
<td>11</td>
<td>Metformin</td>
<td>Lisinopril</td>
<td>Moderate</td>
</tr>
<tr>
<td>12</td>
<td>Sertaline</td>
<td>Aspirin</td>
<td>Moderate</td>
</tr>
<tr>
<td>13</td>
<td>Thyroxine</td>
<td>Metformin</td>
<td>Moderate</td>
</tr>
<tr>
<td>14</td>
<td>Vitamin D</td>
<td>Thyroxine</td>
<td>Moderate</td>
</tr>
<tr>
<td>15</td>
<td>Exenatide</td>
<td>Aspirin</td>
<td>Moderate</td>
</tr>
<tr>
<td>16</td>
<td>Glimepiride</td>
<td>Aspirin</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

### Table 6
Overall predications statistics for 100 sentences annotated by all three annotators.

<table>
<thead>
<tr>
<th>Annotator</th>
<th># of predications (per sentence)</th>
<th>Substance int. (per sentence)</th>
<th>Drug–function (per sentence)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GR</td>
<td>236 (2.36)</td>
<td>156 (2.60)</td>
<td>80 (2.00)</td>
</tr>
<tr>
<td>HK</td>
<td>208 (2.08)</td>
<td>133 (2.22)</td>
<td>75 (1.88)</td>
</tr>
<tr>
<td>MF</td>
<td>223 (2.23)</td>
<td>141 (2.35)</td>
<td>82 (2.05)</td>
</tr>
<tr>
<td>Average</td>
<td>(2.22)</td>
<td>(2.39)</td>
<td>(1.98)</td>
</tr>
</tbody>
</table>

### Table 7
Interannotator agreement.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Overall</th>
<th>Substance int.</th>
<th>Drug–function</th>
</tr>
</thead>
<tbody>
<tr>
<td>GR-HK</td>
<td>0.753</td>
<td>0.776</td>
<td>0.711</td>
</tr>
<tr>
<td>GR-MF</td>
<td>0.688</td>
<td>0.734</td>
<td>0.58</td>
</tr>
<tr>
<td>HK-MF</td>
<td>0.65</td>
<td>0.671</td>
<td>0.611</td>
</tr>
</tbody>
</table>

### Table 8
Gold standard reference.

<table>
<thead>
<tr>
<th># Predications</th>
<th>Substance int. (per sentence)</th>
<th>Drug–function (per sentence)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall (2.00)</td>
<td>689 (2.30)</td>
<td>489 (2.45)</td>
</tr>
</tbody>
</table>

### Table 9
Distribution of predicate types in the gold standard reference.

<table>
<thead>
<tr>
<th>Substance interaction</th>
<th># of predications</th>
<th>Drug–function</th>
<th># of predications</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERACTS_WITH</td>
<td>270</td>
<td>AUGMENTS</td>
<td>78</td>
</tr>
<tr>
<td>STIMULATES</td>
<td>111</td>
<td>PREDISPOSES</td>
<td>44</td>
</tr>
<tr>
<td>INHIBITS</td>
<td>108</td>
<td>CAUSES</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AFFECTS</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DISRUPTS</td>
<td>11</td>
</tr>
</tbody>
</table>

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
6.6. SemRep analysis

6.6.1. Interannotator agreement analysis

Interannotator agreement rates in the range of 0.65–0.75 are considered acceptable, leading us to conclude that our annotations can be reliably used as reference for SemRep evaluation. The annotator responsible for adjudication and further annotation (of 200 sentences) had higher agreement with the other two annotators than those annotators had between themselves; therefore, it is also reasonable to say that the final annotations reflect the middle ground between all annotators.

At the predicate level, interannotator agreement rates are significantly higher than those obtained in our previous annotation study [35]. In that study, the agreement for all predicate types – including but not limited to those discussed in this paper – was found to be 0.536 between annotators GR and HK. The rate of agreement between the same two annotators is 0.753 in the current study. Even more interestingly, in the previous study, predicate types with lowest agreement were DISRUPTS (0.214), STIMULATES (0.238), and AFFECTS (0.308), all of which were considered in the current paper. We found that the agreement for these predicate types increased to 0.67, 0.80 and 0.60, respectively, in the current study. These results indicate that annotation experience leads to more consistent annotations over time. The fact that the current study was limited to a subset of the predicate types could have contributed to higher interannotator agreement as well; however, our study does not provide conclusive evidence for this.

6.6.2. SemRep errors

Analysis of SemRep output for this evaluation set indicated that errors generally fell into two broad categories: (a) shortcomings in the knowledge sources that SemRep depends on and (b) shortcomings in SemRep’s processing of linguistic phenomena.

A significant number of false positive errors were due to incorrect mappings of gene/protein mentions to UMLS concepts or Entrez Gene terms. The two most frequent mapping errors involved Ca2+, the calcium ion, mapping to the gene CA2 (n = 12) and LDL-C, low-density lipoprotein cholesterol, mapping to the gene COG2 (n = 7).

Incorrect argument identification constituted another source of false positive errors or erroneous predications. Failure to identify the correct arguments was due either to missing concepts in the UMLS or to SemRep processing errors, as in the following sentence (3) and semantic predication (4), in which ‘Hypoglycaemia,’ is the subject of AUGMENTS predicate, rather than Insulin:

(3) Hypoglycaemia induced by insulin increased catecholamine secretion, with the adrenaline to noradrenaline ratio significantly higher than in the adrenal gland itself. (PMID: 5152027).

(4) Insulin AUGMENTS catecholamine secretion.

Two major classes of linguistic phenomena addressed inadequately by SemRep are negation and serial coordination. In the following example, SemRep fails to recognize negation in sentence (5) and produces predication (6) that asserts its polar opposite:

(5) The Val66Met polymorphism of the brain-derived neurotrophic factor gene is not associated with risk for schizophrenia and tardive dyskinesia in Han Chinese population. (PMID: 20395113).

(6) Brain-derived neurotrophic factor PREDISPOSES Schizophrenia.

On the other hand, from sentence (7) SemRep extracts an erroneous predication (8) because it is unable to recognize that both subject and object arguments (Ro 25-1553 and vasoactive intestinal peptide (VIP) respectively) are coordinated elements within the same coordination structure.

(7) Studies were conducted to compare the effect of native vasoactive intestinal peptide (VIP), Ro 25-1553 (a cyclic peptide analog of VIP) and salbutamol (a beta2-adrenoceptor agonist) on antigen-induced pathophysiological effects in the guinea pig. (PMID: 7932181).

(8) Ro 25-1553 INTERACTS_WITH Vasoactive Intestinal Peptide.
SemRep produces a predication (10) that may be true but it is not fully supported in the title.

(9) Melanoma risk in association with serum leptin levels and lifestyle parameters: a case-control study (PMID: 17925285).

(10) Leptin PREDISPOSES Melanoma.

We identified this as a false positive because there is no explicit assertion in (9) that there is indeed increased melanoma risk with increased leptin levels. There may be no difference in risk, the risk may be increased or decreased, or the leptin level that is associated with greater risk may be an increase or decrease from normal.

We note that this rejection of inference may not be standard in similar work. For instance, the relationship (12) extracted from the sentence (11) by the system reported in Percha et al. [11] would not be considered a true positive by our system:

(11) How atorvastatin could limit the pro-inflammatory response to thrombin was studied in cultured rat aortic smooth muscle cells. (PMID: 12921859).

(12) Atorvastatin DECREASES F2 (i.e., the gene for Thrombin).

We would reject this assertion because there is no direct relationship asserted between atorvastatin and thrombin but, instead, to an inflammatory response to thrombin that does not necessitate that expression of the F2 gene or the activity of thrombin be decreased. Although such relations may be valid based on additional context from other sentences in the abstract or given prior knowledge, the source sentence itself is not sufficient to support this relationship.

6.6.3. Impact of argument distance

Although SemRep has been shown to perform well with pharmacogenomic relations [33], previous investigation has shown some indication that SemRep precision in predications related to genes can be improved by using only predications with low argument distance [44]. Argument distance is defined as the number of noun phrases between the predicate and the noun phrase selected as the argument of the predicate (subject or object).

When we applied the same type of investigation using our gold standard reference, we found surprisingly different results. For substance interactions, although precision did increase when the maximum argument distance decreased from any size to 3, at smaller distances the trend was reversed and the lowest precision was noted at a distance of 1. Recall was much as expected, steadily decreasing as the distance threshold decreased. For drug–function predications performance was more in line with expectations as precision increased at distance thresholds below 4 and recall decreased generally as maximum distance was reduced. A surprising result of our analysis was that object distance had much more significance than subject distance. This is demonstrated in Fig. 12.

When subject distance is greater than 1, $F$-measure is generally consistent across all subject distance thresholds but steps up for each increase in object distance.

The increase in $F$-measure (and recall without a decline in precision) in high distances may seem counterintuitive. SemRep seems able to identify some of the long range dependencies between the indicators and the arguments due to its underspecified approach, leading to a precision increase at higher argument distances. Long range dependencies are a well-known feature of biomedical literature [43] and have, in fact, led to recent popularity of dependency parsing based approaches to relation extraction from biomedical literature.
6.7. Limitations and future work (application to larger clinical data set)

Our method currently depends on expert review of automatically asserted predictions. This is due in part to a low level of precision in this knowledge domain and expansion of SemRep capabilities continues to be a target for improvement, but at the same time it conveys a benefit in that it allows the expert to determine which types of interactions are most relevant to the patients in their care. After a semantic predication has been selected as interesting, it can be stored and it need not be judged again for the same user/care group combination, thereby allowing its incorporation into any additional relevant pathways.

Because this approach combines two separate assertions linked through a gene or biological function and does not search for stated drug–drug interactions directly, it produces theoretical DDIs that may not have been recognized before nor verified through clinical trials. Application of this methodology to a larger clinical dataset would not only allow for access to a greater range of theoretical DDIs, but also allow biological functions from the proposed interactions to be compared against presence of signs and symptoms in the clinical record to validate and determine incidence of the DDIs. Web-based healthcare data has been explored to discover interactions and could also be combined with our methodology. For instance, White et al. used web search log data to identify hyperglycemia-related terms combined with paroxetine and/or pravastatin in searches on Google, Bing, and Yahoo! [45].

This method could also gain breadth by considering pharmacogenomics to personalize the effects of drugs for the genetic variants of specific patients. Rance et al. [46] describe a methodology for extracting drug–gene interactions specific to genetic variations that could be incorporated to extend the current results to variant-specific interactions. Though current clinical records may be rather sparse in regard to genetic variant assertions, current trends suggest this information is only likely to increase in prevalence.

Step 5 of our methodology requires the manual selection of candidate chains and their component predications. The use of such human intervention is limited not only by capacity (in our case less than 0.28% of potential schemas were actually reviewed), but also by inconsistencies between individual participants. In expanding this approach to a larger data set, this component may need to be replaced by a more automated approach using machine learning techniques. The chains and predications that have already been identified as interesting could serve as training data which might be augmented or supplanted by known DDI data.

7. Conclusion

We present a new methodology for detecting DDIs in clinical data that exploits semantic predications. We expand the search for interactions beyond the traditional paradigm by considering drug effects at the level of biological functions. Because SemRep extracts these from the over 20 million citations in MEDLINE, the search is not limited to a predetermined drug and gene set and has the potential to include any drug or gene contained in the UMLS or Entrez Gene. Several enhancements to SemRep, which can increase the potential of the DDI discovery system, are planned. These include extracting the degree of confidence expressed for a statement in text and the context in which the drug interactions appear.

Ethics statement

The patient data used for this project were de-identified and its use had IRB approval from the University of Minnesota.

Acknowledgments

This research was supported in part by an appointment to the NLM Research Participation Program, which is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine. This project was supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

References


HUGO. <http://www.genenames.org/>.


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