Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Combining automatic table classification and relationship extraction in extracting anticancer drug–side effect pairs from full-text articles

Rong Xu [a,*], QuanQiu Wang [b,*]

[a] Medical Informatics Program, Center for Clinical Investigation, Case Western Reserve University, Cleveland, OH 44106, United States
[b] ThinTek, LLC, Palo Alto, CA 94306, United States

ARTICLE INFO

Article history:
Received 30 May 2014
Accepted 3 October 2014
Available online xxxx

Keywords:
Text mining
Information extraction
Cancer drug side effect
Drug discovery
Drug repositioning
Drug toxicity prediction

ABSTRACT

Anticancer drug-associated side effect knowledge often exists in multiple heterogeneous and complementary data sources. A comprehensive anticancer drug–side effect (drug–SE) relationship knowledge base is important for computation-based drug target discovery, drug toxicity predication and drug repositioning. In this study, we present a two-step approach by combining table classification and relationship extraction to extract drug–SE pairs from a large number of high-profile oncological full-text articles. The data consists of 31,255 tables downloaded from the Journal of Oncology (JCO). We first trained a statistical classifier to classify tables into SE-related and -unrelated categories. We then extracted drug–SE pairs from SE-related tables. We compared drug side effect knowledge extracted from JCO tables to that derived from FDA drug labels. Finally, we systematically analyzed relationships between anti-cancer drug-associated side effects and drug-associated gene targets, metabolism genes, and disease indications. The statistical table classifier is effective in classifying tables into SE-related and -unrelated (precision: 0.711; recall: 0.941; F1: 0.810). We extracted a total of 26,918 drug–SE pairs from SE-related tables with a precision of 0.605, a recall of 0.460, and a F1 of 0.520. Drug–SE pairs extracted from JCO tables is largely complementary to those derived from FDA drug labels; as many as 84.7% of the pairs extracted from JCO tables have not been included a side effect database constructed from FDA drug labels. Side effects associated with anticancer drugs positively correlate with drug target genes, drug metabolism genes, and disease indications.

1. Introduction

Drug-induced side effects are observable phenotypes of drugs manifested at the level of the whole body system and are mediated by a drug interacting with its on- or off-targets through a cascade of downstream pathway perturbations. Systematic and integrated approaches to studying drug-associated side effects have the potential to illuminate the complex pathways of drug-induced toxicities, allowing for the identification of novel drug targets, prediction of unknown drug toxicities, and repositioning of existing drugs for new disease indications.

Computational approaches to drug target discovery, unknown drug toxicities prediction, and drug repositioning have primarily relied on drug molecular structures or functions such as chemical structure, molecular activity, and molecular docking [2,7,11,14,17,18,20,26,34,35]. These computational approaches largely depend on the availability of drug molecular structure or function knowledge bases. Systems approaches would greatly benefit from the vast amount of higher-level clinical phenotype data such as observed drug-related side effects [3,5,6,13]. It has been increasingly recognized that similar side effects of seemingly unrelated drugs can be caused by their common off-targets and that drugs with similar side effects are likely to share molecular targets [5]. Therefore, systems approaches to studying the phenotypic relationships among drugs and integrating the high-level drug phenotypic data with lower-level genetic and chemical data will allow for a better understanding of drug toxicities. Current systems approaches to studying phenotypic relationships among drugs rely exclusively on information extracted from FDA drug labels [3,5,17]. It was recently demonstrated that 39% of serious events associated with targeted anticancer drugs are not reported in clinical trials and 49% are not described in initial FDA drug labeling [21]. For the successful development of phenotype-driven systems approaches to understanding drug-associated side effects, the availability of a comprehensive and machine-understandable drug–side effect (SE) relationship knowledge base is critical.

Drug–SE relationship extraction and mining from multiple heterogeneous and complementary data sources is an active research area. Kuhn et al. developed text mining approaches in constructing

* Corresponding authors.
E-mail addresses: rxx@case.edu (R. Xu), qwang@thintek.com (Q. Wang).

http://dx.doi.org/10.1016/j.jbi.2014.10.002
1532-0464/© 2014 Elsevier Inc. All rights reserved.

a side effect resource (SIDER) from FDA drug labels [15]. Currently, SIDER represents the best source of computable drug side effect association knowledge. Systems approaches to studying this information have led to the prediction of several new drug targets [5,17]. The FDA Adverse Event Reporting System (FAERS) is the spontaneous reporting system overseen by the U.S. FDA and the main resources for post-marketing drug safety surveillance. Mining drug–side effect (drug–SE) relationships from FAERS is a highly active research area. Data mining algorithms such as disproportionality analysis, correlation analysis, multivariate regression, and signal ranking and filtering leveraging external knowledge have been developed to detect adverse drug signals from FAERS [1,12,23,28,29]. Another important information source of drug–SE associations is the vast amount of published biomedical literature. Currently, more than 22 million biomedical abstracts are publicly available on MEDLINE, making it a rich source of side effect information for drugs at all clinical stages, including drugs in pre-marketing clinical trials, post-marketing clinical case reports and clinical trials. Statistical, machine learning and signal ranking approaches have been developed in extracting drug–SE pairs from free-text MEDLINE abstracts [9,22,27,30,31]. Recently, researchers began to explore other data sources for mining drug–SE associations. For example, patient EHRs have emerged as a promising resource for post-marketing drug adverse event discovery [8,19,32]. Health information available on the web and web search log data can also provide valuable information on drug side effects [16,33]. In summary, drug side effect association knowledge exists in multiple heterogeneous and complementary data sources with different formats. The effectiveness of automatic approaches in extracting drug side effect knowledge depends on both data sources and the targeted drugs or SEs. Currently, there exist no universally effective approaches to extract this knowledge from different data sources.

While the main data sources used in previous studies for drug–SE relationship extraction include FDA drug labels, abstracts of published biomedical literature and post-marketing drug safety surveillance systems, there exists a large body of untapped drug side effect knowledge that remains buried in full-text articles. In this study, we developed automatic approaches to extract anticancer drug–SE pairs from the Journal of Clinical Oncology (JCO), specifically the tables imbedded in the full text articles. JCO was established in 1983 and is the official journal of the American Society of Clinical Oncology and the leading journal in oncology. JCO articles include a variety of cancer-related research articles, including clinical trials reporting drug efficacy and toxicity in cancer patients, trial reports evaluating the effectiveness of biomarkers, clinical case reports, and meta-analysis studies, among other article types. JCO articles not only include pivotal clinical trials that have led to drug approval, but also trials that are still in investigational stages, and even failed trials. Side effect knowledge of drugs at different clinical stages is crucial to our understanding of the molecular mechanisms underlying the observed toxicities.

2. Approach

In this pilot study, we focused on extracting drug–SE pairs from tables contained in full-text JCO articles. Unlike the full-text portions of JCO articles, the tables often summarize important information such as patient characteristics, treatment outcomes, and toxicities, with each table often containing only one type of information. We downloaded a total of 13,855 full text articles from JCO and extracted 31,255 tables from the downloaded articles. Since not all tables are related to drug side effects, we first developed a support vector machine classifier to classify tables into SE-related or -unrelated categories. We then extracted drug–SE pairs from the SE-related tables. The precisions of the input lexicons (both drug lexicon and SE lexicon) are critically important for subsequent relationship extraction. We created a clean drug lexicon and a clean SE lexicon through intense manual curation efforts. Using the clean lexicons, we extracted drug–SE pairs from classified tables using dictionary-based approaches. In order to compare our study to existing research efforts in building drug–SE relationship knowledge base, we compared drug–SE pairs that we extracted from JCO tables to those from SIDER, currently the best drug side effect association database constructed from FDA package inserts [15]. To show the potential of extracted drug–SE pairs in developing systems approaches to understand the molecular mechanisms underlying the observed drug phenotypes (toxicities), we linked drugs to their corresponding gene targets, metabolism genes, and disease indications and systematically studied the correlations between drug phenotypes and their known targets, metabolism genes, and disease indications. To the best of our knowledge, this is the first study in combining table classification and relationship extraction in order to extract drug–SE pairs from full-text articles.

3. Data and methods

The system consists of the following steps: (1) we downloaded JCO articles and extracted tables from downloaded articles; (2) we created clean drug and SE lexicons; (3) we classifies tables into drug toxicity-related and non-related categories; (4) we extracted drug–SE pairs from classified tables using manually curated, clean drug and SE lexicons; (5) we compared drug–SE pairs extracted from JCO tables to those derived from FDA drug labels; and (6) we systematically analyzed the relationships between cancer drug-associated side effects with drug gene targets, drug metabolism, and disease indications (Fig. 1).

3.1. Download JCO articles and extract tables

We downloaded a total of 13,855 JCO full text articles (1.5 GB) published between 1983 and 2013 from the Case Western Reserve University Intranet. We then extracted a total of 31,255 tables from these downloaded JCO articles. A typical JCO clinical trial paper often contains multiple tables describing trial participants, treatment response, survival, prognostic factors, treatment costs, or treatment-related adverse events. While some articles contain no tables, others include many tables. For example, the article entitled "Procarbazine, Lomustine, and Vincristine (PCV) Chemotherapy for Anaplastic Astrocytoma: A Retrospective Review of Radiation Therapy Oncology Group Protocols Comparing Survival With Carmustine or PCV Adjuvant Chemotherapy" (PMID 10550132) contains as many as 14 tables.

The content associated with each extracted table includes the article title, table content, and table legend. While the side effect information is included in the table content and table legends, the drug information is often included in the article titles. A typical example of a SE-related table is shown in Fig. 2. We used the publicly available information retrieval library Lucene (http://lucene.apache.org) to create a search engine with indices created on article titles, table contents, and table legends. Each table was assigned a unique identification number.

3.2. Create clean drug and SE lexicons

3.2.1. Create clean anticancer drug lexicon

We first compiled a comprehensive drug lexicon from the Unified Medical Language System (UMLS) (2011AB version) [4] using terms with the following semantic types: "Pharmacologic