Journal of Biomedical Informatics 45 (2012) 870–878

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Systematic identification of pharmacogenomics information from clinical trials

Jiao Li, Zhiyong Lu *

National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, United States

ARTICLE INFO

Article history:
Received 29 April 2011
Accepted 11 April 2012
Available online 24 April 2012

Keywords:
Text mining
Clinical outcome
Pharmacogenomics
Clinical trial

ABSTRACT

Recent progress in high-throughput genomic technologies has shifted pharmacogenomic research from candidate gene pharmacogenetics to clinical pharmacogenomics (PGx). Many clinical related questions may be asked such as 'what drug should be prescribed for a patient with mutant alleles?' Typically, answers to such questions can be found in publications mentioning the relationships of the gene–drug–disease of interest. In this work, we hypothesize that ClinicalTrials.gov is a comparable source rich in PGx related information. In this regard, we developed a systematic approach to automatically identify PGx relationships between genes, drugs and diseases from trial records in ClinicalTrials.gov. In our evaluation, we found that our extracted relationships overlap significantly with the curated factual knowledge through the literature in a PGx database and that most relationships appear on average 5 years earlier in clinical trials than in their corresponding publications, suggesting that clinical trials may be valuable for both validating known and capturing new PGx related information in a more timely manner. Furthermore, two human reviewers judged a portion of computer-generated relationships and found an overall accuracy of 74% for our text-mining approach. This work has practical implications in enriching our existing knowledge on PGx gene–drug–disease relationships as well as suggesting crosslinks between ClinicalTrials.gov and other PGx knowledge bases.

1. Introduction

Clinical outcomes in response to drugs can be significantly different among individuals, in terms of treatment efficacy and drug toxicity. Although many clinical variables of individuals (*e.g.*, liver function, disease severity, and drug interactions) potentially cause the variability of drug effects, it is now recognized that genetic polymorphisms can have an even greater influence on drug efficacy and safety [1]. Pharmacogenomics (PGx) studies elucidate the inherited nature of variability of drug effects in the context of genomics. Recent progress in high-throughput genomic technologies has significantly enhanced the identification of genetic variations associated with drug absorption, distribution, metabolism, excretion, and target action. The consequent explosion of data has raised challenges in PGx data description, storage, and integration. Meanwhile, pharmacogenomics impacts at many stages along the drug discovery and development pipeline, from target identification in early-stage research to post-marketing surveillance in phase IV clinical trials [2]. The consequent diversity of data types has raised challenges to capture more attributes of genotype and phenotype data as well as more complex relationships between them.

PharmGKB [3], the pharmacogenomics knowledge base, is a widely used resource in PGx studies. It collects PGx related genotype and phenotype data with manually annotated relationships between genes, polymorphisms, drugs, and diseases, as well as provides summarized information on important PGx genes and drug pathways. Over the last decade, PharmGKB has collected and annotated PGx data from a variety of different sources but the scientific literature remains its major source [4]. PharmGKB has developed structures to tag and describe relationships with PGx categories: clinical outcome, pharmacodynamics and drug responses, pharmacokinetics, molecular and cellular functional assays, and genotype [5]. Recent PGx text mining efforts have mainly focused on automatically extracting these relationships from the scientific literature [4,6].

Meanwhile, PGx research has shifted from candidate gene pharmacogenetics to clinical pharmacogenomics [4]. To investigate the clinical applications of PGx studies clinical trials are designed and conducted. In the context of drug development, clinical trials are necessary steps to determine if a drug is safe and effective (see Fig. 1). It usually takes many years for a new drug to pass through Phase I, II, and III before it is approved by the national regulatory authority (*e.g.*, the Food and Drug Administration in the United States). In each phase of clinical trials, studies are separately conducted and approved. In the past, researchers have investigated

Abbreviations: PGx, pharmacogenomics; PharmGKB, pharmacogenomics knowledge base.

* Corresponding author.

E-mail address: zhiyong.lu@nih.gov (Z. Lu).

1532-0464 Published by Elsevier Inc. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.jbi.2012.04.005

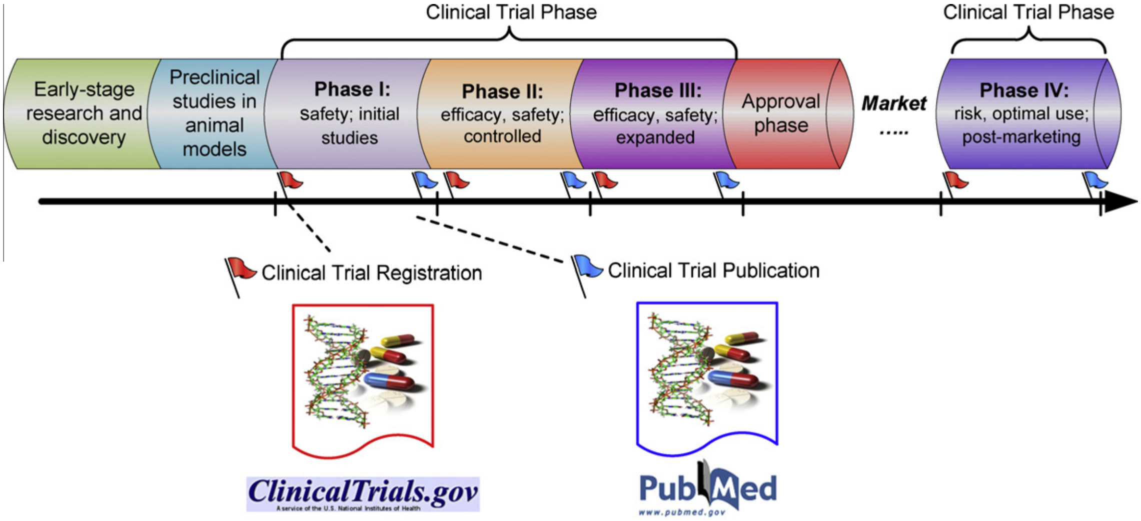
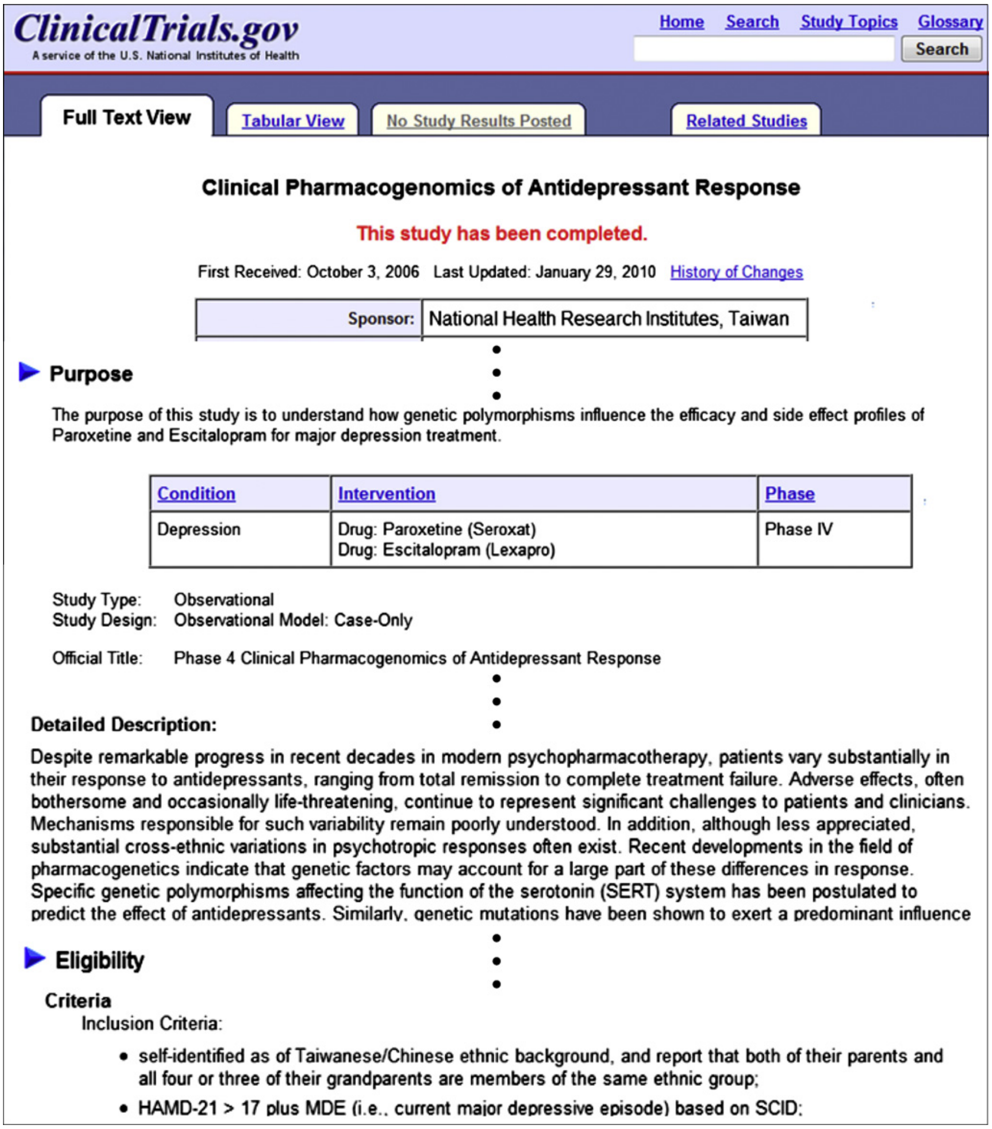

Clinical Trial Phase
Clinical Trial Phase
Early-stage research and discovery
Preclinical studies in animal models
Phase I: safety; initial studies
Phase II: efficacy, safety; controlled
Phase III: efficacy, safety; expanded
Approval phase
Market .....
Phase IV: risk, optimal use; post-marketing
Clinical Trial Registration
Clinical Trial Publication
ClinicalTrials.gov
A service of the U.S. National Institutes of Health
PubMed
www.pubmed.gov

Fig. 1. Clinical trial registration and publication along the drug development pipeline.



Clinical Pharmacogenomics of Antidepressant Response

This study has been completed.

First Received: October 3, 2006 Last Updated: January 29, 2010 History of Changes

Sponsor:	National Health Research Institutes, Taiwan

•
•
•

Purpose

The purpose of this study is to understand how genetic polymorphisms influence the efficacy and side effect profiles of Paroxetine and Escitalopram for major depression treatment.

Condition	Intervention	Phase
Depression	Drug: Paroxetine (Seroxat) Drug: Escitalopram (Lexapro)	Phase IV

Study Type: Observational
Study Design: Observational Model: Case-Only

Official Title: Phase 4 Clinical Pharmacogenomics of Antidepressant Response

•
•
•

Detailed Description:

Despite remarkable progress in recent decades in modern psychopharmacotherapy, patients vary substantially in their response to antidepressants, ranging from total remission to complete treatment failure. Adverse effects, often bothersome and occasionally life-threatening, continue to represent significant challenges to patients and clinicians. Mechanisms responsible for such variability remain poorly understood. In addition, although less appreciated, substantial cross-ethnic variations in psychotropic responses often exist. Recent developments in the field of pharmacogenetics indicate that genetic factors may account for a large part of these differences in response. Specific genetic polymorphisms affecting the function of the serotonin (SERT) system has been postulated to predict the effect of antidepressants. Similarly, genetic mutations have been shown to exert a predominant influence

•
•
•

Eligibility

Criteria

Inclusion Criteria:

- self-identified as of Taiwanese/Chinese ethnic background, and report that both of their parents and all four or three of their grandparents are members of the same ethnic group;
- HAMD-21 > 17 plus MDE (i.e., current major depressive episode) based on SCID;

Fig. 2. Snapshot for a clinical trial record in ClinicalTrials.gov.

the persistent gap between the number of trials conducted and the number for which results are published. They found that up to 37% of clinical trials never resulted in a scientific publication, and that the published articles reporting trial outcomes may not be consistent with protocols [7]. Therefore, in order to promote the transparency of clinical trials, several policy recommendations and regulations have been created for trial registration when a trial is launched [8]. For instance, in the United States, the Food and Drug Administration Modernization Act [9] requires all trials for drugs for serious or life-threatening diseases and conditions be registered in ClinicalTrials.gov [10], a clinical trial registry data bank developed and maintained by the National Library of Medicine (NLM), part of the National Institutes of Health (NIH). As of August 2010, ClinicalTrials.gov, the largest of its kind in the world, contains over 100,000 trials from over 170 countries and is used by approximately 65,000 visitors each day. In ClinicalTrials.gov, each registered record provides information about a trial's purpose, condition, intervention, detailed description, eligibility (who may participate), location, status, etc. Most of the above information is described in natural language. Fig. 2 shows an example of a clinical trial record in ClinicalTrials.gov. Like other research databases [11], ClinicalTrials.gov captures important scientific and clinical investigations in biomedicine. As a result, the knowledge buried in those trial records has shown to be valuable for researchers, clinicians, and the pharmaceutical industry [12–14].

In this study, we hypothesize that ClinicalTrials.gov is a comparably rich data source to the biomedical literature for PGx clinical outcome related information (i.e., how genes affect drug responses in patients with specific diseases). In this regard, we developed a text-mining approach to systematically recognize PGx relevant relationships between genes, drugs, and diseases from both trial record metadata and descriptions (in free text). It should be noted that many relationships mentioned in clinical trials may still be under investigation (i.e., not yet concluded). Despite this fact, they are not selected without cause or randomly. Rather, they are carefully designed and conducted based on the accumulated knowledge from preliminary studies [15,16]. Therefore, these relationships are reasonable candidates for potential inclusion in the databases of PGx studies. In addition, the speculative relationships themselves are valuable for pharmacosurveillance and pharmacovigilance studies [17]. Moreover, mining PGx information from clinical trials as opposed to the scientific literature has one major advantage: detecting the important information in a more timely manner. That is, a relationship mentioned in a trial may not appear in the literature until several years later. This is expected because it takes time for a trial to be conducted, concluded, and published. To investigate this time lag issue further, we analyzed 8588 PubMed® citations that were manually linked to 7224 trials [18] in ClinicalTrials.gov and computed the time lags between the trials and their resulting publications. As shown in Fig. 3a, we found that the average time difference between a trial's start date and publication date is approximately 5 years. Once a trial is completed, majority of them (∼62%) have their results published within 2–3 years (see Fig. 3b). For instance, to study how genetic polymorphisms influence the efficacy and side effect profiles of Paroxetine and Escitalopram for major depression treatments, a Phase IV clinical trial entitled 'Clinical pharmacogenomics of antidepressant response' was launched and registered in ClinicalTrials.gov in 2006 (See Fig. 2; NCT number = NCT00384020). The trial was completed 4 years later in 2010 and shortly thereafter, the trial was published in a journal article entitled 'Genetic polymorphisms of cytochrome P450 enzymes influence metabolism of the antidepressant escitalopram and treatment response' [19].

Note that in Fig. 3a there are articles that were published in the same year as their trials started (i.e., zero year difference between the trial start date and publication date). We looked into these cases and found that some of these articles are actually describing the study rationale and protocol rather than study results (*e.g.*, article 'PMID = 19828019' and its corresponding trial 'NCT0086251'). Also, some reported trial start dates in ClinicalTrials.gov are likely to be errors. For example, the article (PMID = 18761748) published in September 2008 is linked to a trial (NCT00147966) whose registered start date is June 2008, which is likely to be an error in this case.

2. Related work

In this work, we propose a text-mining approach to identify PGx relevant gene–drug–disease relationships from registered trial records. The related work includes manual curation of gene–drug–disease relationships in PharmGKB, text mining techniques for extracting PGx concepts and relationships, and other text-mining applications to clinical trial records.

2.1. Curated gene–drug–disease relationships in PharmGKB

In PharmGKB, the gene–drug–disease relationships are identified based on human curation and further classified into one of

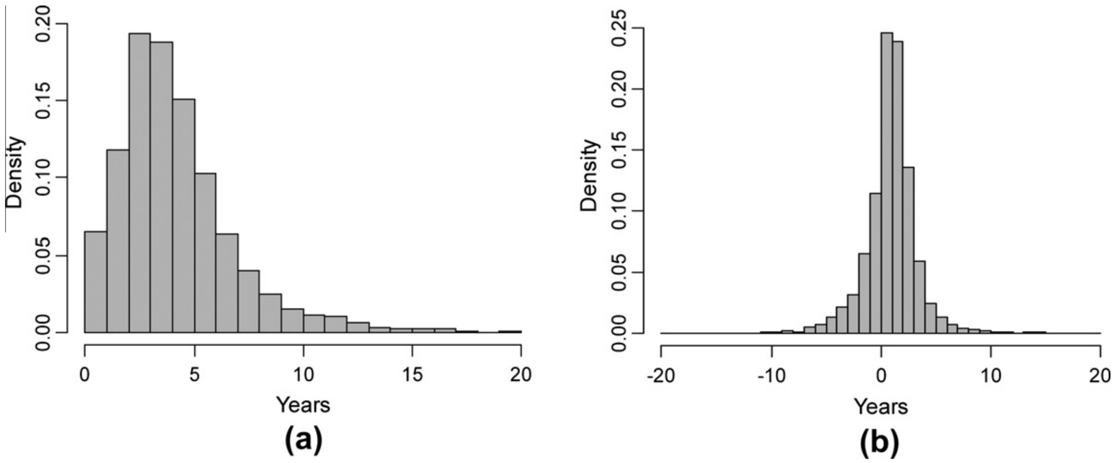


Fig. 3. Time lag between clinical trial and publication. (a) Time lag from trial start date to publication date. On average, a publication occurs 5 years after its corresponding trial starts. (b) Time lag from trial completion date to publication date (data to the left of 0 years suggest that some publications occur before the completion of their trials). Majority of the publications occur 2–3 years after the completion of their corresponding trials.

the five general PGx categories: clinical outcome (CO), pharmacodynamics and drug responses (PD), pharmacokinetics (PK), molecular and cellular functional assays (FA), and genotype (GN). The data in the clinical outcome category demonstrates that the genetic variability in the context of a drug effect significantly changes medical outcomes. For example, the gene 'TYMS', drug 'methotrexate', and disease 'precursor cell lymphoblastic leukemia–lymphoma' were identified for curation in PharmGKB based on a relevant article 'Polymorphism of the thymidylate synthase gene and outcome of acute lymphoblastic leukaemia' [20]. Subsequently, the above gene–drug–disease relationship was classified into the clinical outcome category. As of August 2010, PharmGKB covers 1621 such gene–drug–disease relationships categorized as clinical outcome.

2.2. Text mining techniques for extracting PGx concepts and relationships

Concept identification serves as a prerequisite for many subsequent tasks of biomedical text mining like relationship extraction [21]. In PGx studies, the key concepts include gene, gene variant, drug and disease. Text mining tools have been developed for identifying these concepts such as GAPSCORE for identifying genes from PGx articles [22] and MutationFinder for identifying gene variants [23]. The relationships between identified PGx concepts can be nontyped (*e.g.*, relationship between gene 'TYMS', drug 'methotrexate', and disease 'precursor cell lymphoblastic leukemia–lymphoma' is discussed in article PMID = 11937185) or specific (*e.g.*, gene 'TYMS' variants affect the clinical outcome of 'precursor cell lymphoblastic leukemia–lymphoma' patient treated with 'methotrexate'). Some attempts have been made for PGx relationship extraction. For example, Garten and Altman developed an ontology-based tool, Pharmspresso, for extracting PGx information from full text articles by identifying concepts (such as genes, drugs, polymorphisms, and diseases) and relationships (such as action, association and comparison) [24]. Ahlers et al. developed a rule-based method for extracting specific PGx relationships such as 'genetic etiology' and 'pharmacological effects' from PubMed abstracts [25]. Theobald et al. computed conditional probabilities of PGx relationships between drugs, diseases, and genes by analyzing their co-occurrences in PubMed abstracts [26]. Coulet et al. developed a method to identify PGx relationships using syntactic rules and to organize these relationships in an ontology that maps diverse sentence structures and vocabularies to common semantics [27].

Research on applying text mining techniques in PGx studies is gaining attention and has achieved significant improvement in the recent years. A review of text-mining progress in PGx information extraction can be found in [6]. Recent workshops devoted to this domain were held in the Pacific Symposium on Biocomputing, where the 2010 and 2011 workshop themes were respectively 'extraction of genotype–phenotype–drug relationships form texts: from entity recognition to bioinformatics application' [28] and 'mining the pharmacogenomics literature' [29].

2.3. Other text-mining applications to clinical trial records

Clinical trials provide valuable information about the efficacy/toxicity of medical intervention. Text-mining techniques have been applied to published randomized clinical trial literature for extracting patient demographic information such as trial size and disease/symptom descriptors [13]. To enable semantic representation and search for clinical research eligibility criteria, some text mining studies have focused on extracting information from the narrative descriptions of eligibility criteria in trial records [30,31].

At present, users can search for trials in ClinicalTrials.gov by entering keywords in the search box. The lack of unambiguous names for entities (*e.g.*, intervention, condition, and gene) affects the retrieval of all relevant trials that meet users' specifications. For example, more than 60% of the studies in ClinicalTrials.gov about heart attacks do not contain the phrase 'heart attack' but use a different term such as myocardial infarction [32]. To solve this issue, the embedded search engine of ClinicalTrials.gov expands user queries using synonyms derived from the Unified Medical Language System (UMLS) [33] and rank the retrieval results based on a probabilistic model [34]. This query expansion feature enables users to retrieve trials which use the term 'myocardial infarction' in the condition description as 'heart attack' related ones. However, it remains 'myocardial infarction' and 'heart attack' in the contexts of trial records not identified as the same disease concept. This makes ClinicalTrials.gov difficult to link to other related resources (*e.g.*, PharmGKB). An attempt to use a standardized nomenclature for representing various clinical research eligibility entities was reported by Luo et al. [35,36].

3. Methods

The goal of this study is to systematically identify clinical PGx information from clinical trial records. Fig. 4 shows an overview of our workflow. We collected 93,661 clinical trial records from ClinicalTrials.gov as of August 2010. We first preprocessed these records and identified sections of interest. Second, we used a dictionary-based method to identify PGx concepts (i.e., diseases, drugs and genes) from the preprocessed trial records. Our gene–drug–disease relationship extraction is based on their co-occurrence in one trial record. Finally, we indexed the trial records with the identified PGx concepts. Hence, given a target PGx gene, our approach can return related diseases and drugs with corresponding trials. Similarly, given a target pair of PGx gene and drug, our approach can return trials in which the PGx pair is or was under investigation.

3.1. Preprocessing clinical trial records

In ClinicalTrials.gov, each trial record is divided into sections, and each section is described in free-style texts (see Fig. 2). The condition section includes information on the disease, disorder, syndrome, illness, or injury being studied in a trial. The intervention section includes information on the drug, vaccine, procedure, device, or other potential treatment being investigated in a trial. The study description section describes the study hypothesis, design, and all the information on trial intended for the lay public. These three sections were identified as important for this work and extracted for further PGx concept identification. Specifically, the condition section was used for disease identification, intervention section for drug, and study description section for gene.

3.2. Extracting gene–drug–disease relationships

We used a dictionary-based method to identify genes, drugs, and diseases from the preprocessed trial records. Three PharmGKB dictionaries were collected in August 2010, containing 3197 diseases, 2984 drugs, and 26,216 genes respectively. Each concept and its synonyms in the dictionary are assigned an internal PharmGKB identifier. For example, the drug concept 'imatinib' in PharmGKB, together with its list of synonyms 'Gleevec', 'Glivec', 'Imatinib Mesylate', and 'Imatinib Methansulfonate' are assigned a PharmGKB_Id = 'PA10804'. Both name and synonyms were used for identifying PGx entities in trial records. The PGx concept identification and normalization facilitate PGx information retrieval

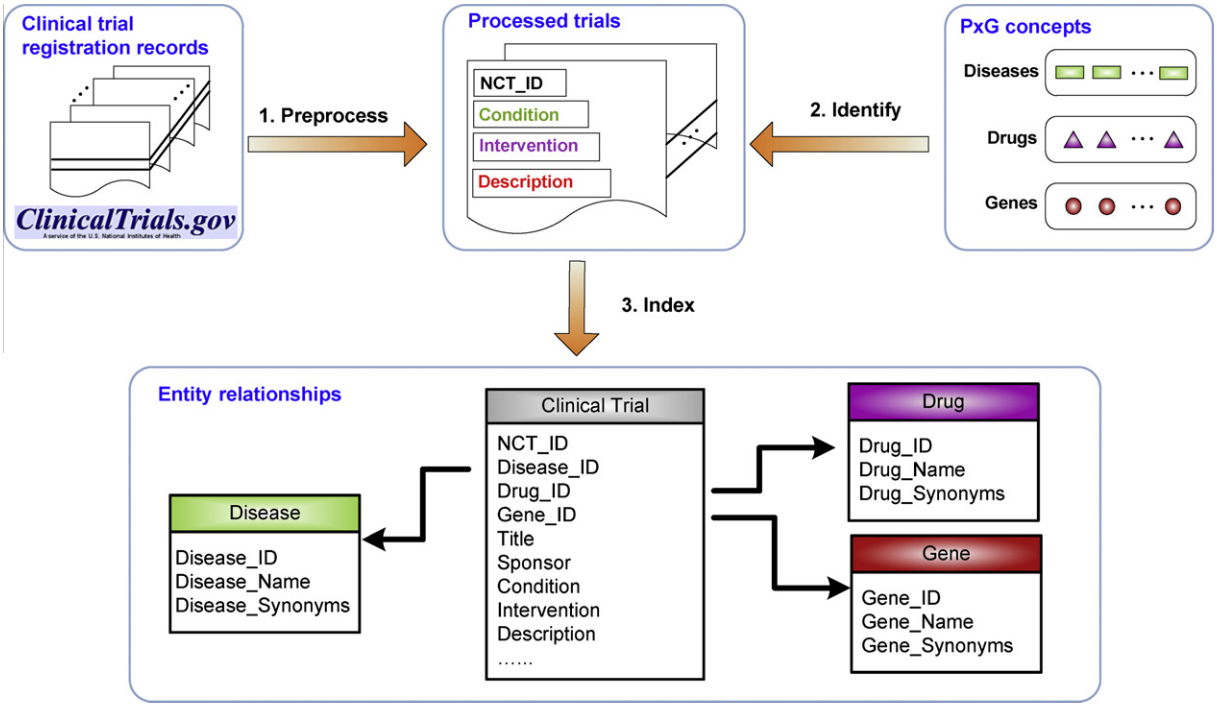


Fig. 4. Workflow for mining clinical trial records.

from ClinicalTrials.gov. Furthermore, this makes the linking analysis between ClinicalTrials.gov and PharmGKB feasible.

For gene–drug–disease relationship extraction, we used a co-occurrence based method. We assume that there is a clinical outcome association between gene, drug and disease (i.e., how gene affects drug responses in patients with specific disease/condition) if they co-occur in the same trial record in ClinicalTrials.gov.

3.3. Indexing clinical trials

We systematically compiled the extracted PGx concepts and relationships with their identifiers linking to the corresponding trial records (see Fig. 4). To facilitate the retrieval of PGx information from clinical trials, we built an index for the PGx concepts and trial records such that given a PGx gene, our approach first looks up the gene dictionary for its identifier, and then readily retrieves all the trials containing that gene identifier. Similarly, given a PGx gene–drug pair, our approach first looks up the gene and drug dictionary respectively for their identifiers, and then retrieves all the trials in which both identifiers are present.

3.4. Hypothesis testing and method evaluation

To test our hypothesis that ClinicalTrials.gov is a comparable source rich in PGx related information, we first compared our results (i.e., extracted PGx relationships between genes, drugs and diseases) in trial records against the ones found in PharmGKB and in PubMed, respectively.

Second, to assess the performance of our text-mining approach, we manually reviewed a subset of automatically extracted relationships. Specifically, two human annotators (JL and ZL) were asked to manually assign one of the following categories to 100 extracted gene–drug–disease relationships: the relationship was not mentioned in the trial record (Category I); the relationship was explicitly mentioned in the trial record with or without supporting publications (Categories II and III). When computing accuracy for our method, both Categories II and III were considered as correct extractions.

4. Results

4.1. Comparative evaluation of ClinicalTrials.gov

For contrasting ClinicalTrials.gov with PharmGKB and PubMed, we compared their coverage of 3-way gene–drug–disease PGx relationships, which were obtained based on the input of 26 PGx gene–drug pairs [37] from the PharmGKB website.

Given these 26 PGx gene–drug pairs, our approach was able to identify 348 clinical trials and 240 3-way PGx relationships. By querying the given PGx pairs in PubMed [38] while limiting the publication type to be 'Clinical Trial' and MeSH [39] (Medical Subject Headings) terms to be 'Genetic Variation' or 'Genotype', 1162 3-way relationships were retrieved in 448 PubMed citations. Finally, we found 261 such 3-way relationships curated in PharmGKB.

Fig. 5 shows a detailed comparison of the 3-way gene–drug–disease relationships found in the clinical trials (blue[1] circle), PubMed abstracts (green circle) and PharmGKB (red circle). 124 (51.7%) and 68 (28.3%) of the relationships found in ClinicalTrials.gov were also present in PubMed and PharmGKB, respectively. For the common 51 drug–gene–disease relationships which were found in all three sources, approximately 75% of them occurred earlier in trials than in PharmGKB or PubMed.

Our approach also identified 99 gene–drug–disease relationships which are currently missing in both PharmGKB and PubMed. Our further analysis shows that majority (65%) of them were found from ongoing trials (*e.g.*, recruiting or active but not recruiting). For example, the 'UGT1A1'–'irinotecan'–'Gastrointestinal Cancer' relationship was extracted from a Phase I trial (NCT00654160) which was launched in 2008 and expected to be completed in 2015. In

[1] For interpretation of color in Fig. 5, the reader is referred to the web version of this article.

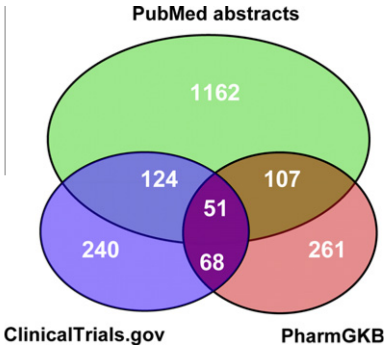


Fig. 5. Comparison of gene–drug–disease relationships identified from different sources. A total of 240 relationships were found in ClinicalTrials.gov. 124 and 68 such relationships were found to be overlapping with 1162 results in PubMed and 261 results in PharmGKB, respectively. Fifty-one relationships were found in all three sources.

this trial, researchers proposed to study UGT1A1 genotype-based dosing of irinotecan when given together with fluorouracil and leucovorin in treating patients with advanced gastrointestinal cancer. As of August 2010, this 'UGT1A1'–'irinotecan'–'Gastrointestinal Cancer' relationship is present in neither PharmGKB nor PubMed.

4.2. Assessment of our automatic approach

For the 240 identified gene–drug–disease relationships by our method, 100 of them were randomly selected for manual review and classification. 74 were judged to be correct extractions: 30 in Category II and 44 in Category III (see category details in Section 3.4). Hence, our text-mining approach achieves an accuracy of 74%.

Table 1 shows 10 examples of correctly identified relationships, as well as their supporting statements and corresponding publications (when available) in the trials. In our evaluation, the first seven relationships were classified to be Category II and the other 3 Category III. Take the 'UGT1A1'–'irinotecan'–'Lung Cancer' relationship for example. Our method extracted this relationship from a Phase III clinical trial (NCT00045162) which proposed to determine the association between UGT1A1 polymorphisms and irinotecan-assoicated toxic effect in patients with lung cancer. After 7 years in 2009, the pharmacogenomics results of this trial were published (PMID = 19349543), reporting that UGT1A1 (G-3156A)A/A (drug metabolism) was associated with IP (Irinotecan plus cisplatin) related neutropenia. As of August 2010, these 10 relationships were missing in PharmGKB. Thus, we believe the relationships identified by our approach are valuable for inclusion to related PGx knowledge bases.

Table 1
Examples of correctly identified gene–drug–disease relationships currently absent in PharmGKB.

Gene	Drug	Disease	Relevant clinical trial and study statement	Corresponding trial results published in the literature
EGFR	Gefitinib	Head and Neck Cancer	NCT00083057 (**2004**): Gefitinib, paclitaxel, and radiation therapy in treating patients with head and neck cancer; molecular targets analysis of EGFR downstream signaling pathway	**PMID**: 19879702 (**2010**): Only 1 patient demonstrated a reduction in phosphorylated EGFR, decreased downstream signaling, and reduced cellular proliferation after initiating GEF
EGFR	Gefitinib	Ovarian Cancer	NCT00049556 (**2002**): This phase II trial is studying how well gefitinib works in treating patients with ovarian cancer. Correlate the biologic modulation of EGFR by this drug with outcome and toxic effects in these patients	**PMID**: 17330838 (**2007**): The results from this study demonstrated that efitinib inhibited the phosphorylation of EGFR in EOC (epithelial ovarian cancer) tumor cells, providing proof of target in a clinical setting
EGFR	Erlotinib	Biliary Cancer	NCT00356889 (**2006**): Determine the presence of EGFR mutations in tumor tissue and correlate this with outcome	**PMID**: 20530271 (**2010**): Low repeats (<16) in EGFR intron 1 polymorphism and G > G k-ras Q38 genotype (wild type) were associated with improved outcomes
UGT1A1	Irinotecan	Head and Neck Cancer	NCT00040807 (**2002**): Correlate UGT1A1 genotype with the toxic effects of irinotecan and docetaxel in these patients	**PMID**: 19634157 (**2009**) The authors explored the association between polymorphisms in the UGT1A1 gene and race, neutropenia, diarrhea, and any toxicity among 35 patients. There were no statistically significant differences in the pattern of worst degree toxicity, or in the grade intensity of neutropenia, or diarrhea by TA repeat category
UGT1A1	Irinotecan	Neuroblastoma	NCT00093353 (**2004**): Correlate UGT1A1 genotype with the occurrence of dose-limiting diarrhea in Neuroblastoma patients treated with irinotecan, temozolomide, and cefixime	**PMID**: 19171709 (**2009**): No association was seen between UGT1A1 genotype and toxicity in this small study
UGT1A1	Irinotecan	Lung Cancer	NCT00045162 (**2002**): Determine the association between UGT1A1 polymorphisms and irinotecan-associated toxic effects in the patients with lung cancer	**PMID**: 19349543 (**2009**): UGT1A1 (G-3156A)A/A (drug metabolism) was associated with IP (Irinotecan plus cisplatin)-related neutropenia.
G6PD	Chloroquine	Malaria	NCT00118794 (**2004**): Evaluate the extent to which the risks associated with the use of chlorproguanil-dapsone in settings without G6PD screening might outweigh the benefits to malaria treatment	**PMID**: 21666744 (**2011**) The authors found evidence of an interaction of treatment group with parasite density, suggesting that failure to rapidly eliminate parasitaemia may have explained the anaemia after chlorproguanil-dapsone in G6PD normal subjects
UGT1A1	Irinotecan	Advanced Gastric Cancer	NCT01136031 (**2008**): Study the accurate maximum tolerated dose (MTD) of the paclitaxel and irinotecan combination regimen after considering the UGT1A1 polymorphism of patients	**None**
CYP2C9	Celecoxib	Colorectal Cancer	NCT00685568 (**2002**): Assess the influence of polymorphism CYP2C9 on age of onset, phenotype or number of colorectal polyps	**None**
DPYD	Capecitabine	Pancreatic Cancer	NCT00303927 (**2005**): Explore the association between capecitabine exposure at steady-state, allelic variants in candidate gene dihydropyrimidine dehydrogenase and drug response in this patient population	**None**

5. Discussion

Our approach shows that ClinicalTrials.gov is rich in revealing gene–drug–disease relationships for PGx studies. Absent from the current PGx knowledge base (PharmGKB), many of the identified PGx relationships are associated with potential clinical outcomes. In this section, we will discuss the issues of coverage and time lag, the practical implications of this research, and the limitations of our approach, as well as future work.

5.1. Coverage

Although we observed a statistically significant overlap between our results and curated facts in PharmGKB (hypergeometric p-value <0.05), some curated PGx relationships were not detected from clinical trials. This is mainly due to the incompleteness of trial registration, especially for the trials held outside of the United States. For example, the relationship between gene 'CYP2C9', drug 'tamoxifen' and disease 'breast cancer' was studied in a clinical trial in Turkey which was not registered in ClinicalTrials.gov. On the other hand, the study results were already published in an article entitled 'Tamoxifen inhibits cytochrome P450 2C9 activity in breast cancer patients' (PMID = 17024799). As a result, the clinical outcome investigation on the 'CYP2C9'–'tamoxifen'–'breast cancer' relationship was curated based on the publication but not found in ClinicalTrials.gov by our approach. Good news is that ClinicalTrials.gov is now making efforts to collaborate with other countries in creating a universal registration system [32]. This endeavor would promote the accessibility of clinical trials in all countries in the future.

5.2. Time lag

As mentioned in Section 4.1, approximately 75% of the 3-way drug–gene–disease PGx relationships were identified earlier in trials than in publications. For the remaining 25% of the relationships, we found two main reasons why they were found otherwise (i.e., earlier in publications than in trials). First, it is due to the fact that the earliest trial of an identified relationship is not registered in ClinicalTrials.gov. For example, the relationship between gene 'SULT1A1', drug 'tamoxifen', and disease 'breast cancer' was curated in PharmGKB based on a supporting article (PMID = 15024382) published in 2004 but its corresponding trial is missing in ClinicalTrials.gov. On the other hand, a different trial reporting the same relationship was registered in ClinicalTrials.gov in 2008 (NCT00667121). Hence, the first appearance of the 'SULT1A1'–'tamoxifen'–'breast cancer' relationship in trial records was dated as 2008 by our method—4 year behind the earliest publication year.

Second, it is due to the discrepancy between the nature of clinical trials and curation scope of PharmGKB. A PGx related clinical trial is designed to study the direct relationships between genes, drugs and diseases, (i.e., how genes affect drug responses in patients with specific diseases/conditions). However, both direct and indirect relationships are captured by PharmGKB [40]. For example, in PharmGKB, the clinical outcome annotation for the relationship between gene 'CYP3A4', drug 'pantoprazole', and disease 'Gastroesophageal Reflux Disease (GERD)' is curated based on an article (PMID = 16961157) published in 2006. However, in ClinicalTrials.gov the earliest trial for investigating this relationship was not registered until 2009 (NCT00744419). Therefore, in this case the 'CYP3A4'–'pantoprazole'–'GERD' relationship was detected 3 years ahead in the literature than in ClinicalTrials.gov. However, our further examination shows that the curated article (PMID = 16961157) is a review rather an original research report. In that review article, several genes (CYP2C19 and CYP3A4), drugs (amoxicillin, esomeprazole, pantoprazole, etc.), and diseases (Gastroesophageal Reflux and Peptic Ulcer) were discussed but the exact relationship between 'CYP3A4', 'pantoprazole', and 'GERD' was not reported.

Note that in Fig. 3a we show that for each trial related article, its publication date is always after its corresponding trial start date. But with respect to PGx relationships, owning to the aforementioned reasons, some may be found earlier in publications than in trials.

5.3. Practical implications of this research

As mentioned earlier, anyone using extracted relationships from this research should be cautioned that some of those relationships are still under investigation and thus not concluded. Nonetheless, we believe these speculative relationships are still valuable for inclusion to relevant knowledge bases (perhaps with special remarks). Below, we use PharmGKB as a representative PGx knowledge base and show two potential practical uses of our research findings:

First, we recommend building cross-links between PharmGKB and ClinicalTrials.gov. Doing so would allow PharmGKB users to readily identify clinical trials in which relevant PGx genes are under investigation for different conditions and interventions. On the other hand, through linking to PharmGKB, ClinicalTrials.gov users can be exposed to the most comprehensive knowledge of PGx concepts such as gene variants and genetic tests.

Second, relationships found in ClinicalTrials.gov but currently missing in PharmGKB may be considered for future curation. In this regard, we have two specific recommendations for prioritizing the list of candidate relationships: (a) based on our analysis, any extracted relationships that are associated with multiple supporting trials should be of high priority; and (b) any relationships that are associated with completed and published clinical trials should be of high priority. For example, the relationship between gene 'EGFR', drug 'gefitinib', and disease 'Head and Neck Cancer' is associated with four clinical trials (i.e., NCT00083057, NCT00088907, NCT00820417 and NCT00169221). Moreover, the study status of one trial (NCT00083057) is indicated as 'completed' and its results are published. Thus, the 'EGFR'–'gefitinib'–'Head and Neck Cancer' relationship should be of high priority for curation consideration.

5.4. Limitations of our approach and future work

In this study, we used a dictionary-based method for gene, disease, and drug identification for directly associating with the PharmGKB vocabulary. Like any other dictionary-based method, our approach favored precision but failed to identify entity variants not covered by the used dictionaries. Also, due to name ambiguity between entity types, we may occasionally have identified false positives in our results. For example, the PGx gene symbol 'TPMT' is also the abbreviation of drug 'topiramate'. This ambiguity directly led to an error in gene identification from trial (NCT00884884) in which TPMT is indicated as the short form of the antiepileptic drug 'topiramate'.

In relationship extraction, we used a co-occurrence based method for identifying relationships between genes, drugs, and diseases. Although this method has been successfully applied in a number of studies such as [41–43], it has certain limitations: (a) not all co-occurred relationships are actually meaningful (accounting for 26% of the errors in relationship extraction); and (b) we cannot characterize the types of relationships extracted.

In the future, we plan on (a) improving the methods for PGx concept identification and relationship extraction using more sophisticated NLP techniques such as dependency parsing [44];

(b) designing a method for ranking extraction results by combining features like relevant trial status and numbers; (c) developing a robust method for linking clinical trials to their corresponding publications when they are not manually supplied by the trial investigators; (d) developing an automatic method to detect specific gene variants and allele changes which affect drug response reported in trial results and further link them to a standardized gene variation database such as dbSNP [45].

6. Conclusions

The clinical trial is at a critical juncture in the drug development pipeline, connecting previous studies on molecular mechanism with a final decision of approval. We successfully developed a systematic approach to automatically identify clinical PGx information from registered clinical trials. In this study, we collected 93,661 clinical trial records from ClinicalTrials.gov and used a dictionary-based method to identify and normalize PGx concepts (i.e., diseases, drugs and genes) in the texts of the collected trial records. In relationship extraction, we used a co-occurrence based method for identifying relationships between genes, drugs, and diseases. To facilitate the retrieval of PGx information from clinical trials, we built an index for the PGx concepts and the trials collected in our study. Hence, given a pair of PGx gene–drug relationship, our approach can return trials in which the PGx pair is studied under different conditions and controls. In comparative evaluation, we show that ClinicalTrials.gov is a rich source of PGx gene–drug–disease relationships. Manual review shows that our automatic identification method achieves an accuracy of 74%. By comparing our results with the relationships identified from PubMed abstracts and in PharmGKB, we found that our approach can potentially enrich current resources and accelerate the dissemination of clinical outcome information of pharmacogenomics.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. The authors would like to thank Dr. John Wilbur for his helpful comments on this manuscript and Dr. Bastien Rance for his valuable discussion on the related work. The authors would also like to thank the PharmGKB team for clarifying their curation scope and discussing the usefulness of our work.

References

[1] Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. Science 1999;286(5439):487–91.

[2] Penny MA, McHale D. Pharmacogenomics and the drug discovery pipeline: when should it be implemented? Am J Pharmacogenomics 2005;5(1):53–62.

[3] Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics research network and knowledge base. Pharmacogenomics J 2001;1(3):167–70.

[4] Thorn CF, Klein TE, Altman RB. Pharmacogenomics and bioinformatics: PharmGKB. Pharmacogenomics 2010;11(4):501–5.

[5] Altman RB, Flockhart DA, Sherry ST, Oliver DE, Rubin DL, Klein TE. Indexing pharmacogenetic knowledge on the World Wide Web. Pharmacogenetics 2003;13(1):3–5.

[6] Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics 2010;11(10):1467–89.

[7] Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291(20):2457–65.

[8] Zarin DA, Tse T. Medicine. Moving toward transparency of clinical trials. Science 2008;319(5868):1340–2.

[9] The Food and Drug Administration Modernization Act of 1997. Pub. L. No. 105-115, 113; 1997.

[10] ClinicalTrials.gov. <http://clinicaltrials.gov/>.

[11] Hurdle JF, Botkin J, Rindflesch TC. Leveraging semantic knowledge in IRB databases to improve translation science. AMIA Annu Symp Proc; 2007. p. 349–53.

[12] Rennie D. Trial registration: a great idea switches from ignored to irresistible. JAMA 2004;292(11):1359–62.

[13] Xu R, Garten Y, Supekar KS, Das AK, Altman RB, Garber AM. Extracting subject demographic information from abstracts of randomized clinical trial reports. Stud Health Technol Inform 2007;129(Pt 1):550–4.

[14] Xu R, Supekar K, Huang Y, Das A, Garber A. Combining text classification and Hidden Markov Modeling techniques for categorizing sentences in randomized clinical trial abstracts. AMIA Annu Symp Proc; 2006. p. 824–8.

[15] Orloff J, Douglas F, Pinheiro J, Levinson S, Branson M, Chaturvedi P, et al. The future of drug development: advancing clinical trial design. Nat Rev Drug Discov 2009;8(12):949–57.

[16] Kramer JA, Sagartz JE, Morris DL. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. Nat Rev Drug Discov 2007;6(8):636–49.

[17] Introduction to drug utilization research. Oslo, Norway: World Health Organization; 2003. p. 8.

[18] ClinicalTrials.gov identifier to be added to MEDLINE®/PubMed® data. <http://www.nlm.nih.gov/pubs/techbull/mj05/mj05_ct.html>.

[19] Tsai MH, Lin KM, Hsiao MC, Shen WW, Lu ML, Tang HS, et al. Genetic polymorphisms of cytochrome P450 enzymes influence metabolism of the antidepressant escitalopram and treatment response. Pharmacogenomics 2010;11(4):537–46.

[20] Krajinovic M, Costea I, Chiasson S. Polymorphism of the thymidylate synthase gene and outcome of acute lymphoblastic leukaemia. Lancet 2002;359(9311):1033–4.

[21] Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 2006;7(2):119–29.

[22] Chang JT, Schutze H, Altman RB. GAPSCORE: finding gene and protein names one word at a time. Bioinformatics 2004;20(2):216–25.

[23] Caporaso JG, Baumgartner Jr WA, Randolph DA, Cohen KB, Hunter L. MutationFinder: a high-performance system for extracting point mutation mentions from text. Bioinformatics 2007;23(14):1862–5.

[24] Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics 2009;10(Suppl. 2(S6)).

[25] Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput; 2007. p. 209–20.

[26] Theobald M, Shah N, Shrager J. Extraction of conditional probabilities of the relationships between drugs, diseases, and genes from PubMed guided by relationships in PharmGKB. Summit on Translat Bioinforma; 2009. p. 124–8.

[27] Coulet A, Shah NH, Garten Y, Musen M, Altman RB. Using text to build semantic networks for pharmacogenomics. J Biomed Inform 2010;43(6):1009–19.

[28] Coulet A, Shah N, Hunter L, Barral C, Altman RB. Extraction of genotype–phenotype–drug relationships from text: from entity recognition to bioinformatics application. Pac Symp Biocomput, 2010. p. 485–7.

[29] Cohen KB, Garten Y, Hahn U, Shah NH. Mining the pharmacogenomics literature – workshop introduction. Pac Symp Biocomput; 2011. p. 362–3.

[30] Tu SW, Peleg M, Carini S, Bobak M, Ross J, Rubin D, et al. A practical method for transforming free-text eligibility criteria into computable criteria. J Biomed Inform 2011;44(2):239–50.

[31] Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. J Am Med Inform Assoc 2011.

[32] Zarin DA, Ide NC, Tse T, Harlan WR, West JC, Lindberg DA. Issues in the registration of clinical trials. JAMA 2007;297(19):2112–20.

[33] Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inform Med 1993;32(4):281–91.

[34] Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 2007;14(3):253–63.

[35] Luo Z, Duffy R, Johnson S, Weng C. Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS. In: AMIA summits on translational science proceedings. San Francisco, USA; 2010. p. 26–30.

[36] Luo Z, Johnson SB, Weng C. Semi-automatically inducing semantic classes of clinical research eligibility criteria using UMLS and hierarchical clustering. In: AMIA annual symposium proceedings. Washington, DC, USA; 2010. p. 487–91.

[37] Well known PGx gene–drug relationships. <http://www.pharmgkb.org/resources/forScientificUsers/well_known_pairs_of_gene-drug_pgx_relationships.jsp>.

[38] PubMed®. <http://www.ncbi.nlm.nih.gov/pubmed/>.

[39] Medical Subject Headings (MeSH®). <http://www.nlm.nih.gov/mesh/>.

[40] How are pharmacogenomics articles annotated in PharmGKB? <http://www.pharmgkb.org/resources/faqs.jsp#FAQs-annotatedPKPD>.

[41] Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics 2004;20(Suppl. 1):i290–6.

[42] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform 2005;74(2–4):289–98.

[43] Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput Biol 2010;6(9).

[44] De Marneffe MC, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. In: The international conference on language resources and evaluation. Genoa, Italy; 2006. p. 449–54.

[45] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. DbSNP: the NCBI database of genetic variation. Nucl Acids Res 2001;29(1):308–11.