TRANSLATIONAL DRUG INTERACTION STUDY USING TEXT MINING TECHNOLOGY

Heng-Yi Wu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics & Computation,
Indiana University
October 2017

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee ________________________________

Josette Jones, Ph.D., Chair

____

Lang Li, Ph.D.

August 15, 2017 ________________________________

Mathew Palakal, Ph.D.

____

Huanmei Wu, Ph.D.

Acknowledgements

I would not have completed my Ph.D. dissertation without the tremendous supports and guidance of many people who have influenced my life and shaped my experiences both personally and professionally.

First, I especially would like to express my sincere gratitude to my mentor Dr. Lang Li. Dr. Lang Li supports me to be a Ph.D. candidate. He introduced me into the area of scientific research and shaped me to be an independent scientist. Besides these, his dignity, diligence, patience, cooperativeness and open-mindedness served as my role model. It is my honor to have Dr. Li as my Ph.D. mentor. Such an experience will surely benefit for the rest of my life. Also, I would also like to thank the members on my thesis committee, Dr. Josette Jones, Dr. Mathew Palakal, and Dr. Huanmei Wu. They provide critical evaluations and crucial suggestions to develop my dissertation. I would like to thank for Dr. Josette Jones for her kindness and enthusiastic advising in Healthcare knowledge. I would like to thank for Dr. Mathew Palakal for the knowledges of Text Mining that I learnt from him. I would like to thank for Dr. Huanmei Wu for providing expert assistances that are essential for me to complete my dissertation.

Besides my committee members, I would like to give thanks to my collaborators who were willing to provide data, comments, inputs and support for project success and scientific discovery. I would like to thank for Dr. Sara K. Quinney, Dr. Desta Zeruesenay, Dr. Lijun Cheng, Dr. Pengyue Zhang, Chien-Wei Chiang, Dr. Zhiping Wang, Shijun Zhang and other researchers who provide contributions in the projects to investigate the drug-drug interaction research. Last but not least, I would like to thank all my professors and friends. Finally and the most, I would like to express my appreciation to my family, especially to my wife, Hsiao-Yun Huang. I remembered how much support you gave me in the past few years during my graduate study. It is your loves and encouragements make this endeavor possible.

Heng-Yi Wu

TRANSLATIONAL DRUG INTERACTION STUDY USING TEXT MINING TECHNOLOGY

Drug-Drug Interaction (DDI) is one of the major causes of adverse drug reaction (ADR) and has been demonstrated to threat public health. It causes an estimated 195,000 hospitalizations and 74,000 emergency room visits each year in the USA alone. Current DDI research aims to investigate different scopes of drug interactions: molecular level of pharmacogenetics interaction (PG), pharmacokinetics interaction (PK), and clinical pharmacodynamics consequences (PD). All three types of experiments are important, but they are playing different roles for DDI research. As diverse disciplines and varied studies are involved, interaction evidence is often not available cross all three types of evidence, which create knowledge gaps and these gaps hinder both DDI and pharmacogenetics research.

In this dissertation, we proposed to distinguish the three types of DDI evidence (in vitro PK, in vivo PK, and clinical PD studies) and identify all knowledge gaps in experimental evidence for them. This is a collective intelligence effort, whereby a text mining tool will be developed for the large-scale mining and analysis of drug-interaction information such that it can be applied to retrieve, categorize, and extract the information of DDI from published literature available on PubMed. To this end, three tasks will be done in this research work: First, the needed lexica, ontology, and corpora for distinguishing three different types of studies were prepared. Despite the lexica prepared in this work, a comprehensive dictionary for drug metabolites or reaction, which is critical to in vitro PK study, is still lacking in pubic databases. Thus, second, a name entity recognition tool will be proposed to identify drug metabolites and reaction in free text. Third, text mining tools for retrieving DDI articles and extracting DDI evidence are developed. In this work, the knowledge gaps cross all three types of DDI evidence can be identified and the gaps between knowledge of molecular mechanisms underlying DDI and their clinical consequences can be closed with the result of DDI prediction using the retrieved drug-gene interaction information such that we can exemplify how the tools and methods can advance DDI pharmacogenetics research.

Josette Jones, Ph.D., Chair

Table of contents

List of Tables

List of Figures

List of Abbreviations

Absorption, Disposition, Metabolism, Excretion, and Transportation (ADMET)

Adverse Drug Reaction (ADR)

Area Under the plasma Concentration-time curve (AUC)

Cascading Style Sheet (CSS)

Clearance (CL)

Comprehensive Perl Archive Network (CPAN)

Cytochrome P450 (CYP)

Drug Interaction Knowledge-Base (DIKB)

Drug-Drug Interactions (DDIs)

Emergency Department (ED)

False Negative (FN)

False Positive (FP)

F-measure (F1)

Hazard Ratio (HR)

Human Metabolome Database (HMDB)

Inclusion-Exclusion Criteria (IEC)

Information Extraction (IE)

Information Retrieval (IR)

Levels-of-Evidence (LOEs)

Mammalian Phenotype (MP)

Mass-Spectrometry (MS)

Metabolic Drug-Drug Interaction (M-DDI)

Metabolism & Transport Drug Interaction Database (DIDB)

Named Entity Recognition (NER)

Natural Language Processing (NLP)

Nuclear Magnetic Resonance (NMR)

Odds Ratio (OR)

Part-of-Speech (PoS)

Pharmacodynamics (PD)

Pharmacogenetic (PG)

Pharmacokinetics (PK)

Physiologically-Based Pharmacokinetics (PBPK)

Precision (P)

Protein-Protein Interaction (PPI)

Recall (R)

Suggested Ontology for Pharmacogenomics (SOPHARM)

Support Vector Machines (SVM)

True Positive (TP)

Food and Drug Administration (FDA)

Chapter 1. Introduction

1.1 Adverse Drug Reaction and Drug-Drug Interaction

Adverse drug reaction (ADR) is one of major causes of morbidity and mortality occurring in clinical care every year. To investigate the crucial problem, US Food and Drug Administration (FDA) found that more than 40% US population is prescribed more than four medications at a single time, which makes more susceptible to ADR (Knapp & Tomita, 1987). A literature search in Medline and Embase database from 1990 to 2006 showed that drug-drug interactions (DDIs) were held responsible for 0.054% of the emergency department (ED) visits, 0.57% of the hospital admissions and 0.12% of the re-hospitalizations (M. L. Becker et al., 2007). It is possible that drug interaction can be beneficial or detrimental. The use of multiple drugs might provide synergism such as increasing the efficacy of therapeutic effect, decreasing dosage but holding the same efficacy to avoid toxicity, or minimizing the drug resistance (Chou, 2006). However, we have more interests in the investigation of negative interaction because pathological significance is often unexpected and hard to be diagnosed. To predispose DDIs, the importance of high risk factors like age, polypharmacy and genetic polymorphisms should be carefully evaluated (Magro, Moretti, & Leone, 2012). In the elder population, DDIs account form 4.8% of the hospital admissions, which is much higher than the proportion of DDI victims within the total population. The reason is directed to the abatement of liver metabolism or kidney function (Juurlink, Mamdani, Kopp, Laupacis, & Redelmeier, 2003; Merle, Laroche, Dantoine, & Charmes, 2005). Genetic polymorphism has profound influence on enzyme function, which might results in increased drug metabolism and absence of drug response. Evidences (Johansson & Ingelman-Sundberg, 2011) suggested that patients affected by genetic polymorphisms will experience severe toxicities upon drug intake.

For economic aspect, the problem of DDI effect or co-medication effect has scaled such heights that it has even led to withdrawing of drugs from the market after approval. The 1990s saw the withdrawal of more than 11 drugs as shown in (Ajayi, Sun, & Perry, 2000). In 2007, the biopharmaceutical industry invested roundabout $58.8 billion for the research and development as the withdrawing of drugs (DiMasi & Grabowski, 2007) is a major setback to the industry as the deployment of a single drug compound is estimated at $200 million.

1.2 Drug-Drug Interaction Mechanisms and In Vitro and In Vivo Drug Interaction Studies

Drug-drug interaction can result when a substance affects the activity of a drug or its metabolites when these two drugs are administrated at the same time. The simultaneous administration of two drugs, which causes synergistic or antagonistic effect, might lead to the alternation of medication effectiveness or some harming effects on patient body. Those potential influences on human body should be noticed to prevent from a high risk of multiple interactions because the number of approved drugs increases. To preclude the possibility of hazardous interaction, understanding the significant scientific principles or mechanisms of drug-drug interaction is important.

Due to the continued growth in drug development and the insight into molecular biology, we come to realize that transporter and enzyme played an important role in drug elimination, which inspired a clue to dig the mechanisms surrounding drug-drug interaction. In brief, there are two major molecular mechanisms of drug-drug interaction, enzyme-based drug metabolism and transporter-based drug transportation (Pang, Rodrigues, & Peter, 2010). To study DDI with P450-mediated drug metabolism, the investigation of how a drug inhibit or induce another drug can learn from how this drug is metabolized as well as which enzymes are catalyzing the main metabolic pathway. If an enzyme that is responsible for the metabolism of one drug is induced or inhibited by another drug, then the bioavailability of original drug will be changed, which might result in being toxic or less effective. For transporter-based drug transportation, transporter is important to drug deposition. Only drugs can be metabolized after they are transported into liver cells. To know how transporter-mediated DDI happen, the knowledge of which transporters inhibited by the investigational drug or the affinity of substrate with drug transporter can also suggest a potential for drug-drug interaction (Use, 2012).

There are two basic types of drug interaction, pharmacokinetics (PK) and pharmacodynamics (PD). In short, PK investigates the activity of drug combinations with drug absorption, disposition, metabolism, excretion, and transportation (ADMET), which describes how these five criteria influence drug level (concentration). Pharmacokinetically speaking, potentiative or reductive combinations are respectively correlated to positive or negative modulation of drug transport, permeation, distribution, localization, or metabolism. Potentiative modulation of drug transport will enhance drug absorption via the disruption of transport carrier, increase drug concentration in plasma by inhibiting metabolic process, and stimulate or inhibit the metabolism of drugs into active or inactive form. On the other hand, reductive modulation provides contrasting perspectives to potentiative modulation. The reductive modulation of drug transport typically blocks drug absorption, decreases drug concentrantion in plasma, and reduces drug metabolism activity (Jia et al., 2009). Those information brings to systematically investigate the physiological and biochemical mechanisms of drug exposure in multiple tissue types, cells, animals, and human subjects (M. Rowland & Tozer, 1995), which links preclinical and clinical phase of drug development. If the PK can be interpreted as the dose-concentration relationship, pharmacodynamics (PD) can be defined as the mechanism of drug action and relationship between drug concentration and effect. A drug's pharmacodynamics effect ranges widely from the molecular signals (such as its targets or downstream biomarkers) to clinical symptoms (such as the efficacy or side effect endpoints). To classify its therapeutic effects, it can be synergistic, additive, or antagonistic if the effect is greater than, equal to, or less than the summed effects of drug combinations (Jia et al., 2009).

As stated in the previous section, the complicated transporter-enzyme interplay in the deposition of drug leads to the difficulty for the identification of DDIs in drug administration and drug development. Thus, understanding the molecular mechanism underlying different types of drug interaction could facilitate the discovery of novel DDI. Recently, in-vitro technologies can qualitatively provide an insight into the potential DDI based on the observation of enzyme kinetic parameters. Via ADME screening efforts as well as the assessment of CYP inhibition, the choice of test compound inhibiting the metabolism of one probe substrate for an enzyme in the in-vitro experiment can be fulfilled to carry out the prediction of in-vivo DDI. (Wienkers & Heath, 2005) addressed basic principles of in-vitro inhibition prediction underlying the generation of in-vitro drug metabolism data and suggested several factors that introduced error or uncertainty into a quantitative prediction of in-vivo DDI based on in-vitro derived PK parameters. In (Rostami-Hodjegan & Tucker, 2004), three factors authors recommended for the ideal model to predict metabolic drug-drug interaction (M-DDI) should be an accurate measurement in the average increase in the area under the plasma concentration-time curve (AUC) of a victim drug following administration of a perpetrator drug, the plasma binding displacement interaction and the impact of the concentration-time profile of the inhibitor. To evaluate the potential for M-DDI, (Rostami-Hodjegan & Tucker, 2004) developed an in silico software SIMCYP, which incorporate extensive data on demographics, disease states, anatomical, physiological, genetic, biochemical variables, and input of information on in-vitro drug metabolism and transport.

1.3 Computational Drug Interaction Prediction and Drug Interaction Text Mining

1.3.1 Overview of Computational Drug Interaction Prediction

The evaluation of the potential risk of DDI is of importance in patient safety since drug-drug interactions can raise the danger of patients and the cost of healthcare system. According to the guidance for industry from the Food and Drug Administration (Huang, Temple, Throckmorton, & Lesko, 2007), study design, data analysis and implication for dosing and labeling are suggested to deal with drug interaction studies. When studying DDI for a new drug, it usually begins with in vitro study to determine whether a drug is a substrate, inhibitor, or inducer of metabolizing enzymes. The consequence of in vitro investigations can serve as an evidence to screen out the candidate potential drug pairs for additional in vivo study. To conduct an in vivo DDI study for an investigating drug, a quantitative analysis to mathematically describe the kinetics of drug metabolism involved in ADME process is needed. The basic model for the initial assessment of DDI based on in vitro and in vivo studies can be achieved by physiologically-based pharmacokinetics (PBPK) modeling. From published in vitro experiments and in vivo studies, (Chien et al., 2006; Li, 2007; Li et al., 2007; Quinney et al., 2010; Wang, Kim, Quinney, Zhou, & Li, 2010; Yu, Kim, Wang, Hall, & Li, 2008; J. Zhou, Qin, Sara, et al., 2009; J. Zhou, Qin, Yu, et al., 2009) had developed Bayesian models and computational algorithms to construct physiological based pharmacokinetic (PBPK) models for DDI prediction.

Another common way to explore novel DDI is literature-based discovery. The hidden knowledge among information embedded in publications can be dug out through finding connections between articles. To this end, many researchers took advantage of some commercial or public databases as resource, such as Metabolism & Transport Drug Interaction Database (DIDB) (Hachad, Ragueneau-Majlessi, & Levy, 2010), PharGKB (Hewett et al., 2002), and DrugBank (Knox et al., 2011) which provided extensive lists of DDI information published in articles, clinical files or biomedical research reports. (Gottlieb, Stein, Oron, Ruppin, & Sharan, 2012) proposed a computational framework INDI to infer and explore DDI by calculating similarity measurement between drug pair via diverse feature measurements i.e chemical-based, ligand-based, side-effect based, annotation-based, sequence-based, and etc. However, the problem of data inconsistancy arose when using different databases. Some significant scientific evidences associated with DDI are limited or lacking in some existing databases. This deficiency is hard to prevent because the tasks of data collections are manually accomplished by different research groups or professional experts. To conquer this problem, employing the technologies from Information Retrieval (IR) or Natural Language Processing (NLP) can be a solution to help extract data more efficiently and consistently.

1.3.2 Biomedical Text Mining

Text mining refers to the process of deriving high-quality information from text, which relies on Natural Language Processing (NLP). To translate the text into computer-readable language, there are some basic steps of NLP (Nadkarni, Ohno-Machado, & Chapman, 2011), including sentence splitting, tokenization, part-of-speech, name entity recognition, shallow parsing, and syntactic parsing. In this section, we do not go into the details of techniques for NLP tools. The attentions will be paid more on the tasks of corpus construction, information retrieval (IR) or information extraction (IE), which employs highly scalable statistics-based techniques to index and search large volume of text efficiently.

Extracting facts from texts is the goal of text mining systems. The range of extraction tasks can be narrow from retrieving potentially relevant articles by sophisticated keyword search or classifying papers into different ontological types (IR), recognizing biological entities or concepts in text, detecting relations between biological entities (IE) and broader to document summarization or question answering (beyond IE) (Zweigenbaum, Demner-Fushman, Yu, & Cohen, 2007). To fulfill those tasks in biomedical domain, name entity recognition is an initial processing step because the significant knowledge is usually centered on the mechanism of biological activities which are described by nominalized verbs and nouns within sentences. Therefore, to identify text that satisfies various types of information needs is an important first step toward accurate text mining. But how to utilize the identified entities for improving text mining is challenging. One solution to this problem is an annotated corpus. The corpus annotated with such information allows real usage within text to be taken into account. The annotated sentence then can be represented in syntactic and semantic format, which shows the different levels of scientific characteristics. However, the strategy of constructing corpus is diverse. It differs with the purpose of text mining task and the methodology we used in extracting information. (J. D. Kim, Ohta, Tateisi, & Tsujii, 2003) introduces GENIA corpus with linguistically rich annotations for biomedical articles. The value of GENIA corpus comes from its annotations. All biologically meaningful terms are semantically annotated with descriptors from GENIA ontology. (Wilbur, Rzhetsky, & Shatkay, 2006) suggests the basic guideline and criteria of corpus construction and annotation task for facilitating the training components of IE system by using machine learning method. Another value of annotated corpus is being a gold standard that facilitates the evaluation of approach. The success of practical applications crucially depends on the quality of extraction results, which against the access of gold standard reference.

1.3.3 Relationship Extraction

Within information extraction (IE) methods, we are more interesting in relationship extraction. The goal of relationship extraction is to detect the prespecified type of relation between a pair of entities of given types. A relation is typically represented as a pair of entities, linked by an arc that is either directed or undirected. The arc is given a label usually corresponding to a semantic type. In biology, the type of entities can be very specific such gene, protein, or drug, while the type of relationship can be referred from some particular verbs, including transcribe, repress, or inhibit.

To effectively extract relationship, analysis of sentence structure is necessary. The use of semantic processing or deep parsing techniques that analyze both the syntactic and semantic structure of texts can benefit relation extraction. Several approaches had been reported in literature to extract relation of interest. Generally, there are three main approaches for relationship extraction: co-occurrence-based, rule-based, and machine-learning based approaches. (Muller, Kenny, & Sternberg, 2004) employ co-occurrence-based method, which is the simplest way to capture relationships relying on co-occurrence of two entities to derive a relation. Rule-based approaches (Feldman, Regev, Finkelstein-Landau, Hurvitz, & Kogan, 2002; K., R., & R., 2007) are to take advantage of linguistic technology to grasp syntactic structure or semantic meaning for understanding the relationship from the unstructured text. (Feldman et al., 2002) employed a NP1-Verb-NP2 template to identify the relation between two domain-specific entities. (K. et al., 2007) constructed a set of domain-specific rules and applies them to dependency parse tree to capture different forms of expressing a given relationship. Finally, classifiers using machine-learning approaches such as Support Vector Machines (SVM) (Qian & Zhou, 2012) often used for relation extraction. This method needs laborious efforts to define grammars or rules and text in training dataset is manually tagged by a human expert. This text mining method use the training data to automatically learn the "rules" so it can mine wanted information or identify the necessary knowledge (Airola et al., 2008; Chen, Liu, & Manderick, 2009; Pyysalo et al., 2008; Tikk, Thomas, Palaga, Hakenberg, & Leser, 2010).

The comparison among different methods is not easy because each method obtains its inherited pros and cons. Co-occurrence method provides highest recall but poor precision among three. A large amount of false positive relations are returned whenever the sentence is sophisticated with more than two entities or two key entities co-occurred in each single sentence but it does not state their relationship. Thus, co-occurrence method is more suitable to use as a simple baseline method for performance comparison. Rule-based method achieves better precision in extracting binary relationships due to the more precise rule conditions for defining relationship. But when it meets the complex sentence with various coordinates and relational clauses, the performance turn down obviously (D. Zhou & He, 2008). In general, machine learning-based method performs the best among methods. As an evidence in BioCreative challenge (M., F., & A., 2009), the frameworks using supervised machine learning algorithm outperformed existing methods in detecting protein-protein interaction. One important advantage is system can predict categories for unseen samples. However, this advantage is heavily relying on annotated corpus (Segura-Bedmar, Martinez, & de Pablo-Sanchez, 2011b). Therefore, it can be also a big disadvantage because of the need for huge learning set.

1.3.4 Literature Review for Extracting Drug-Drug Interaction

Different approaches had been developed for extracting biomedical relationships such as protein-protein interactions. From the experience of previous researches centered on protein-protein interaction (PPI) (Airola et al., 2008; Chen et al., 2009; Pyysalo et al., 2008; Qian & Zhou, 2012; Tikk et al., 2010), few approaches have been proposed to the problem of detecting DDI. To promote the development of DDI extraction tools, DDIExtraction 2011, the first challenge task on drug-drug interaction extraction, was held in 2011 at Spain. In this workshop, they provided evidence for the most effective methods available to solve specific problems and reveal the performance on these problems. In competition, most participants proposed systems using classifiers SVM or RLS. Their choices verified that machine learning can outperform other methods in relation extraction. Observed from results, approaches based on kernel methods achieved better performance than the classical feature-based methods (Segura-Bedmar, Martınez, & Sánchez-Cisneros, 2011a). Thus, the advantages of kernel-based method using machine learning classifier are spotlighted in this workshop.

In literature, some articles are outstanding in DDI extraction. The co-chairs of DDIExtraction 2011 (I. Segura-Bedmar, P. Martinez, et al., 2011b) proposed a hybrid approach, which combines shallow parsing and pattern matching to extract relation between drugs based on annotated corpus. It utilizes the proposed syntactic patterns to split the sentence into clauses from which relations are extracted by matching patterns. The ability of dealing with complicated sentence is the advantage of this method. Complexity can be diminished by separating a long sentence into simplified clauses and by the detection of the apposition and coordinate structure. But there is one gap in the extraction of DDI information if used in pharmacokinetics or pharmacogenetics articles. Only exploring DDI based on literal denotation will lead to the missing detection of actual DDI information due to the lacking of scientific knowledge. In (B. Percha, Garten, & Altman, 2012), DDIs are identified by aggregating drug-gene interactions which are extracted via rule-based method. The extracted interactions are then normalized and mapped into their standardized ontology to form the semantics network. The network could be useful to find potential DDI and the types of relationship.

Differed from (B. Percha et al., 2012) extracting DDI via the perspective of pharmacogenetics, (L. Tari, Anwar, Liang, Cai, & Baral, 2010) developed a method that combined text mining and automated reasoning to extract DDIs with the support of enzyme domain knowledge. This work focused on the discovery of DDIs through the integration of biological knowledge with biological facts from published literature by using text mining and automated reasoning approach. The novelty of this method is not only to extract DDI pairs from explicitly mentioned text like the typical extraction approach (I Segura-Bedmar, Martínez, P., de Pablo-Sánchez, C., 2011) but also enable to discover potential DDIs by automated reasoning. In this work, the biological knowledge includes the two relationships between drugs and metabolism enzymes, including how the drug is catalyzed by an enzyme and how the induction or inhibition of metabolism enzymes by another drug. On the other hand, the biological facts are curated from literature. Since the biological factors curated from literature can meet the relationship of biological knowledge and identify which enzyme is induced or inhibited by specific drug or the enzymes which are responsible for the metabolism of that drug, the DDI evidence can be acquired through the use of logic representation of the domain knowledge and automated reasoning. The distinct capability of extracting drug-enzyme interactions to infer DDIs that are not explicitly stated in text through automated reasoning is the beauty of this work.

Similar to (L. Tari, Anwar, S., Liang, S., Cai, J., Baral, C., 2010), Boyce (R. Boyce, Collins, Horn, & Kalet, 2009b) applied a pharmacological knowledge base called Drug Interaction Knowledge-base (DIKB) to predict significant interactions in a validation set. Differed from

Tari’s work, this approach obtains the ability of leveraging the available scientific evidence within a domain for supporting important drug package insertions. It benefits from a key component of the system, a rule-based theory of how drugs interact by metabolic inhibition. Based on the background knowledge in DIKB and the rule-based theory, it distinguished between assertion instance (a clear statement of some property) and insertion types (such as X substrate-of Y). Therefore, experts can calculate their confidence scores for drug-mechanism assertions by defining combinations of evidence types from each assertion type in the system’s evidence taxonomy. Then, the system ranks the evidence-type combinations by the relative amount of confidence score, called levels-of-evidence (LOEs).

1.4 Knowledge Gap for Drug-Drug Interaction Studies

Since the attention paid more to drug interactions has increased, current research aims to investigate different scopes of drug interactions, including pharmacogenetics (PG), pharmacokinetics (PK), and pharmacodynamics (PD) (R. Boyce, Collins, Horn, & Kalet, 2009a; R. Boyce et al., 2009b; Hennessy & Flockhart, 2012). All three types of studies are critical, but they get insight into the truth in different aspects. PG studies explore the inherited genetic difference in drug metabolism pathways, which can influence individual response to drugs (Klotz, 2007). PK studies examine how an organism affects a drug in terms of absorption, distribution, metabolism, and excretion; whereas PD studies is used to know how the drugs affect the organism (Knollmann, 2011). As diverse disciplines and varied studies are involved, interaction evidence is often not available cross all three types of evidence, which create knowledge gaps and these gaps hinder both DDI and pharmacogenetics research.

Owing to the development of automatic text mining technology, it should be a good opportunity to aggregate and tap into our collective scientific knowledge from biomedical literature and potentiate translational drug interaction research. However, most of current works largely treat DDI as if it is studied at a single scale. For a given drug interaction, the three types of evidence are typically not reported together and are often not all available. We refer to lack of evidence along any of the three types as a knowledge gap. As current automatic extraction methods treat all DDI reports similarly, without distinguishing experimental evidence at different scales, the problem of knowledge gaps in DDI evidence has not been hitherto considered or addressed. This may explain the relatively low performance so far achieved by general-purpose DDI extraction (e.g.F1≈0.34-0.75) (Segura-Bedmar, Crespo, de Pablo-Sanchez, & Martinez, 2010; I. Segura-Bedmar, Martínez, P., de Pablo-Sánchez, C., 2011).

1.5 Proposed Solutions

To conquer this issue, three aims are proposed to close such gaps in DDI evidence by using informatics methods to integrate and tap into our collective scientific knowledge.

1.5.1 Scope of Aim 1: Lexica, Ontology, and Corpora for DDI Evidence

The purpose of this aim is to prepare the required resources that can be used to distinguish the three types of DDI evidence (in vitro PK, clinical PK, and clinical PD studies). In this work, DDI lexica, ontology and corpora pertaining to three types of studies are developed. a) The lexica contain terminologies pertaining to drugs, study design, drug/enzyme/transporter relationships, ADR, DDI models and their parameters. b) A comprehensive PK ontology was constructed to build the relationship between concepts for in vitro PK and in vivo PK studies, which can provide background knowledge for text mining tools. c) Two different corpora are prepared: The first corpus is constructed to be the golden standard corpus. The second corpus is prepared to be the training data for large-scale text mining. For golden standard corpus, using our ontology and lexica, DDI information in entity level, sentence level, and entity-relationship level were indicated and annotated with the type of evidence. For training corpus, 300 DDI relevant abstracts for each evidence type and 800 DDI irrelevant abstracts including single drug, drug-nutrition, PD related and randomly selected abstracts are manually curated from Medline database. The detail of this aim will be described in Chapter 2 and Chapter 3.

1.5.2 Scope of Aim 2: Named Entity Recognition for Drug Metabolite and Reaction

Although there exist many well-established dictionaries for drug names, such as DrugBank, MeSH terms, Rx-Norm, NDC, PubChem, etc, the existing resources that contain the terminologies of drug metabolites are very limited. To enrich the information available in public database and differentiate from metabolome, the purpose of this aim is to propose a NER tool for drug metabolite and reaction from biomedical literature. In this work, 1) metabolite-rich corpora and 2) a comprehensive lexical repository, including a drug name dictionary and the lexicon of general nomenclatures (prefix/suffix) for drug metabolite are constructed. 2) A named entity recognition tool to annotate drug metabolites and reactions in scientific text, utilizing an integrated dictionary and machine learning algorithms, is developed. The detail of this aim will be described in Chapter 4.

1.5.3 Scope of Aim 3: Evidence-based Text Mining Tools for DDI

The purpose of this aim is to develop a text mining pipeline for large-scale mining and analysis of drug-interaction information such that it can be applied to retrieve, categorize, and extract the information of investigated DDI pairs from published literature available on PubMed and to identify all knowledge gaps in experimental evidence among them. To implement a large-scale screening on whole Medline database, training datasets consist of hundreds of positive and negative abstracts. Second, we develop a suite of text mining tools to explicitly identify each type of DDI evidence, namely in vitro PK, in vivo PK and clinical PD. The suite consists of: 1) Information Retrieval (IR): Document-level classifiers for retrieval of PubMed abstracts presenting each type of DDI evidence; 2) Information extraction (IE) tools for tagging drugs, drug metabolites, interaction verbs, and experimental endpoints and identifying interacting drug pairs from abstracts. 3) Identify knowledge gap of DDI Evidence: Identify and automatically annotate interaction-evidence obtained from PubMed abstracts based on their likelihood to convey reliable evidence of each type. Display the Venn Diagram of DDI evidence cross all types of studies and ranked and annotated list to experts who will use it to select drug-pairs with strong in-vivo and clinical interaction evidence but insufficient in-vitro evidence. The detail of this aim will be described in Chapter 5.

1.5.4 The Theoretical Model for This Project

As shown in Figure 1.1, Aim 1 provides the resource of a drug name dictionary and In Vitro PK corpus to Aim 2 and offers the lexica and corpus of DDI evidences for AIM 3. With the resource that Aim 1 provides, Aim 2 proposed a NER tool to discover drug metabolite and reaction from literature, which is valuable to enrich the limited resource in public database, improve the identification of drug metabolite in pharmacological articles, and increase the coverage of information extraction. For Aim 3, the lexica, ontology and corpus from Aim 1 are critical components for text mining. The lexica and ontology can used to annotate all aspects of DDI and create features to distinguish different DDI evidences from diverse studies. The corpus facilitated DDI text mining from the literature.

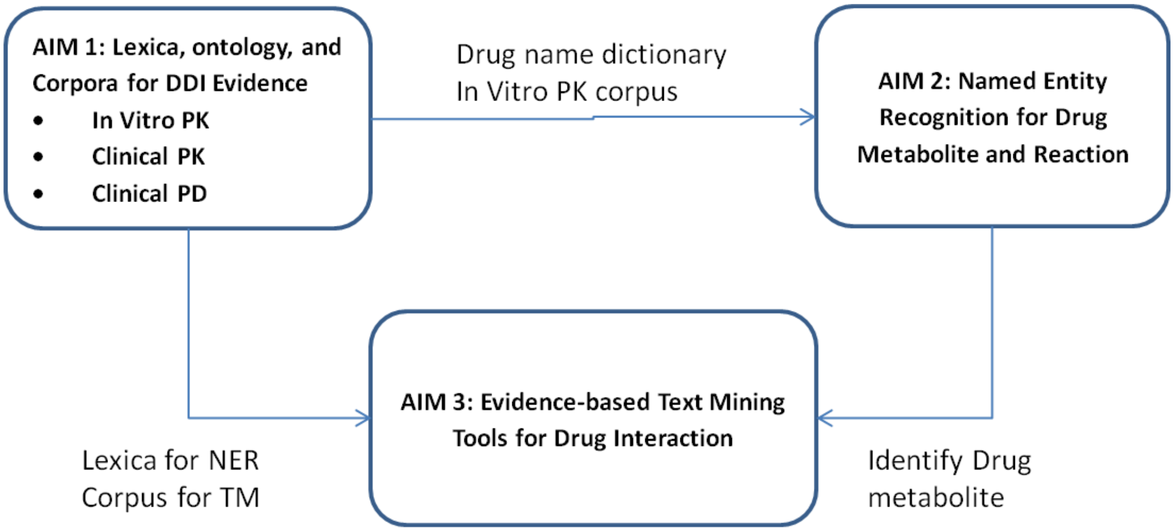

AIM 1: Lexica, ontology, and Corpora for DDI Evidence
• In Vitro PK
• Clinical PK
• Clinical PD
Drug name dictionary
In Vitro PK corpus
AIM 2: Named Entity Recognition for Drug Metabolite and Reaction
AIM 3: Evidence-based Text Mining Tools for Drug Interaction
Lexica for NER
Corpus for TM
Identify Drug metabolite

Figure 1.1 Theoretical Model of This Project

1.6 Impact of This Project

1.6.1 Impact of Aim 1

In this work, lexica, ontology, and corpus are constructed. By bringing terminologies of diverse types of experiments together from different databases into an unified resource, the integrated lexica allow researchers from diverse background and disciplines to conduct data collection and distinguish DDI evidence from different types of studies. The DDI ontology is built to interpret raw text in biomedical articles by the descriptors with a standardized format and organized into hierarchical structure. Such an advantage allows complex text to be represented with semantic and consistent manner. DDI corpus construction is an important first step towards more accurate text mining, which allows utilizing scarce resources to annotate text as a training corpus for machine learning. In addition, the corpus can be widely useful to the biomedical data-mining research community for exploring the information of drug interaction and lead to the development of practical and useful resources. Overall, this aim contributes an integrated lexicon for collecting terminologies of diverse DDI studies, an ontology for interpreting DDI terminologies with a semantic format, and a set of corpus for implementing DDI text mining.

1.6.2 Impact of Aim 2

The importance of drug metabolite to DDI: Drug metabolism, distribution, and excretion are the primary pharmacokinetics research areas. A drug's pharmacokinetics (PK) involves not only the parent compound, but also its metabolites (Malcolm Rowland, Tozer, & Rowland, 2011). In some instances, an active drug metabolite can retain enough or even dominate its intrinsic activity at target receptor and contribute to the pharmacological effects. Certain drugs such as codeine and losartan have active metabolites (morphine and EXP3174 respectively) that are responsible for more therapeutic action than their parent drugs (Obach, 2013). On the other hand, pro-drugs, formulated in an inactive form, are designed to be metabolized inside the body to form the active drugs (Hacker, Messer II, & Bachmann, 2009). A salient example is tamoxifen. As a pro-drug, tamoxifen itself is not an active compound to treat breast cancer. Instead, his metabolites, 4-OH-tamoxifen and endoxifen are potent inhibitors to estrogen alpha (Desta, Ward, Soukhova, & Flockhart, 2004; Johnson et al., 2004; Lee, Ward, Desta, Flockhart, & Jones, 2003; Stearns et al., 2003). Drug interactions always make the PK research even more complicated. One drug's metabolism, distribution, and excretion can be changed by another drug, and sometimes the other drug's metabolites, too. A notable example is itraconazole. Itraconazole itself is a potent CYP3A inhibitor, so are its metabolites, such as hydroxy-itraconazole, keto-itraconazole, and N-desalkyl-itraconazole (Isoherranen, Kunze, Allen, Nelson, & Thummel, 2004). Therefore, all of the CYP3A substrates' metabolism, such as midazolam, triazolam, and etc, are inhibited by itraconazole and its metabolites, if they are taken together. Pharmacogenetics, another forefront of pharmacology research, also has major impact on drug metabolism products. Using the previous tamoxifen example, tamoxifen active metabolite, endoxifen, is generated through CYP2D6 enzyme. Among breast cancer patients with CYP2D6 loss functional variants (e.g. *4, *5, and*10), the patients usually have very limited tamoxifen metabolite, endoxifen concentration (Stearns et al., 2003). Therefore, drug metabolites and their parent drugs are equally important in pharmacokinetics research.

Improve the deficiency of drug metabolite in public database: Although there are a number of well-established dictionaries for drug names, such as DrugBank, MeSH terms, Rx-Norm, NDC, PubChem, etc, there is very limited naming system for drug metabolites. In particular, we want to make a distinction between metabolome and drug metabolites. The metabolome is considered to be the collection of all metabolites in a biological cell, tissue, organ, or organism. Metabolome may include both endogenous metabolites that are naturally produced by an organism (such as amino acids, organic acids, nucleic acids, fatty acids, amines, sugars, vitamins, co-factors, pigments, antibiotics, etc.) as well as exogenous chemicals (such as drugs, environmental contaminants, food additives, toxins and other xenobiotics) that are not naturally produced by an organism (Nordstrom, O'Maille, Qin, & Siuzdak, 2006; Wishart, 2007). Therefore, ideally, metabolome shall include drug metabolites. However, due to the limitation of Mass-Spectrometry (MS) or Nuclear Magnetic Resonance (NMR) biotechnologies, metabolome studies and drug metabolisms studies are conducted in very different protocols. Drug metabolites rarely can be found from metabolome studies, either because they are different metabolites, or their names are totally different. For instance, the highly populated Human Metabolome

Database (HMDB) (Wishart et al., 2013), in which reports data on >29,000 endogenous metabolites, 2485 drugs, and 948 drug metabolites, is not known in the pharmacokinetics research fields at all. Other examples are DrugBank 4.0 (Law et al., 2014) and ChEBI (Degtyarenko et al., 2008), comprising of only 1,445 and 111 drug metabolites respectively, which are much less than the total number of generic drugs (8,184).

Contribution to pharmacology research community: The main purpose of Aim 2 is to construct a tool for recognizing drug metabolite and reaction from text. In this task, an annotation guideline, a gold standard corpus and a NER tool for drug metabolite were constructed. The annotation guideline provides well-defined rules for annotators to recognize and differentiate drug metabolite from metabolome and classify the types of drug metabolite representation in text. With the proposed annotation guideline, a high quality gold standard corpus is finely annotated by three domain experts. This corpus covers all types of drug metabolite representations, which facilitates the next step machine learning to discover drug metabolites from unknown text. With such a gold standard corpus, an innovative drug metabolite NER tool was developed to capture drug metabolite and reaction defined in the proposed annotation guideline. The main impact of this project is to enrich the limited resource of drug metabolite in public database, improve the identification of drug metabolite in pharmacological articles, and increase the coverage of information extraction for drug interaction.

1.6.3 Impact of Aim 3

Explore evidence at multiple scales of drug-drug interaction: Evidence-based assessment of published drug-drug interaction information appears to be feasible and would help clinicians and patients. Assessment of this information on a large scale requires significant resources and the resolution of a number to technical issues. To meet this problem, recent interest in automatic DDI identification focuses primarily on extraction of interacting drug-pairs from multiple resources, including biomedical literature, EHRs, and FDA labeling (Segura-Bedmar, Martinez, & de Pablo-Sanchez, 2011a; L. Tari, Anwar, S., Liang, S., Cai, J., Baral, C., 2010). However, current work largely treats DDI as if it is studied at a single scale. To expedite progress through DDI information it is essential to note that experimental evidence for DDI ranges in scale from intracellular biochemistry to human populations, and can be categorized into three main types (Hennessy & Flockhart, 2012): in vitro, in vivo and clinical. While clinical evidence may ultimately trigger DDI alerts, it does not provide insight into molecular mechanisms underlying the interactions. The latter is vital for determining drug absorption, distribution, metabolism, excretion, and targeting, which enable investigating less risky alternatives.

Due to the diverse disciplines involved in DDI studies (Hennessy & Flockhart, 2012), for a given drug interaction, the three types of evidence are typically not reported together and are often not all available. We refer to lack of evidence along any of the three types as a knowledge gap. As current automatic extraction methods treat all DDI reports similarly, without distinguishing experimental evidence at different scales, the problem of knowledge gaps in DDI evidence has not been hitherto considered or addressed. This may explain the relatively low performance so far achieved by general-purpose DDI extraction (e.g.F1≈0.34-0.75) (I. Segura-Bedmar, P. Martinez, et al., 2011a).

Large-scale, comprehensive text-mining for identifying knowledge gaps for DDI: Biomedical text from a broad variety of publications and multiple sources forms an increasingly important basis for integrating our collective scientific knowledge and enabling knowledge discovery (Albright et al., 2013; Savova et al., 2010). This provides the opportunity to tackle the problem of missing DDI evidence and integrate dispersed published evidence to close knowledge gaps in experimental DDI evidence. In contrast to the traditional drug interaction research, which usually focuses on one drug a time, we propose to utilize text mining to approach the problem at the entire "bibliome" scale. That is, we propose to apply our tools to all published experimental reports in PubMed, the 20,446 available FDA prescription-drug labels (R. D. Boyce, Collins, Clayton, Kloke, & Horn, 2012), as well as take advantage of other available resources such as DrugBank (Segura-Bedmar I, 2010) in our lexica- and tool-development.

Translation between molecular and clinical research: Clinical decisions typically stem from in vivo and clinical evidence. However, studying molecular interaction mechanisms in vitro is essential for understanding the hazards of specific drugs given certain genetic polymorphisms and for exploring potential alternative treatments. Since translational DDI research aims to link between knowledge of molecular mechanisms underlying DDI and their clinical consequences, it is of paramount importance to identify knowledge gaps that prevent such translation. Therefore, an essential and fundamental step toward developing reliable clinical decision systems requires Comprehensive drug-interaction evidence of all three types.

Indeed, the quality of outcome in the proposed work might be imperfect and there are a great deal of false positives in the result. Even the DDI pairs are truly mentioned in a specific study, they can be only considered as potential drug interactions. A complete evaluation of the potential for the drug interaction is needed to decide whether the potential interaction exists and, if so, whether the potential for such interaction indicates the needs for dosage adjustment or additional therapeutic monitoring. However, the value of this work is that the aggregative result can provide an exciting opportunity to promote translation of molecular to clinical research. We will be able to follow-up by experimentally testing potentially problematic DDIs that our proposed research uncovers. This will further facilitate the downstream development of more effective clinical decision systems.

Improve the coverage of DDI evidence in public databases: Recently, more and more research studies utilized DDI evidences from existing public databases for drug interaction study. However, an overlapping analysis between Drugbank and Micromedex showed that there are around 25% of disagreements (Wong, Ko, & Chan, 2008). The lack of comprehensive scientific evidences complicates the process of verifying the discrepancies. Therefore, to explore the mechanism behind drug interaction, it will supply the more necessary scientific evidence to validate DDIs.

Chapter 2. An Integrated Pharmacokinetics Ontology and Corpus

2.1 Background

Pharmacokinetics (PK) is a very important translational research field, which studies drug absorption, disposition, metabolism, excretion, and transportation (ADMET). PK systematically investigates the physiological and biochemical mechanisms of drug exposure in multiple tissue types, cells, animals, and human subjects (Malcolm Rowland & Tozer, 1995). There are two major molecular mechanisms of a drug's PK: metabolism and transportation. The drug metabolism mainly happens in the gut and liver; while drug transportation exists in all tissue types. If the PK can be interpreted as how a body does on the drug, pharmacodynamics (PD) can be defined as how a drug does on the body. A drug's pharmacodynamics effect ranges widely from the molecular signals (such as its targets or downstream biomarkers) to clinical symptoms (such as the efficacy or side effect endpoints) (Malcolm Rowland & Tozer, 1995).

Drug-drug interaction (DDI) is another important pharmacology concept. It is defined as whether one drug's PK or PD response is changed due to the presence of another drug. PD based drug interaction has a wide range of interpretations (i.e. from molecular markers to clinical endpoints). PK based drug interaction mechanism is very well defined: metabolism enzyme based and transporter based DDIs. Pharmacogenetic (PG) variations in a drug's PK and PD pathways can also affect its responses (Malcolm Rowland & Tozer, 1995). In this paper, we focus our discussion on the PK, PK based DDI, and PK related PG.

Although significant efforts have been invested to integrate biochemistry, genetics, and clinical information for drugs, significant gaps exist in the area of PK. For example DrugBank (http://www.drugbank.ca/) doesn't have in vitro PK and its associated DDI data; DiDB (http://www.druginteractioninfo.org/) doesn't have sufficient PG data; and PharmGKB (http://www.pharmgkb.org/) doesn't have sufficient in vivo and in vitro PK and its associated DDI data. As an alternative approach to collect PK from the published literature, text mining has just started to be explored (Malcolm Rowland & Tozer, 1995 ; Segura-Bedmar, Martinez, & de Pablo-Sanchez, 2011d; L. Tari et al., 2010; Wang et al., 2009).

From either database construction or literature mining, the main challenge of PK data integration is the lack of PK ontology. This paper developed a PK ontology first. Then, a PK corpus was constructed. It facilitated DDI text mining from the literature.

2.2 Material and Methods

2.2.1 PK Ontology Construction and Content

PK Ontology is composed of several components: experiments, metabolism, transporter, drug, and subject (Table 2.1). Our primary contribution is the ontology development for the PK experiment, and integration of the PK experiment ontology with other PK-related ontologies.

Categories	Description	Resources
Pharmacokinetics Experiments	Pharmacokinetics studies and parameters. There are two major categories: in vitro experiments and in vivo studies.	Manually accumulated from text books and literatures.
Transporters	Drug transportation enzymes	http://www.tcdb.org
Metabolism Enzymes	Drug metabolism enzymes	http://www.cypalleles.ki.se/
Drugs	Drug names	http://www.drugbank.ca/
Subjects	Subject description for a pharmacokinetics study. It is composed three categories: disease, physiology, and demographics	http://bioportal.bioontology.org/ontologies/42056 http://bioportal.bioontology.org/ontologies/39343 http://bioportal.bioontology.org/ontologies/42067

Table 2.1 PK Ontology Categories

Experiment specifies in vitro and in vivo PK studies and their associated PK parameters. Table 2.2 presents definitions and units of the in vitro PK parameters. The PK parameters of the single drug metabolism experiment include Michaelis-Menten constant (K_m), maximum velocity of the enzyme activity (V_{max}), intrinsic clearance (CL_{int}), metabolic ratio, and fraction of metabolism by an enzyme (fm_{enzyme}) (Segel, 1975). In the transporter experiment, the PK parameters include apparent permeability (Papp), ratio of the basolateral to apical permeability and apical to basolateral permeability (Re), radioactivity, and uptake volume (International Transporter et al., 2010). There are multiple drug interaction mechanisms: competitive inhibition, non-competitive inhibition, uncompetitive inhibition, mechanism based inhibition, and induction (Rostami-Hodjegan A, 2004). IC_{50} is the inhibition concentration that inhibits to 50% enzyme activity; it is substrate dependent; and it doesn't imply the inhibition mechanism. K_i is the inhibition rate constant for competitive inhibition, noncompetitive inhibition, and uncompetitive inhibition. It represents the inhibition concentration that inhibits to 50% enzyme activity, and it is substrate concentration independent. K_{deg} is the degradation rate constant for the enzyme. K_I is the concentration of inhibitor associated with half maximal Inactivation in the mechanism based inhibition; and K_{inact} is the maximum degradation rate constant in the presence of a high concentration of inhibitor in the mechanism based inhibition. E_{max} is the maximum induction rate, and EC_{50} is the concentration of inducer that is associated with the half maximal induction.

Experiment Types	Parameters	Description	Unit	References
Single Drug Metabolism Experiment	K_m	Michaelis-Menten constant.	mg L^{-1}	Segel p28.
	V_{max}	Maximum velocity of the enzyme activity.	mg h^{-1} mg^{-1}	Segel p19
	CL_{int}	Intrinsic metabolic clearance is defined as ratio of maximum metabolism rate, Vmax, and the Michaelis-Menten constant, Km.	ml h^{-1} mg^{-1}	RT p165
	Metabolic ratio	Parent drug/metabolite concentration ratio	NA	
	fm_{enzyme}	Fraction of drug systemically available that is converted to a metabolite through a specific enzyme.	NA	RT xiii
Single Drug Transporter Experiment	Papp	The apparent permeability of compounds across the monolayer cells.	cm/sec	Transport Consortium
	Re	Re is the ratio of basolateral to apical over apical to basolateral.	NA	Transport Consortium
	Radioactivity	Total radioactivity in plasma and bile samples is measured in a liquid scintillation counter	dpm/mg	Transport Consortium
	Uptake Volume	The amount of radioactivity associated with the cells divided by its concentration in the incubation medium.	ul/mg	Transport Consortium
Drug Interaction Experiment	IC_{50}	Inhibitor concentration that inhibits to 50% of enzyme activity.	mg L^{-1}	
	K_i	Inhibition rate constant for competitive inhibition, noncompetitive inhibition, and uncompetitive inhibition.	mg L^{-1}	Segel p103
	K_{deg}	The natural degradation rate constant for the Enzyme.	h^{-1}	Rostami-Hodjegan and Tucker
	K_I	The concentration of inhibitor associated with half maximal Inactivation in the mechanism based inhibition.	mg L^{-1}	Rostami-Hodjegan and Tucker
	K_{inact}	The maximum degradation rate constant in the presence of a high concentration of inhibitor in the mechanism based inhibition.	h^{-1}	Rostami-Hodjegan and Tucker
	E_{max}	Maximum induction rate	Unit free	Rostami-Hodjegan and Tucker
	EC_{50}	The concentration of inducer that is associated with the half maximal induction.	mg L^{-1}	Rostami-Hodjegan and Tucker
Type of Drug Interactions	Competitive/noncompetitive/ uncompetitive/mechanism based inhibition and induction.	Rostami-Hodjegan and Tucker		

Table 2.2 In Vitro PK Parameters

The in vitro experiment conditions are presented in Table 2.3. Metabolism enzyme experiment conditions include buffer, NADPH sources, and protein sources. In particular, protein sources include recombinant enzymes, microsomes, hepatocytes, and etc. Sometimes, genotype information is available for the microsome or hepatocyte samples. Transporter experiment conditions include bi-directional transporter, uptake/efflux, and ATPase. Other factors of in vitro experiments include pre-incubation time, incubation time, quantification methods, sample size, and data analysis methods. All these info can be found in the FDA website (http://www.abclabs.com/Portals/0/FDAGuidance_DraftDrugInteractionStudies2006.pdf).

Experimental Conditions:	drugs	Substrate, metabolite, and inhibitor/inducer			FDA Drug Interaction Guidance, 2006.
Metabolism Enzymes	Buffer	Salt composition			
		EDTA concentration			
		MgCl$_2$ concentration Cytochrome b$_5$ concentration			
	NADPH source	Concentration of exogenous NADPH added isocytrate dehydrogenase + NADP			
	protein	Non-recombinant enzymes	Microsomes (human liver microsomes, human intestine microsomes, S9 fraction, cytosol, whole cell lysate, hepatocytes.		
		Recombinant enzymes	Enzyme name	mg/mL or uM	
			genotype		
Transporters	Bi-Directional	CHO; Caco-2 cells; HEK-293; Hepa-RG; LLC; LLC-PK1 MDR1 cells; MDCK; MDCK-MDR1 cells; Suspension Hepatocyte			
	Transport				
	Uptake/efflux	tumor cells, cDNA transfected cells, oocytes injected with cRNA of transporters			
	ATPase	membrane vesicles from various tissues or cells expressing P-gp, Reconstituted P-gp			
Other factors	Pre-incubation time				
	Incubation time				
	Quantification methods	HPLC/UV, LC/MS/MS, LC/MS, radiographic			
	Sample size				
	Data Analysis	log-linear regression, plotting; and nonlinear regression			

Table 2.3 In Vitro Experiment Conditions

The in vivo PK parameters are presented in Table 2.4. All of the information are summarized from two text books (Gibaldi M, 1982; Malcolm Rowland & Tozer, 1995). There are several main classes of PK parameters. Area under the concentration curve parameters are (AUC_{inf}, AUC_{SS}, AUC_{t}, AUMC); drug clearance parameters are (CL, CL_{b}, CL_{u}, CL_{H}, CL_{R}, CL_{po}, CL_{IV}, CL_{int}, CL_{12}); drug concentration parameters are (C_{max}, C_{SS}); extraction ratio and bioavailability parameters are (E, E_{H}, F, F_{G}, F_{H}, F_{R}, f_{e}, f_{m}); rate constants include elimination rate constant k, absorption rate constant ka, urinary excretion rate constant ke, Michaelis-Menten constant Km, distribution rate constants (k_{12}, k_{21}), and two rate constants in the two-compartment model (λ_{1}, λ_{2}); blood flow rate (Q, Q_{H}); time parameters (t_{max}, $t_{1/2}$); volume distribution parameters (V, V_{b}, V_{1}, V_{2}, V_{ss}); maximum rate of metabolism, Vmax; and ratios of PK parameters that present the extend of the drug interaction, (AUCR, CL ratio, Cmax ratio, C_{ss} ratio, $t_{1/2}$ ratio).

Category	Name	Description	Unit	reference
PK parameter	AUC$_{inf}$	Area under the drug concentration time curve.	mg h L^{-1}	RT p37
	AUC$_{SS}$	Area under the drug concentration time curve within a dosing curve at steady state.	mg h L^{-1}	RT pxi
	AUC$_{t}$	Area under the drug concentration time curve from time 0 to t.	mg h L^{-1}	RT p37
	AUMC	Area under the first moment of concentration versus time curve.	mg^{2} h L^{-2}	RT p486
	AUCR	AUC ratio (drug interaction parameter).	Unit free	
	CL	Total clearance: defined as the proportionality factor relating rate of elimination to the plasma drug concentration.	ml h^{-1}	RT p23
	CL$_{b}$	Blood clearance: defined as the proportionality factor relating rate of elimination to the blood drug concentration.	ml h^{-1}	RT p160
	CL$_{u}$	Unbound clearance: defined as the proportionality factor relating rate of elimination to the unbounded plasma drug concentration.	ml h^{-1}	RT p163
	CL$_{H}$	Hepatic portion of the total clearance.	ml h^{-1}	RT p161
	CL$_{R}$	Renal portion of the total clearance.	ml h^{-1}	RT p161
	CL$_{po}$	Total clearance of drug following an oral dose.	ml h^{-1}	
	CL$_{IV}$	Total clearance of drug following an IV dose.	ml h^{-1}	
	CL$_{int}$	Intrinsic metabolic clearance is defined as ratio of maximum metabolism rate, Vmax, and the Michaelis-Menten constant, Km.	ml h^{-1}	RT p165
	CL$_{12}$	Inter-compartment distribution between the central compartment and the peripheral compartment.	ml h^{-1}	
	CL ratio	Ratio of the clearance (drug interaction parameter).	Unit free	
	C$_{max}$	Highest drug concentration observed in plasma following administration of an extravascular dose.	mg L^{-1}	RT pxii
	C$_{max}$ ratio	The ratio of C$_{max}$ (drug interaction parameter).	Unit free	
	C$_{ss}$	Concentration of drug in plasma at steady state during a constant rate intravenous infusion.	mg L^{-1}	RT pxii
	C$_{ss}$ ratio	The ratio of C$_{ss}$ (drug interaction parameter).	Unit free	
	E	Extraction ratio is defined as the ratio between blood clearance, CL$_{b}$, and the blood flow.	Unit free	RT p159
	E$_{H}$	Hepatic extraction ratio.	Unit free	RT p161
	F	Bioavailability is defined as the proportion of the drug reaches the systemic blood.	Unit free	RT p42
	F$_{G}$	Gut-wall bioavailability.	Unit free	
	F$_{H}$	Hepatic bioavailability.	Unit free	RT p167
	F$_{R}$	Renal bioavailability.	Unit free	RT p170
	fe	Fraction of drug systemically available that is excreted unchanged in urine.	Unit free	RT pxiii
	fm	Fraction of drug systemically available that is converted to a metabolite.	Unit free	RT pxiii
	fu	Ratio of unbound and total drug concentrations in plasma.	Unit free	RT pxiii
	k	Elimination rate constant.	h^{-1}	RT pxiii
	K$_{12}$, k$_{21}$	Distribution rate constants between central compartment and peripheral compartment.	h^{-1}	
	ka	Absorption rate constant.	h^{-1}	RT pxiii
	ke	Urinary excretion rate constant.	h^{-1}	RT pxiii
	km	Rate constant for the elimination of a metabolite.	h^{-1}	RT pxiii
	Km	Michaelis-Menten constant.	mg L^{-1}	RT pxiii
	MRT	Mean time a molecular resides in body.	h	RT pxiv
	Q	Blood flow.	L h^{-1}	RT pxiv
	Q$_{H}$	Hepatic blood flow.	L h^{-1}	RT pxiv
	t$_{max}$	Time at which the highest drug concentration occurs following administration of an extravascular dose.	h	RT pxiv
	t$_{1/2}$	Half-life of the drug disposition.	h	RT pxiv
	t$_{1/2}$ ratio	Half-life ratio (drug interaction parameter).	Unit free	
	t$_{1/2,\alpha}$	Half-life of the fast phase drug disposition.	h	
	t$_{1/2,\beta}$	Half-life of the slow phase drug disposition.	h	
	V	Volume of distribution based on drug concentration in plasma.	L	RT pxiv
	V$_{b}$	Volume of distribution based on drug concentration in blood.	L	RT pxiv
	V$_{1}$	Volume of distribution of the central compartment.	L	RT pxiv
	V$_{2}$	Volume of distribution of the peripheral compartment.	L	
	V$_{ss}$	Volume of distribution under the steady state concentration.	L	RT pxiv
	Vmax	Maximum rate of metabolism by an enzymatically mediated reaction.	mg h^{-1}	RT pxiv
	λ_{1}, λ_{2}	Disposition rate constants in a two-compartment model.	h^{-1}	GP p84
Pharmacokinetics Models	Non-Compartment	Use drug concentration measurements directly to estimate PK parameters, such as AUC, CL, C$_{max}$, T$_{max}$, t$_{1/2}$, F, and V.		GP p409
	One Compartment	The whole body is a homogeneous compartment, and the distribution of the drug from the blood to tissue is very fast. It assumes either a first order or a zero order absorption rate and a first order eliminate rate.		RT p34 GP p1
	Two Compartment	It assumes the whole body can be divided into two compartments: central compartment and peripheral compartment. It assumes either a first order or a zero order absorption rate and a first order eliminate and distribution rates.		GP p84
Study Designs	Hypothesis	Bioequivalence, drug interaction, pharmacogenetics, and disease conditions.		
	Design	Single arm or multiple arms; cross-over or fixed order design; with or without randomization; with or without stratification; prescreening or no-prescreening; prospective or retrospective studies; and case reports or cohort studies.		
	Sample size	The number of subjects, and the number of plasma or urine samples per subject.		
	Time points	Sampling time points and dosing time points.		
	Sample types	Blood, plasma, and urine.		
	Dose	Subject specific doses.		
Quantifi. methods	HPLC/UV, LC/MS/MS, LC/MS, radiographic			

Table 2.4 In Vivo PK Studies

It is also shown in Table 2.4 that two types of pharmacokinetics models are usually presented in the literature: non-compartment model and one or two-compartment models. There are multiple items need to be considered in an in vivo PK study. The hypotheses include the effect of bioequivalence, drug interaction, pharmacogenetics, and disease conditions on a drug's PK. The design strategies are very diverse: single arm or multiple arms, cross-over or fixed order design, with or without randomization, with or without stratification, pre-screening or no-pre-screening based on genetic information, prospective or retrospective studies, and case reports or cohort studies. The sample size includes the number of subjects, and the number of plasma or urine samples per subject. The time points include sampling time points and dosing time points. The sample type includes blood, plasma, and urine. The drug quantification methods include HPLC/UV, LC/MS/MS, LC/MS, and radiographic.

CYP450 family enzymes predominantly exist in the gut wall and liver. Transporters are tissue specific. Table 2.5 presents the tissue specific transports and their functions. Probe drug is another important concept in the pharmacology research. An enzyme's probe substrate means that this substrate is primarily metabolized or transported by this enzyme. In order to experimentally prove whether a new drug inhibits or induces an enzyme, its probe substrate is always utilized to demonstrate this enzyme's activity before and after inhibition or induction. An enzyme's probe inhibitor or inducer means that it inhibits or induces this enzyme primarily. Similarly, an enzyme's probe inhibitor needs to be utilized if we investigate whether a drug is metabolized by this enzyme. Table 2.6 presents all the probe inhibitors, inducers, and substrates of CYP enzymes. Table 2.7

presents all the probe inhibitors, inducers, and substrates of the transporters. All these information were collected from industry standard (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm064982.htm), reviewed in the top pharmacology journal (Huang et al., 2007).

Gene	Aliases	Tissue type	Function
ABCB1	P-gp, MDR1	Intestinal enterocyte, kidney proximal tubule, hepatocyte (canalicular), brain endothelia	Efflux
ABCG2	BCRP	Intestinal enterocyte, hepatocyte (canalicular), kidney proximal tubule, brain endothelia, placenta, stem cells, mammary gland (lactating)	Efflux
SLCO1B1	OATP1B1, OATP-C, OATP2, LST-1	Hepatocyte (sinusoidal)	Uptake
SLCO1B3	OATP1B3, OATP-8	Hepatocyte (sinusoidal)	Uptake
SLC22A2	OCT2	Kidney proximal tubule	Uptake
SLC22A6	OAT1	Kidney proximal tubule, placenta	Uptake
SLC22A8	OAT3	Kidney proximal tubule, choroid plexus, brain endothelia	Uptake

Table 2.5 Tissue Specific Transporters

CYP Enzymes	Inhibitors	Inducers	Substrates
CYP1A2	Ciprofloxacin, enoxacin, fluvoxamine, Methoxsalen, mexiletine, oral contraceptives, phenylpropanolamine, thiabendazole, vemurafenib, zileuton, acyclovir, allopurinol, caffeine, cimetidine, daidzein, disulfiram, Echinacea, famotidine, norfloxacin, propafenone, propranolol, terbinafine, ticlopidine, verapamil	Montelukast, phenytoin, smokers versus non-smokers, moricizine, omeprazole, phenobarbital	Alosetron, caffeine, duloxetine, melatonin, ramelteon, tacrine, tizanidine, theophylline, tizanidine
CYP2B6	Clopidogrel, ticlopidine prasugrel	Efavirenz, rifampin, nevirapine	Bupropion, efavirenz
CYP2C8	Gemfibrozil, fluvoxamine, ketoconazole, trimethoprim	Rifampin	Repaglinide, Paclitaxel
CYP2C9	Amiodarone, fluconazole, miconazole, oxandrolone, capecitabine, cotrimoxazole, etravirine, fluvastatin, fluvoxamine, metronidazole, sulfinpyrazone, tigecycline, voriconazole, zafirlukast	Carbamazepine, rifampin, aprepitant, bosentan, phenobarbital, St. John's wort	Celecoxib, Warfarin, phenytoin
CYP2C19	Fluconazole, fluvoxamine, ticlopidine, esomeprazole, fluoxetine, moclobemide, omeprazole, voriconazole, allicin (garlic derivative), armodafinil, carbamazepine, cimetidine, etravirine, human growth hormone (rhGH), felbamate, ketoconazole, oral contraceptives	Rifampin, artemisinin	Clobazam, lansoprazole, omeprazole, Smephenytoin, S-mephenytoin
CYP3A	Boceprevir, clarithromycin, conivaptan, grapefruit juice, indinavir, itraconazole,	Avasimibe, carbamazepine, phenytoin, rifampin, St. John's wort, bosentan, efavirenz, etravirine, modafinil, nafcillin, amprenavir, aprepitant, armodafinil, clobazamechinacea, pioglitazone, prednisone, rufinamide, vemurafenib	Alfentanil, aprepitant, budesonide, buspirone, conivaptan, darifenacin, darunavir, dasatinib, dronedarone, eletriptan, eplerenone, everolimus, felodipine, indinavir, fluticasone, lopinavir, lovastatin, lurasidone, maraviroc, midazolam, nisoldipine, quetiapine, saquinavir, sildenafil, simvastatin, sirolimus, tolvaptan, tipranavir, triazolam, ticagrelor, vardenafil, Alfentanil, astemizole, cisapride, cyclosporine, dihydroergotamine, ergotamine, fentanyl, pimozide, quinidine, sirolimus, tacrolimus, terfenadine
	ketoconazole, lopinavir/ritonavir, mibefradil, nefazodone, nelfinavir, posaconazole, ritonavir, saquinavir, telaprevir, telithromycin, voriconazole, amprenavir, aprepitant, atazanavir, ciprofloxacin, crizotinib, darunavir/ritonavir, diltiazem, erythromycin, fluconazole, fosamprenavir, grapefruit juice, imatinib, verapamil, alprazolam, amiodarone, amlodipine, atorvastatin, bicalutamide, cilostazol, cimetidine, cyclosporine, fluoxetine, fluvoxamine, ginkgo, goldenseal, isoniazid, lapatinib, nilotinib, oral contraceptives, pazopanib, ranitidine, ranolazine, tipranavir/ritonavir, ticagrelor, zileuton		
CYP2D6	Bupropion, fluoxetine, paroxetine, quinidine, cinacalcet, duloxetine, terbinafine,	NA	Atomoxetine, desipramine, dextromethorphan, metoprolol, nebivolol, perphenazine, tolterodine, venlafaxine, Thioridazine, pimozide
	amiodarone, celecoxib, clobazam, cimetidine, desvenlafaxine, diltiazem, diphenhydramine, echinacea, escitalopram, febuxostat, gefitinib, hydralazine, hydroxychloroquine, imatinib, methadone, oral contraceptives, pazopanib, propafenone, ranitidine, ritonavir, sertraline, telithromycin, verapamil, vemurafenib		

Table 2.6 In Vivo Probe Inhibitors/Inducers/Substrates of CYP Enzymes

Transporter	Inhibitor	Inducer	Substrate
P-gp	Amiodarone, azithromycin, captopril, carvedilol, clarithromycin, conivaptan, cyclosporine, diltiazem, dronedarone, erythromycin, felodipine, itraconazole, ketoconazole, lopinavir and ritonavir, quercetin, quinidine, ranolazine, ticagrelor, verapamil	Avasimibe, carbamazepine, phenytoin, rifampin, St John's wort, tipranavir/ritonavir	Aliskiren, ambrisentan, colchicine, dabigatran etexilate, digoxin, everolimus, fexofenadine, imatinib, lapatinib, maraviroc, nilotinib, posaconazole, ranolazine, saxagliptin, sirolimus, sitagliptin, talinolol, tolvaptan, topotecan
BCRP	Cyclosporine, elacridar (GF120918), eltrombopag, gefitinib	NA	Methotrexate, mitoxantrone, imatinib, irrinotecan, lapatinib, rosuvastatin, sulfasalazine, topotecan
OATP1B1	Atazanavir, cyclosporine, eltrombopag, gemfibrozil, lopinavir, rifampin, ritonavir, saquinavir, tipranavir	NA	Atrasentan, atorvastatin, bosentan, ezetimibe, fluvastatin, glyburide, SN-38 (active metabolite of irinotecan), rosuvastatin, simvastatin acid, pitavastatin, pravastatin, repaglinide, rifampin, valsartan, olmesartan
OATP1B3	Atazanavir, cyclosporine, lopinavir, rifampin, ritonavir, saquinavir	NA	Atorvastatin, rosuvastatin, pitavastatin, telmisartan, valsartan, olmesartan
OCT2	Cimetidine, quinidine	NA	Amantadine, amiloride, cimetidine, dopamine, famotidine, memantine, metformin, pindolol, procainamide, ranitidine, varenicline, oxaliplatin
OAT1	Probenecid	NA	Adefovir, captopril, furosemide, lamivudine, methotrexate, oseltamivir, tenofovir, zalcitabine, zidovudine
OAT3	Probenecid, cimetidine, diclofenac	NA	Acyclovir, bumetanide, ciprofloxacin, famotidine, furosemide, methotrexate, zidovudine, oseltamivir acid, (the active metabolite of oseltamivir), penicillin G, pravastatin, rosuvastatin, sitagliptin

Table 2.7 In Vivo Probe Inhibitors/Inducers/Substrates of Selected Transporters

Metabolism The cytochrome P450 superfamily (officially abbreviated as CYP) is a large and diverse group of enzymes that catalyze the oxidation of organic substances. The substrates of CYP enzymes include metabolic intermediates such as lipids and steroidal hormones, as well as xenobiotic substances such as drugs and other toxic chemicals. CYPs are the major enzymes involved in drug metabolism and bioactivation, accounting for about 75% of the total number of different metabolic reactions (FP, 2008). CYP enzyme names and genetic variants were mapped from the Human Cytochrome P450 (CYP) Allele Nomenclature Database (http://www.cypalleles.ki.se/). This site contains the CYP450 genetic mutation effect on the protein sequence and enzyme activity with associated references.

Transport Proteins are proteins which serves the function of moving other materials within an organism. Transport proteins are vital to the growth and life of all living things. Transport proteins involved in the movement of ions, small molecules, or macromolecules, such as another protein, across a biological membrane. They are integral membrane proteins; that is they exist within and span the membrane across which they transport substances. Their names and genetic variants were mapped from the Transporter Classification Database (http://www.tcdb.org). In addition, we also added the probe substrates and probe inhibitors to each one of the metabolism and transportation enzymes (see prescribed description).

Drug names was created using the drug names from DrugBank 3.0 (Knox et al., 2011). DrugBank consists of 6,829 drugs which can be grouped into different categories of FDA-approved, FDA approved biotech, nutraceuticals, and experimental drugs. The drug names are mapped to generic names, brand names, and synonyms.

Subject included the existing ontologies for human disease ontology (DOID), suggested Ontology for Pharmacogenomics (SOPHARM),, and mammalian phenotype (MP) from http://bioportal.bioontology.org (see Table 2.1). The PK ontology was implemented with Protégé (Rubin, Noy, & Musen, 2007) and uploaded to the BioPortal ontology platform.

2.2.2 PK Corpus

A PK abstract corpus was constructed to cover four primary classes of PK studies: clinical PK studies (n = 56); clinical pharmacogenetic studies (n = 57); in vivo DDI studies (n = 218); and in vitro drug interaction studies (n = 210). The PK corpus construction process is a manual process. The abstracts of clinical PK studies were selected from our previous work, in which the most popular CYP3A substrate, midazolam was investigated (Wang et al., 2009). The clinical pharmacogenetic abstracts were selected based on the most polymorphic CYP enzyme, CYP2D6. We think these two selection strategies represent very well all the in vivo PK and PG studies. In searching for the drug interaction studies, the abstracts were randomly selected from a PubMed query, which used probe substrates/inhibitors/inducers for metabolism enzymes reported in Table 2.6.

Once the abstracts have been identified in four classes, their annotation is a manual process (Figure 2.1). The annotation was firstly carried out by three master level annotators (Shreyas Karnik, Abhinita Subhadarshini, and Xu Han), and one Ph.D. annotator (Lang Li). They have different training backgrounds: computational science, biological science, and pharmacology. Any differentially annotated terms were further checked by Sara K. Quinney and David A. Flockhart, one Pharm D. and one M.D. scientists with extensive pharmacology training background. Among the disagreed annotations between these two annotators, a group review was conducted (Drs Quinney, Flockhart, and Li) to reach the final agreed annotations. In addition a random subset of 20% of the abstracts that had consistent annotations among four annotators (3 masters and one Ph.D.), were double checked by two Ph.D. level scientists.

A structured annotation scheme was implemented to annotate three layers of pharmacokinetics information: key terms, DDI sentences, and DDI pairs (Figure 2.2). DDI sentence annotation scheme depends on the key terms; and DDI annotations depend on the key terms and DDI sentences. Their annotation schemes are described as following.

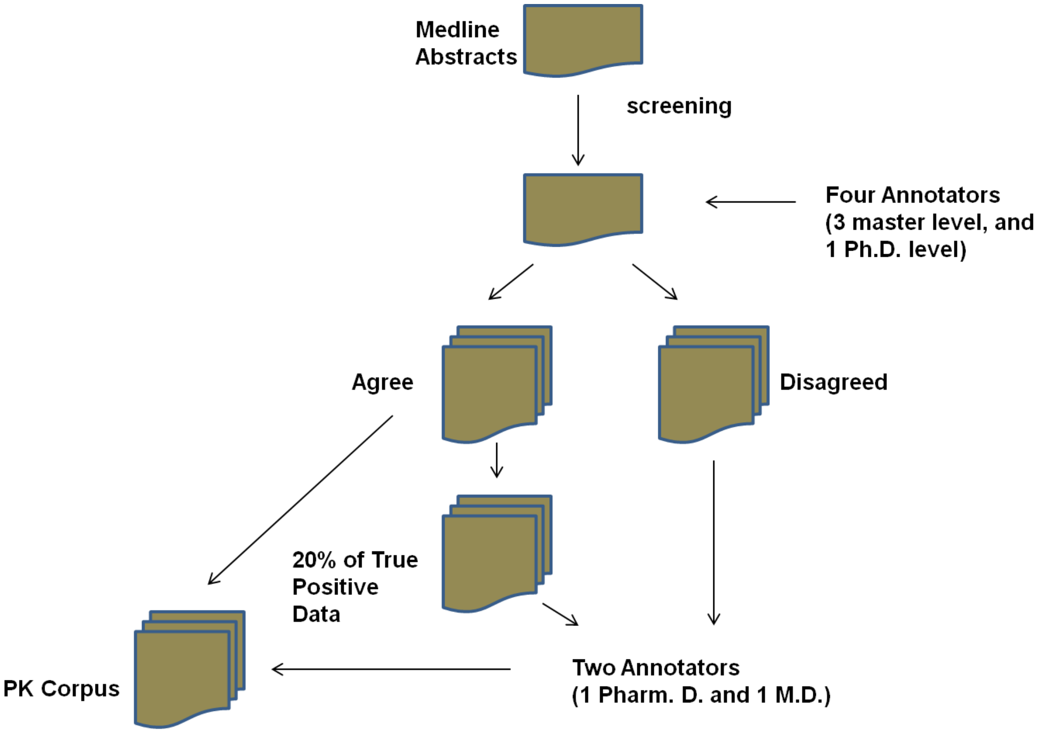


Figure 2.1 PK Corpus Annotation Flow Chart

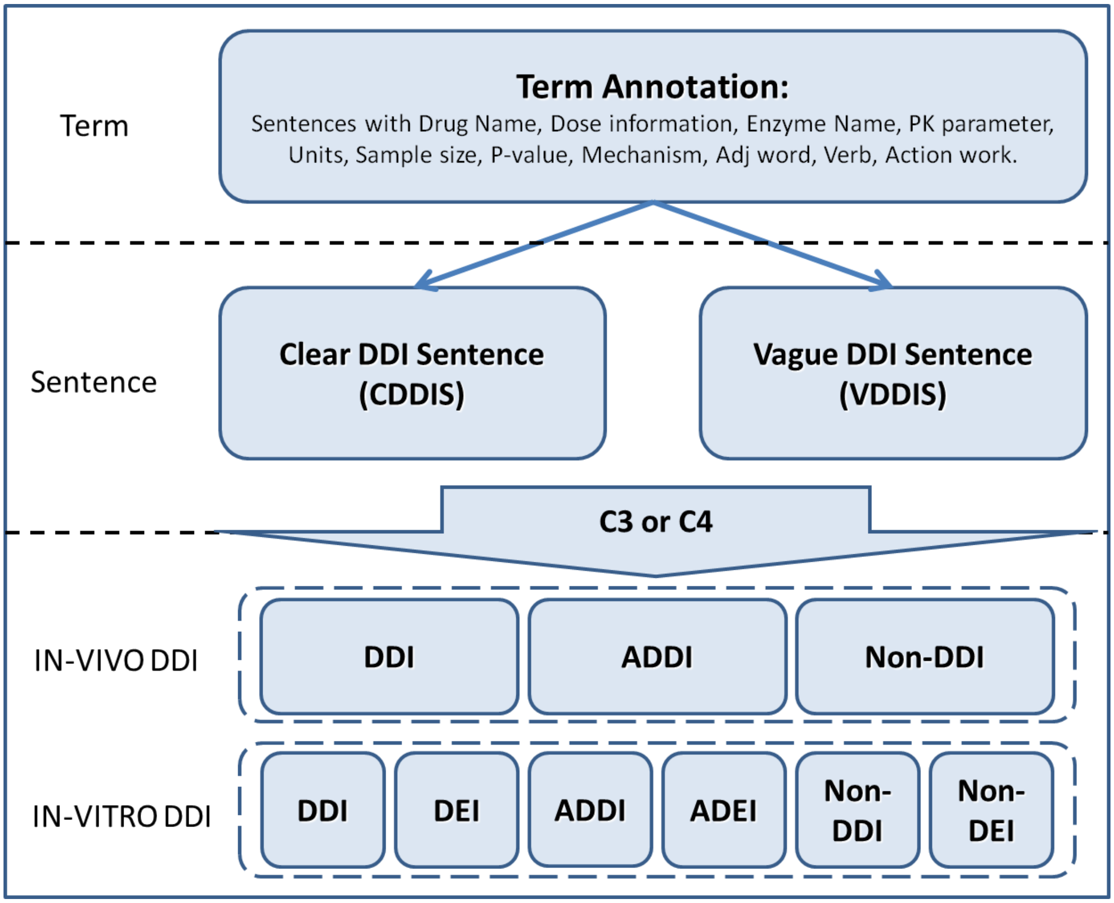

Term
Term Annotation:
Sentences with Drug Name, Dose information, Enzyme Name, PK parameter, Units, Sample size, P-value, Mechanism, Adj word, Verb, Action work.
Sentence
Clear DDI Sentence (CDDIS)
Vague DDI Sentence (VDDIS)
C3 or C4
IN-VIVO DDI
DDI
ADDI
Non-DDI
IN-VITRO DDI
DDI
DEI
ADDI
ADEI
Non-DDI
Non-DEI

Figure 2.2 A Three Level Hierarchical PK and DDI Annotation Scheme.

Key terms include drug names, enzyme names, PK parameters, numbers, mechanisms, and change. The boundaries of these terms among different annotators were judged by the following standard.

- Drug names were defined mainly on DrugBank 3.0 (Knox et al., 2011). In addition, drug metabolites were also tagged, because they are important in in vitro studies. The metabolites were judged by either prefix or suffix: oxi, hydroxyl, methyl, acetyl, N-dealkyl, N-demethyl, nor, dihydroxy, O-dealkyl, and sulfo. These prefixes and suffixes are due to the reactions due to phase I metabolism (oxidation, reduction, hydrolysis), and phase II metabolism (methylation, sulphation, acetylation, glucuronidation) (LL, BA, & BC, 2011).

- Enzyme names covered all the CYP450 enzymes. Their names are defined in the human cytochrome P450 allele nomenclature database, http://www. cypalleles.ki.se/. The variations of the enzyme or gene names were considered. Its regular expression is (?:cyp|CYP|P450|CYP450)?[0–9][a-zA-Z][0–9](?:*[0–9])?$.

- PK parameters were annotated based on the defined in vitro and in vivo PK parameter ontology in Table 2.2 and Table 2.4. In addition, some PK parameters have different names, CL = clearance, t1/2 = half-life, AUC = area under the concentration curve, and AUCR = area under the concentration curve ratio.

- Numbers such as dose, sample size, the values of PK parameters, and p-values were all annotated. If presented, their units were also covered in the annotations.

- Mechanisms denote the drug metabolism and interaction mechanisms. They were annotated by the following regular expression patterns: inhibit(e(s|d)?|ing|ion(s)?|or)$, catalyz(e(s|d)?|ing)$, correlat(e(s|d)?|ing|ion(s)?)$, metaboli(z(e(s|d)?|ing)|sm)$, induc(e(s|d)?|ing|tion(s)?|or)$, form((s|ed)?|ing|tion(s)?|or)$, stimulat(e(s|d)?|ing|ion(s)?)$, activ(e(s)?|(at)(e(s|d)?|ing|ion(s)?))$, and suppress(e(s|d)?|ing|ion(s)?)$.

- Change describes the change of PK parameters. The following words were annotated in the corpus to denote the change: strong(ly)?, moderate(ly)?, high(est)?(er)?, slight(ly)?, strong(ly)?, moderate(ly)?, slight(ly)?, significant(ly)?, obvious(ly)?, marked(ly)?, great(ly)?, pronounced(ly)?, modest(ly)?, probably, may, might, minor, little, negligible, doesn’t interact, affect((s|ed)?|ing|ion(s)?)?$, reduc(e(s|d)?|ing|tion(s)?)$, and increas(e(s|d)?|ing)$.

The middle level annotation focused on the drug interaction sentences. Because two interaction drugs were not necessary all presented in the sentence, sentences were categorized into two classes:

- Clear DDI Sentence (CDDIS): two drug names (or drug-enzyme pair in the in vitro study) are in the sentence with a clear interaction statement, i.e. either interaction, or non-interaction, or ambiguous statement (i.e. such as possible or might and etc.).

- Vague DDI Sentence (VDDIS): One drug or enzyme name is missed in the DDI sentence, but it can be inferred from the context. Clear interaction statement also is required.

Once DDI sentences were labeled, the DDI pairs in the sentences were further annotated. Because the fundamental difference between in vivo DDI studies and in vitro DDI studies, their DDI relationships were defined differently. In in vivo studies, three types of DDI relationships were defined (Table 2.8): DDI, ambiguous DDI (ADDI), and non-DDI (NDDI). Four conditions are specified to determine these DDI relationships. Condition 1 (C1) requires that at least one drug or enzyme name has to be contained in the sentence; condition 2 (C2) requires the other interaction drug or enzyme name can be found from the context if it is not from the same sentence; condition 3 (C3) specifies numeric rules to defined the DDI relationships based on the PK parameter changes; and condition 4 (C4) specifies the language expression patterns for DDI relationships. Using the rules summarized in Table 2.8, DDI, ADDI, and NDDI can be defined by $C1 \wedge C2 \wedge (C3 \vee C4)$. The priority rank of in vivo PK parameters is AUC > CL > $t_{1/2}$ > C_{max}. In in vitro studies, six types of DDI relationships were defined (Table 2.8). DDI, ADDI, NDDI were similar to in vivo DDIs, but three more drug-enzyme relationships were further defined: DEI, ambiguous DEI (ADEI), and non-DDI (NDEI). C1, C2, and C4 remained the same for in vitro DDIs. The main difference is in C3, in which either Ki or IC50 (inhibition) or EC50 (induction) were used to defined DDI relationship quantitatively. The priority rank of in vitro PK parameters is Ki > IC50. Table 2.9 presented eight examples of how DDIs or DEIs were determined in the sentences.

DDI relationship	C1	C2	C3**	C4**
IN VIVO STUDY				
DDI	Yes	Yes	The PK parameter with the highest priority* must satisfy p-value <0.05 and FC > 1.50 or FC < 0.67	Significant, obviously, markedly, greatly, pronouncedly and etc.
Ambiguous DDI (ADDI)			The PK parameter with the highest priority* in the conditions of p-value <0.05 but 0.67 < FC < 1.50; or FC >1.50 or FC <0.67, but p-value > 0.05.	Modestly, moderately, probably, may, might, and etc.
Non-DDI (NDDI)			The PK parameter with the highest priority*are in the condition of p-value > 0.05 and 0.67 < FC < 1.50	Minor significance, slightly, little or negligible effect, doesn't interact etc.
IN VITRO STUDY				
DDI	Yes	Yes	(0< Ki < 10 or 0< EC50 < 10 microM, and p-value <0.05)	Significant, obviously, markedly, greatly, pronouncedly and etc.
DEI				
Ambiguous DDI (ADDI)			(10 < Ki < 100 or 10 < EC50 < 100 microM, and p-value <0.05 or vice versa)	Modestly, moderately, probably, may, might, and etc.
Ambiguous DEI (ADEI)				
Non-DDI (NDDI)			(Ki > 100 microM or EC50 > 100 microM, and p-value >0.05)	Minor significance, slightly, little or negligible effect, doesn't interact etc.
Non-DEI (NDEI)				

Table 2.8 DDI Definitions in Corpus

Note:

C1: At least one drug or enzyme name has to be contained in the sentence.

C2: Need to label the drug name if it is not from the same sentence.

C3: PK-parameter and value dependent.

C4: Significance statement.

*Priority issue: When C3 and C4 occur and conflict, C3 dominates the sentence.**For the priority of PK parameters: AUC > CL > $t_{1/2}$ > $C_{max;}$; the priority of in vitro PK parameters: Ki>IC50.

PMID	DDI sentence	Relationship and commend
20012601	The pharmacokinetic parameters of verapamil were significantly altered by the co-administration of lovastatin compared to the control.	Because of the words, “significantly”, (Verapamil, lovastatin) is a DDI.
20209646	The clearance of mitoxantrone and etoposide was decreased by 64% and 60%, respectively, when combined with valspodar.	Because of the fold changes were less than 0.67, (mitoxantrone, valspodar.) and (etoposide, valspodar) are DDIs.
20012601	The (AUC (0-infinity)) of norverapamil and the terminal half-life of verapamil did not significantly changed with lovastatin coadministration.	Because of the words, “not significantly changed”, (verapamil, ovastatin) is a NDDI.
17304149	Compared with placebo, itraconazole treatment significantly increase the peak plasma concentration (Cmax) of paroxetine by 1.3 fold (6.7 2.5 versus 9.0 3.3 ng/mL, P≤0.05) and the area under the plasma concentration-time curve from zero to 48 hours [AUC(0–48)] of paroxetine by 1.5 fold (137 73 versus 199 91 ng*h/mL, P≤0.01).	AUC has a higher rank than Cmax, and it had a 1.5 fold-change and less than 0.05 p-value, thus, (itraconazole, paroxetine) is a DDI.
13129991	The mean (SD) urinary ratio of dextromethorphan to its metabolite was 0.006 (0.010) at baseline and 0.014 (0.025) after St John’s wort administration (P=.26)	The change in PK parameter is more than 1.5 fold but P-value is >0.05. Thus, (dextromethorphan, St John’s wort) is an ADDI.
19904008	The obtained results show that perazine at its therapeutic concentrations is a potent inhibitor of human CYP1A2.	Because of words, “potent inhibitor”, (perazine, CYP1A2) is a DEI.
19230594	After human hepatocytes were exposed to 10 microM YM758, microsomal activity and mRNA level for CYP1A2 were not induced while those for CYP3A4 were slightly induced.	Because of words, “not induced” and “slightly induced”, (YM758, CYP1A2) and (YM758, CYP1A2) are NDEIs.
19960413	From these results, DPT was characterized to be a competitive inhibitor of CYP2C9 and CYP3A4, with K(i) values of 3.5 and 10.8 microM in HLM and 24.9 and 3.5 microM in baculovirus-insect cell-expressed human CYPs, respectively.	Because K was larger than 10microM, (DPT, CYP2C9) and (DPT, CYP3A4) are ADEIs.

Table 2.9 Examples of DDI Definitions

Key Terms	Annotation Categories	Frequencies	Krippendorff's alpha
	Drug	8633	0.953
	CYP	3801	
	PK Parameter	1508	
	Number	3042	
	Mechanism	2732	
	Change	1828	
	Total words	97291	
DDI sentences	CDDI sentences	1191	0.921
	VDDI sentences	120	
	Total sentences	4724	
DDI Pairs	DDI	1239	0.905
	ADDI	300	
	NDDI	294	
	DEI	565	
	ADEI	95	
	NDEI	181	
	Total Drug Pairs	12399	

Table 2.10 Annotation Performance Evaluation

Krippendorff's alpha (Klaus Krippendorff, 2004) was calculated to evaluate the reliability of annotations from four annotators. The frequencies of key terms, DDI sentences, and DDI pairs are presented in Table 2.10. Their Krippendorff's alphas are 0.953, 0.921, and 0.905, respectively. Please note that the total DDI pairs refer to the total pairs of drugs within a DDI sentence from all DDI sentences.

The PK corpus was constructed by the following process. Raw abstracts were downloaded from PubMed in XML format. Then XML files were converted into GENIA corpus format following the gpml.dtd from the GENIA corpus (J. D. Kim et al., 2003). The sentence detection in this step is accomplished by using the Perl module Lingua::EN::Sentence, which was downloaded from The Comprehensive Perl Archive Network (CPAN, www.cpan.org). GENIA corpus files were then tagged with the prescribed three levels of PK and DDI annotations. Finally, a cascading style sheet (CSS) was implemented to differentiate colours for the entities in the corpus. This feature allows the users to visualize annotated entities. We would like to acknowledge that a DDI Corpus was recently published as part of a text mining competition DDIExtraction 2011 (http://labda.inf.uc3m.es/ DDIExtraction2011/dataset.html). Their DDIs were clinical outcome oriented, not PK oriented. They were extracted from DrugBank, not from PubMed abstracts. Our PK corpus complements to their corpus very well.

2.3 Utility

2.3.1 Example 1 An Annotated Tamoxifen Pharmacogenetics Study

This example shows how to annotate a pharmacogenetics studies with the PK ontology. We used a published tamoxifen PG study (Borges et al., 2010). The key information from this tamoxifen PG trial was extracted as a summary list. Then the pre-processed information was mapped to the PK ontology (column 2 in Table S1). This PG study investigates the genetics effects (CYP3A4, CPY3A5, CYP2D6, CYP2C9, CYP2B6) on the tamoxifen pharmacokinetics outcome (tamoxifen metabolites) among breast cancer patients. It was a single arm longitudinal study (n = 298), patients took SOLTAMOX^{TM} 20mg/day, and the drug steady state concentration was sampled (1, 4, 8, 12) months after the tamoxifen treatment. The study population was a mixed Caucasian and African American. In additional file 1: Table S1, the trial summary is well organized by the PK ontology.

2.3.2 Example 2 Midazolam/Ketoconazole Drug Interaction Study

This was a cross-over three-phase drug interaction study (Chien et al., 2006) (n = 24) between midazolam (MDZ) and ketoconazole (KTZ). Phase I was MDZ alone (IV 0.05 mg/kg and PO 4mg); phase II was MDZ plus KTZ (200mg); and phase III was MDZ plus KTZ (400mg). Genetic variable include CYP3A4 and CYP3A5. The PK outcome is the MDZ AUC ratio before and after KTZ inhibition. Its PK ontology based annotation is shown in additional file 1: Table S1 column three.

2.3.3 Example 3 In Vitro Pharmacokinetics Study

This was an in vitro study (Williams et al., 2002), which investigated the drug metabolism activities for 3 enzymes, such as CYP3A4, CYP3A5, and CYP3A7 in a recombinant system. Using 10 CYP3A substrates, they compared the relative contribution of 3 enzymes among 10 drug's metabolism. Its PK ontology based annotation is shown in additional file 1: Table S2.

2.3.4 Example 4 A Drug Interaction Text Mining Example

We implemented the approach described by (Airola et al., 2008) for the DDI extraction. Prior to performing DDI extraction, the testing and validation DDI abstracts in our corpus was pre-processed and converted into the unified XML format (Airola et al., 2008). The following steps were conducted:

- Drugs were tagged in each of the sentences using dictionary based on DrugBank. This step revised our prescribed drug name annotations in the corpus. One purpose is to reduce the redundant synonymous drug names. The other purpose is only keep the parent drugs and remove the drug metabolites from the tagged drug names from our initial corpus, because parent drugs and their metabolites rarely interacts. In addition, enzymes (i.e. CYPs) were also tagged as drugs, since enzyme-drug interactions have been extensively studied and published. The regular expression of enzyme names in our corpus was used to remove the redundant synonymous gene names.

- Each of the sentences was subjected to tokenization, Part-of-Speech (PoS) tags and dependency tree generation using the Stanford parser (De Marneffe, 2006).

- C_2^n drug pairs form the tagged drugs in a sentence were generated automatically, and they were assigned with default labels as no-drug interaction. Please note that if a sentence had only one drug name, this sentence didn't have a DDI. This setup limited us considering only CDDI sentence in our corpus.

- The drug interaction labels were then manually flipped based on their true drug interaction annotations from the corpus. Please note that our corpus had annotated DDIs, ADDIs, NDDIs, DEIs, ADEIs, and NDEIs. Here only DDIs and DEIs were labeled as true DDIs. The other ADDIs, NDDIs, DEIs, and ADEIs were all categorized into the no-drug interactions.

Then sentences were represented with dependency graphs using interacting components (drugs) (Figure 2.3). The graph representation of the sentence was composed of two items: i) One dependency graph structure of the sentence; ii) a sequence of PoS tags (which was transformed to a linear order "graph" by connecting the tags with a constant edge weight). We used the Stanford parser (De Marneffe, 2006) to generate the dependency graphs. Airola et al. proposed to combine these two graphs to one weighted, directed graph. This graph was fed into a support vector machine (SVM) for DDI/non-DDI classification. More details about the all paths graph kernel algorithm can be found in (Airola et al., 2008). A graphical representation of the approach is presented in Figure 2.3.

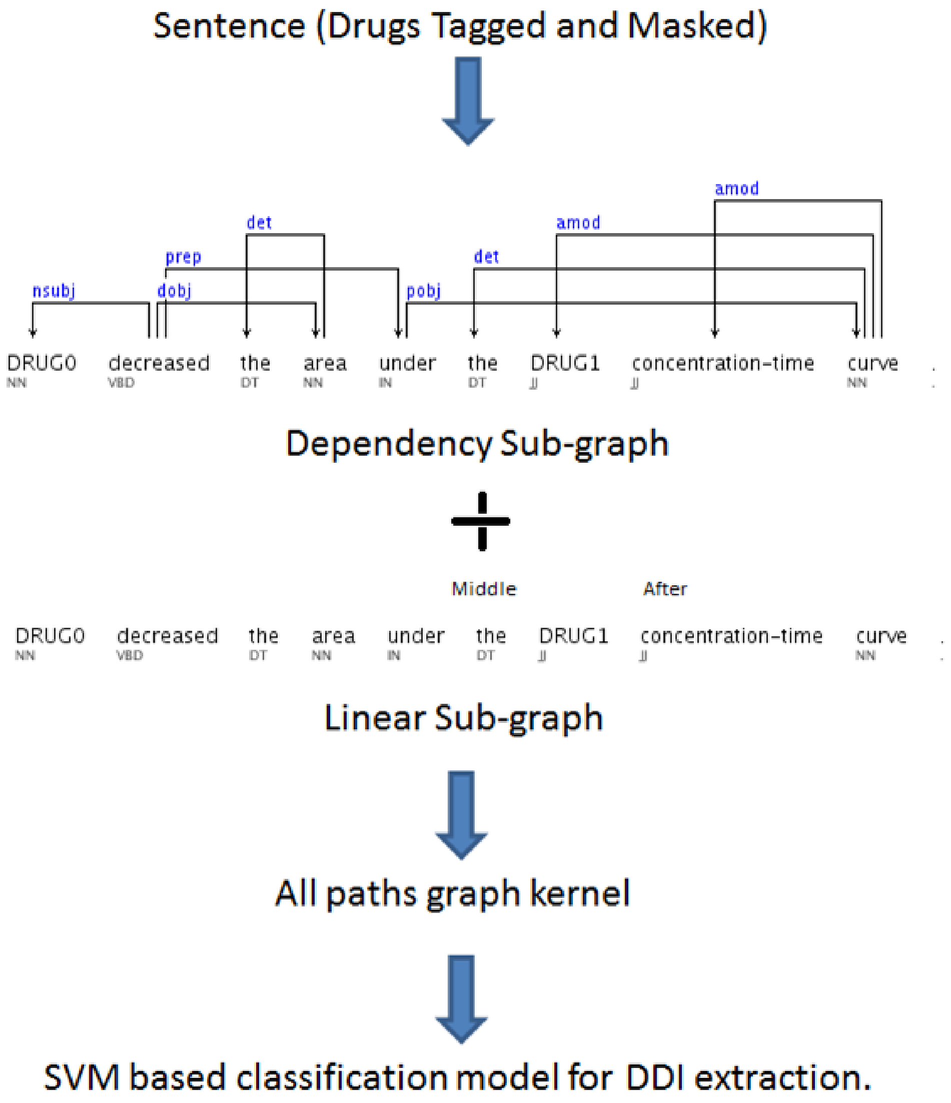


Figure 2.3 Drug Interaction Extraction Algorithm Flow Chart

DDI extraction was implemented in the in vitro and in vivo DDI corpus separately. Table 2.11 presented the training sample size and testing sample size in both corpus sets. Then Table 2.12 presents the DDI extraction performance. In extracting in vivo DDI pairs, the precision, recall, and F-measure in the testing set are 0.67, 0.79, and 0.73, respectively. In the in vitro DDI extraction analysis, the precision, recall, and F-measure are 0.47, 0.58, and 0.52 respectively in the in vitro testing set. In our early DDI research published in the DDIExtract 2011 Challenge (Karnik, Subhadarshini, Wang, Rocha, & Li, 2011), we used the same algorithm to extract both in vitro and in vivo DDIs at the same time, the reported F-measure was 0.66. This number is in the middle of our current in vivo DDI extraction F-measure 0.73 and in vitro DDI extraction F-measure 0.52.

Error analysis was performed in testing samples. Table 2.13 summarized the results. Among the known reasons for the false positives and false negatives, the most frequent one is that there are multiple drugs in the sentence, or the sentence is long. The other reasons include that there is no direct DDI relationship between two drugs, but the presence of some words, such as dose, increase, and etc., may lead to a false positive prediction; or DDI is presented in an indirect way; or some NDDI are inferred due to some adjectives (little, minor, negligible).

Datasets	Abstracts	Sentences	DDI Pairs	True DDI Pairs
in vivo DDI training	174	2112	2024	359
in vivo DDI testing	44	545	574	45
in vitro DDI training	168	1894	7122	783
in vitro DDI testing	42	475	1542	146

Table 2.11 DDI Data Description

Datasets	Precision	Recall	F-measure
in vivo DDI Training	0.67	0.78	0.72
in vivo DDI Testing	0.67	0.79	0.73
in vitro DDI Training	0.51	0.59	0.55
in vitro DDI Testing	0.47	0.58	0.52

Table 2.12 DDI Extraction Performance

No.	Error Categories	Error type	Frequency		Examples
			In vivo	In vitro	
1	There are multiple drugs in the sentence, and the sentence is long.	FP	6	34	PMID: 12426514. In 3 subjects with measurable concentrations in the single-dose study, rifampin significantly decreased the mean maximum plasma concentration (C(max)) and area under the plasma concentration-time curve from 0 to 24 h [AUC(0–24)] of praziquantel by 81% (P <.05) and 85% (P <.01), respectively, whereas rifampin significantly decreased the mean C(max) and AUC(0–24) of praziquantel by 74% (P <.05) and 80% (P <.01), respectively, in 5 subjects with measurable concentrations in the multiple-dose study
		FN	2	17	PMID: 10608481. Erythromycin and ketoconazole showed a clear inhibitory effect on the 3-hydroxylation of lidocaine at 5 microM of lidocaine (IC50 9.9 microM and 13.9 microM, respectively), but did not show a consistent effect at 800 microM of lidocaine (IC50 >250 microM and 75.0 microM, respectively).
2	There is no direct DDI relationship between two drugs, but the presence of some words, such as dose, increase, and etc. may lead to a false positive prediction	FP	6	14	PMID: 17192504. A significant fraction of patients to be treated with HMR1766 is expected to be maintained on warfarin
3	DDI is presented in an indirect way.	FN	2	19	PMID: 11994058. In CYP2D6 poor metabolizers, systemic exposure was greater after chlorpheniramine alone than in extensive metabolizers, and administration of quinidine resulted in a slight increase in CLoral.
4	Design issue. Some NDDI are inferred due to some adjectives (little, minor, negligible)	FP	1	3	PMID: 10223772. In contrast,the effect of ranitidine or ebrotidine on CYP3A activity in vivo seems to have little clinical significance.
5	Unknown	FP	5	44	PMID: 10383922. CYP1A2, CYP2A6, and CYP2E1 activities were not significantly inhibited by azelastine and the two metabolites.
		FN	6	26	PMID: 10681383. However, the most unusual result was the interaction between testosterone and nifedipine.

Table 2.13 DDI Extraction Error Analysis from Testing DDI Sets

2.4 Conclusions and Discussions

A comprehensive PK ontology was constructed. It annotates both in vitro PK experiments and in vivo PK studies. Using our PK ontology, a PK corpus was also developed. It consists of four classes of PK studies: in vivo PK studies, in vivo PG studies, in vivo DDI interaction studies, and in vitro DDI studies. This PK corpus is a highly valuable resource for text mining drug interactions relationship.

We previously had developed entity recognition algorithm or tools to tag PK parameters and their associated numerical data (Wang et al., 2009). We had shown that for one drug, midazolam, we have achieved very high accuracy and recall rate in tagging PK parameter, clearance (CL), and its associated numerical values. However, using our newly developed PK corpus, we cannot regain such a good performance in a more general class of drugs and PK parameters. This area will need much further investigation.

We would like to acknowledge that a DDI Corpus was recently published as part of a text mining competition DDIExtraction 2011 (http://labda.inf.uc3m.es/ DDIExtraction2011/dataset.html). Their DDIs were clinical outcome oriented, not PK oriented. They were extracted from DrugBank, not from PubMed abstracts. Our PK corpus complements to their corpus very well.

Chapter 3. Clinical Pharmacodynamics Drug Interaction Corpus

3.1 Background

Drug-drug interactions pose a significant challenge in current medicine, leading to adverse drug reactions, emergency room visits and hospitalizations (L. B. Becker, Kallewaard, M., Caspers, P.W., Visser, L.E., Leufkens, H.G., et al. , 2007; M. J. Hall, DeFrances, Williams, Golosinskiy, & Schwartzman, 2010; Nisha, 2010). Translational DDI research aims to link between knowledge of molecular mechanisms underlying DDIs and their clinical consequences. Three types of evidence indicate drug interaction (Hennessy & Flockhart, 2012): in vivo, in vitro and clinical. While clinical evidence forms the ultimate DDI alert, in-and-of itself it does not provide information about molecular mechanisms underlying interactions, and as such, does not suggest alternative treatments that circumvent DDIs. Therefore, for newly developed drugs, the FDA requires in vitro and in vivo DDI studies (L. Zhang, Reynolds, K. S., Zhao, P., Huang, S.M., 2010; L. Zhang, Zhang, Y., Zhao, P., Huang, S.M., 2009). An integrated simultaneous view of all three types of DDI evidence was shown effective in reducing false DDI predictions (R. Boyce, Collins, C., Horn, J., and Kale, I., 2009; R. Boyce, Collins, C., Horn, J., Kalet, I., 2009). However, due to the diversity of disciplines involved in the study of DDI along the different levels (Hennessy & Flockhart, 2012), the three types of evidence are not all available nor are they presented together.

The medical informatics and text mining research communities have invested much effort toward developing standards and tools to extract drug interaction evidence from a variety of sources, such as FDA labels and Medline abstracts (R. Boyce et al., 2009b; Herrero-Zazo, Segura-Bedmar I Fau - Martinez, Martinez P Fau - Declerck, & Declerck, 2013; Kolchinsky, Lourenco A Fau - Li, Li L Fau - Rocha, & Rocha, 2013; Segura-Bedmar, Martinez P Fau - de Pablo-Sanchez, & de Pablo-Sanchez, 2011; L. Tari et al., 2010; H.-Y. Wu, Chiang, & Li, 2014; H. Y. Wu, Karnik S Fau - Subhadarshini, et al., 2013). A fundamental step of mining drug interaction evidence is to construct gold-standard annotations of existing drug interaction evidence. Herrero-Zazo et al created a large DDI corpus, based on 792 entries from the DrugBank database and 233 Medline abstracts (Herrero-Zazo et al., 2013). This comprehensive DDI corpus contains both pharmacokinetics and pharmacodynamics DDI evidence. Independently and at about the same time, our group has published another DDI corpus (H. Y. Wu, Karnik S Fau - Subhadarshini, et al., 2013), based on 218 Medline abstracts discussing in-vivo DDI evidence (where DDI led to drug concentration change), and 210 abstracts discussing in-vitro DDI evidence (where DDI changes enzyme activities). Our work also included an ontology for characterizing pharmacokinetics aspects of DDI, as well as the specific evidence in the context of in vitro experiments and of in vivo studies. That ontology clearly differentiated these two types of DDI evidence.

One important type of drug interaction evidence comes from epidemiology studies, in which co-committed drugs are compared to single drugs on their efficacy and/or adverse drug events (ADEs). These clinical pharmacodynamics studies are usually conducted within the large health record databases (Duke et al., 2012). The corpus developed by Herrero-Zaro and Segura-Bedmar did not stress on the pharmaco-epidemiological studies and their statistical DDI evidences, nor did the corpus developed by Wu and Li (H. Y. Wu, Karnik S Fau - Subhadarshini, et al., 2013). This paper aims to address the knowledge gap of epidemiologic DDI evidence in the DDI corpus construction.

3.2 Material and Methods

3.2.1 Overall Data Curating and Annotation Description

In this clinical pharmacodynamics corpus, two curators conducted the clinical pharmacodynamics DDI abstract selection and DDI evidence annotation. One curator is a clinical pharmacology post-doc fellow, who has extensive training in both molecular pharmacology and clinical pharmacology. The other curator is a third year health informatics Ph.D. student, who has extensive skills in text mining and corpus construction. The second curator also has good knowledge on the drug interaction research. In addition, Dr. Lang Li designed and supervised all the corpus construction process. Dr. Li holds the full professorship in the Department of Medical and Molecular Genetic, Division of Clinical Pharmacology, Department of Biostatistics, and Department of Bio-Health Informatics. He has more than 85 published paper related to pharmacology research, and 43 of them are drug interaction related. Dr. Li's extensive expertise in pharmacology, epidemiology, biostatistics, and informatics ensures the scientific relevance and data quality of this corpus construction.

The corpus construction was conducted in three steps (Figure 3.1). During the abstract screening process (Step 1), given a list of keywords, two searching strategies (journal-specific search and PubMed search) were conducted separately for screening abstracts. After removing duplicates between two searches, the abstracts were ready for the validation process (Step 2), inclusion-exclusion criteria (IECs) were predefined to filter the abstracts from Step 1. Based on these IECs, both curators validated the abstracts, and the disconcordant selections were judged by Dr. Li. In the final step, both curators annotated all DDI entities, sentences, and relationships in the abstracts. Detailed description on corpus construction is illustrated below.

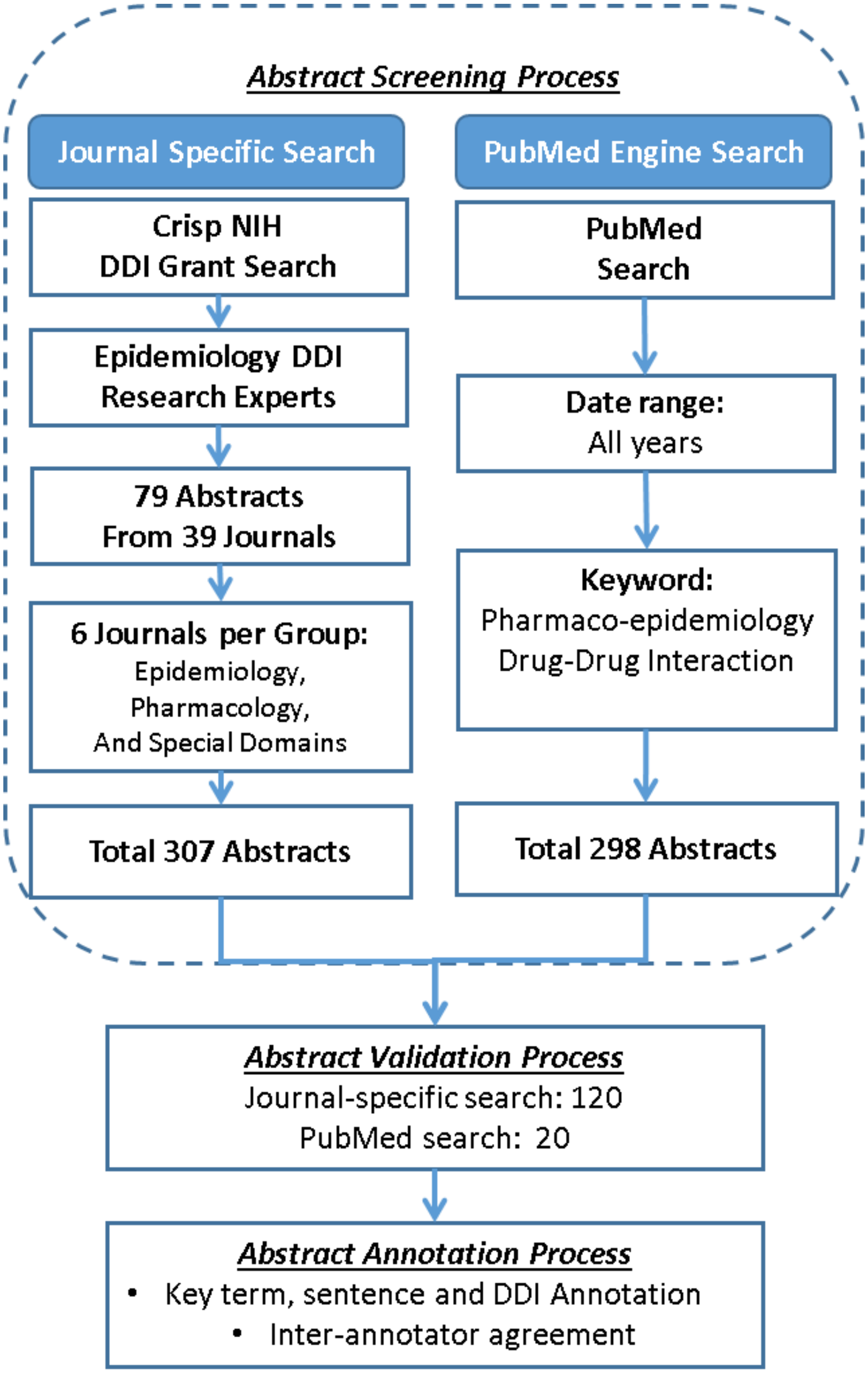


Figure 3.1 Flowchart of Clinical Pharmacodynamics Drug Interaction Abstract Screening, Quality Control, and Annotation Process

Step 1: Abstract Screening Process The DDI abstracts are made up of texts from two different searching strategies: Journal-specific search and PubMed Search.

Journal-specific search started from clinical pharmacodynamics drug interaction research experts, who have active National Institute of Health funded grants focusing on the clinical pharmacodynamics and informatics related drug interaction research. From the NIH grant reporting system, CRISP (http://projectreporter.nih.gov/reporter.cfm), we identified four current active grants: DK102694, GM104483, LM011838, and GM107145. The PI names are Sean Hennessy, Lang Li, Luis Rocha, Hagit Shatkay, Richard Boyce, and Nicholas P. Tatonetti, respectively. The first set of drug interactions abstracts were selected from their publications and the references within their publications. There were 79 epidemiological drug interaction studies and their abstracts from 39 journals (The detail can be seen in Table S3 of Supplementary Material section). To explore extra clinical pharmacodynamics DDI papers, we focused on those 39 journals referenced from the first set of clinical pharmacodynamics DDI papers. These journals were categorized into three groups: pharmacology, epidemiology, and specific disease areas. Then 6 journals were selected from each group. Using the search strings, (“drug-drug interaction”, “clinical pharmacodynamics”, “epidemiology”, “medical record” and “concomitant”), the second set of DDI abstracts were selected from each journal. During this screening process, an abstract was selected as relevant if it mentioned drug interaction based on either health record based retrospective studies or prospective clinical trials or observational studies.

Searching in Medline (PubMed) was another strategy to retrieve clinical pharmacodynamics DDI abstracts. If using the same search terms as in journal specific search, only two abstracts were retrieved from PubMed. Therefore, the key terms were loosen to “(drug-drug interaction)AND(pharmacoepidemiology)”.

Step 2: Abstract Validation Process Having retrieved relevant abstracts from two search pipelines, two curators further reviewed those abstracts. A list of inclusion and exclusion criteria (IEC) were predefined. There were two inclusion criteria: 1) Only abstracts that reported a test of a drug-interaction hypothesis, i.e. testing whether two co-committed drugs show a different clinical outcome than one drug alone, were further selected as the candidate abstracts for the corpus. 2) The second inclusion criterion was expected to place restrictions on those clinical pharmacodynamics related articles. On the other hand, exclusion criteria filtered out those articles that studied single drug study, drug efficacy comparison, co-medication frequency, drug compliance, drug-alcohol/food interaction and the articles related to pharmacokinetics, pharmacodynamics, pharmacogenetics, in vivo, and in vitro studies. In addition, we excluded case report studies in that DDI situation in a single case report cannot represent their DDI effect on population. The detail of IECs information is listed in Table 3.1.

Inclusion	1. Drug1/drug2 together can increase ADR risk or efficacy than the single drug alone.
	2. The distribution of known DDIs in different patient population defined by age, country, disease, or ADR
Exclusion	a) Even two drugs are mentioned, but the comparisons are conducted between two drugs.
	b) Only reported the co-medication frequencies, not their effect.
	c) Drug interaction detection algorithms or software
	d) Compliance of avoiding DDI
	e) Concordance of DDI reporting among different system
	f) Comparison the performance of DDI clinical decision system
	g) Single drug study without mentioning the DDI
	h) Drug-alcohol/food interaction
	i) Drug/test interaction
	j) DDI in PK, PD, PG, in vivo, and in vitro study
	k) Case report study
	l) Review paper

Table 3.1 Inclusion-Exclusion Criteria (IEC) for Abstract Validation Process

Step 3: Annotation Scheme BRAT rapid annotation tool (Stenetorp et al., 2012) was used to annotate the corpus. It has three level annotations: term level, sentence level, and DDI level. This structured three-layer annotation scheme started from the annotation of basic entities (clinical pharmacodynamics related terms), then the sentences related to DDI, and finally interacting drug pairs in the sentences. The procedure of our annotation scheme is further detailed in the follow-up paragraphs.

- The boundaries of the individual terms denoting such entities were defined as follows: Drug names were defined mainly on DrugBank 3.0 (Knox et al., 2011) and MeSH (Rogers, 1963) Term. In addition to ordinary drug names from DrugBank, MeSH contains some classes of compounds typically used in the treatment of a specific category of disease or disorder. For example, selective serotonin re-uptake inhibitors (SSRIs) for depression treatment have many names including, for instance, citalopram, escitalopram, fluoxetine, paroxetine, sertraline. In many clinical pharmacodynamics articles, the term SSRI was used as opposed to any of the generic names.

- Clinical Endpoints and Their Values Clinical endpoints generally refer to phenotypic measurements of evidence of disease (e.g. disease incidence), drug efficacy (e.g. patient overall survival), drug side effects (e.g. myopathy or neuropathy), and laboratory tests (e.g. CK values or blood pressure). These clinical endpoints demonstrate the clinical significance of the drug interaction evidence. Clinical endpoint values are quantifications of the clinical endpoints. For examples, incidence rate for the disease, survival rate for the drug efficacy, and frequency of the drug side effects. Lab test values are usually numerical.

- Statistical Models and Their Values Statistical models are used to evaluate the statistical significance of the drug interaction evidence. Usually, a linear regression model is used to assess the continuous clinical endpoint points, e.g. lab tests; a logistic regression is used to assess the binary clinical endpoints, such as disease or drug side effect; and a Cox proportional hazard model is used to assess the time to event clinical endpoint, such as overall survival. The regression coefficients are always used to quantify the clinical effect of drug interactions. They are named specifically as the odds ratio (OR) and the hazard ratio (HR) in the logistic regression and Cox proportional hazard regression model, respectively. The statistical evidences are presented with p-values. DDI evidence of a p-value less than 5% was classified as a DDI; otherwise it was classified as a NDDI (i.e. non-DDI). For example: Comparing mini-dose warfarin and warfarin plus aspirin, the annual rates of major bleeding was 0.3% and 1.4% respectively, (P = .20). In this case, warfarin and aspirin interaction was classified as a NDDI.

- Change Term describes the change of action in clinical endpoints and their values. The following words were annotated in the corpus to denote its action: increase, alter, elevate, induce, higher, change, decrease, reduce, attenuate, lower, exacerbate, and etc (H. Y. Wu, Karnik S Fau - Subhadarshini, et al., 2013).

The middle level annotation focused on the identification of drug interaction sentences. The criteria of determining DDI sentence is based on the existence of drug pairs with an interaction statement, such as an OR or an HR for the change a clinical endpoint when comparing co-committed drugs to a single-drug treatment. The interaction statement might be either an interaction, or a non-interaction.

Once the DDI sentences were labeled, DDI pairs in the sentences were further annotated. Differed from DDI pair annotation in our pharmacokinetics DDI corpus (H. Y. Wu, Karnik S Fau - Subhadarshini, et al., 2013), DDI relationships in clinical pharmacodynamics corpus are only classified into two classes (DDI and Non-DDI).

3.2.2 Inter-Annotator Agreement

To access the quality of the corpus construction and the consistency of the annotation task, the inter-annotator agreement was measured for the processes of both validation (Step 2) and annotation (Step 3) tasks. In the phase of abstract validation, two curators executed the abstract selection independently following the predefined IECs. Any disagreements were firstly investigated and discussed between two curators. If the consensus was not achieved, the disagreements were resolved by Dr. Li, the supervisor curator. In the annotation task, those partially or none overlapped entities, sentences, and/or DDI relationships, were firstly investigated by two curators first. Any disagreements were finally judged by the supervisor curator.

3.3 Results

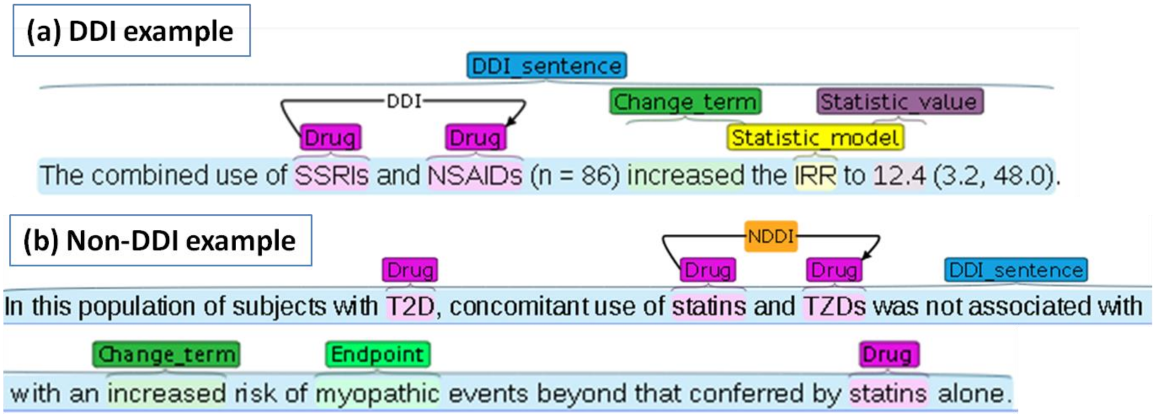
3.3.1 Clinical Pharmacodynamics Drug Interaction Abstract Screening

We retrieved 307 abstracts from relevant journals and 315 abstracts from PubMed engine. After removing 17 duplicates between them, there were 605 unique abstracts.

To ensure the quality of data curation, quality Control and Clinical pharmacodynamics DDI Abstract Validation are implemented by two curators using the IEC defined in Table 3.1. They disagreed only on 14 out of 605 previously selected abstracts. Those disagreements were further reviewed and resolved by the supervisor curator. As the result of this quality control analysis, 465 abstracts from the screening stage were classified as not clinical pharmacodynamics drug interaction studies, while the other 140 abstracts were DDI related. Those none clinical pharmacodynamics DDI abstracts include population studies that investigated the frequency of known drug interactions in the health databases, implementations of drug interaction software, population drug safety study without testing DDI hypothesis, care reports and review articles. In Table S4, one clinical pharmacodynamics DDI example and four different types of misclassified abstracts from our screening step are presented.

3.3.2 Corpus Annotation Representation and Its Statistics

The corpus was constructed on the 140 DDI-related abstracts with Brat rapid annotation tool, (http://caarray.compbio.iupui.edu/brat/#/PharEpi_corpus/). Based on our proposed annotation scheme, the corpora of these abstracts were manually carried out by two independent annotators. Figure 3.2 shows two typical examples of complete sentences describing DDI annotation using the brat rapid tool. In Figure 3.2 Two Examples of DDI Annotation, one interacting pair (SSRIs and NSAIDs) was created according the “Change term” on “their statistical model and value”. In Figure 3.2(b), a non-interaction relationship (statins-TZD) is identified due to the negation in “Change term” of “endpoint” compared with the effect of statins alone. According to both examples, all the semantic entities including drug name, clinical endpoint and its value, statistical models and their values, and change term are necessary and critical to identify a DDI or a non-DDI in a clinical pharmacodynamics study.



The combined use of SSRIs and NSAIDs (n = 86) increased the IRR to 12.4 (3.2, 48.0).

In this population of subjects with T2D, concomitant use of statins and TZDs was not associated with

with an increased risk of myopathic events beyond that conferred by statins alone.

Figure 3.2 Two Examples of DDI Annotation

In the corpus, drug names (magenta), change terms (lime green), statistical models (yellow), and statistical values (dark magenta) were annotated in the term level; and DDI sentences (blue) was annotated in the sentence-level. To indicate drug-drug interaction, an arrow was used to link two interacting drugs. Figure 3.2(a) shows a drug combination increases the IRR to 12.4, which infers the information of DDI. Figure 3.2(b) shows the concomitant use of statin and TZD did not increase the risk of myopathic event. Therefore, it is a NDDI sentence.

Table 3.2 displays the statistics for the agreement of the annotation task in the DDI corpus. It comprises of 2181 annotated entities, 297 DDI sentences, and 393 drug pairs. Among entity annotations, the most common semantic type was drug (41.9%) following by endpoint (17.3%), percentage (11.4%), change term (9.8%), and statistical value (8.9%). Those numbers show the recognition of DDI relationship was dominated by those important entities. In addition, Table 3.2 presents the results for the annotation concordance. The statistics shows very high agreements for term-level annotation (2170/2181), sentence-level annotation (285/297), and DDI-pair annotation (378/393).

Annotation type	Semantic type	Number	Concordance
Key Term Level (2181 terms)	Drug names	913	2170
	Endpoint	378	
	Endpoint Value	33	
	Statistical Model	149	
	Statistical Value	194	
	P-value	52	
	Percentage	249	
	Change Term	213	
Sentence Level (297 sentences)	DDI sentence	297	285
DDI Level (393 pairs)	DDI pair	341	378
	NDDI pair	52	
Number of DDI Abstracts		140	126

Table 3.2 Statistics of Clinical Pharmacodynamics DDI Corpus Annotation

3.4 Discussions

This paper presented a corpus construction for the clinical pharmacodynamics drug interaction evidences published in the Medline. It fills in the gap of current drug interaction corpus development, and will facilitate the future drug interaction text mining, especially for the clinically relevant drug interaction evidences. The following paragraphs discusses the limitations, the challenges, and potential new informatics researches based on this corpus construction.

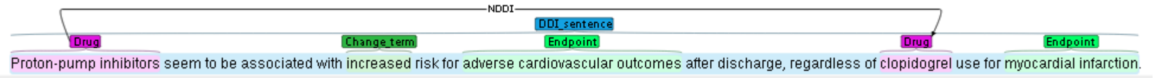
3.4.1 How Much Clinical Pharmacodynamics Drug Interaction Information in The Full Text Do We Miss When We Have Only DDI Information in The Abstracts?

We shall all keep it in mind that the annotated clinical pharmacodynamics drug interaction evidence published in the abstract is only a tip of the iceberg among all the drug interaction information in the full article. Usually an abstract would not illustrate the design of an epidemiology study. Hence, we miss the patient population information specified in its inclusion and exclusion criteria. In the abstract, we also miss the case control match-up definition in the design and covariates justification in the regression models, which balance and justify confound biomedical and demographic variables, respectively. Moreover, we miss the health record database description. Thus, we cannot assess the inherited limitation on the drug interaction evidence. All these additional DDI information from the full article are critical, if the drug interaction evidence will be assessed and implemented into the clinical setting.

3.4.2 Challenges in Pharmaco-epidemiological DDI Abstract Selection and DDI Annotations

During the abstract screen process, we found that DDI information in the clinical pharmacodynamics literatures is imbalanced. Comparing two searching strategies, only 20 abstracts were acquired from the PubMed search, while 120 came out of the journal search. Thus, we would recommend a combo strategy in selecting the DDI abstracts. While the general PubMed search can keep the generality of query, the journal specific search can have high sensitivity.

In annotating clinical pharmacodynamics DDI information, we recognized that there were two difficult situations: term abbreviation and particular sentence representation. Many abstracts used the abbreviation to represent drug names, endpoints, side effect, and disease. They confused the curators frequently and led to errors during the annotation process. Unlike the explicit examples in Figure 3.2, some DDI sentences represent DDI information with indirect statement, and it causes the disagreement between two curators. For example, in Figure 3.3, the DDI sentence stated that "proton-pump inhibitors are associated with increased risk of ADR, except for clopidogrel". One curator missed the link between the drug pair of proton-pump inhibitors and clopidogrel due to the indirect DDI statement.



Proton-pump inhibitors seem to be associated with increased risk for adverse cardiovascular outcomes after discharge, regardless of clopidogrel use for myocardial infarction.

Figure 3.3 The Indirect NDDI Statement and Relationship

3.4.3 Drug Interaction Knowledge Gap Issue

As diverse disciplines and varied studies for DDI are involved, drug interaction evidence is often not available cross all different types of research. It creates the knowledge gap and impedes the translational DDI research from molecular pharmacology to clinical pharmacology. To examplify such gap, we extend current drug interaction corpus to include clinical pharmacodynamics studies. We applied the same construction strategy as our previous drug interaction work (Pharmacokinetics (PK) drug interaction Corpus) to the current version of DDI corpus.

Compared with the PK corpus, the presentation of DDI in clinical pharmacodynamics studies is not the same as that in vivo/in vitro drug interaction experiments. In vivo and in vitro experiments measure different endpoints (e.g. AUC and clearance) and describe the drug response with distinctive models (e.g. one compartment or two compartment models). On the other hand, epidemiology studies investigate the effects of drug interaction on diseases or ADR conditions in predefined patient populations. To identify risk factors for DDI, clinical pharmacodynamics studies focus on the study design, data collection, statistical analysis, and the interpretation of results. Based on the statistics in Table 3.2, there are about two DDI sentences and 2.8 DDI pairs on average in each abstract. In our PK corpus, there are more than 3 sentences (1311/428) and over 4 DDI pairs (1833/428) on average in each abstract. These numbers demonstrate the different sentence structure and DDI representation between clinical pharmacodynamics and pharmacokinetics DDI studies. As the main goal of the DDI corpus construction is to develop text mining tools for the large scale DDI extraction, we anticipate that the text mining algorithms will be different based on different DDI corpora. In our early work, we demonstrated that the text mining algorithms and their performances were different between in vitro PK and in vivo PK DDI corpora. We expect the similar trend of text mining performances between clinical pharmacodynamics DDI studies and PK drug interaction studies.

3.5 Conclusion

In summary, clinical pharmacodynamics DDI corpus was constructed via streamlined abstract screening, validation, and annotation processes. The DDI abstracts were screened based on both journal based search and PubMed general search; the DDI abstracts was validated based on predefined IECs; and a three-layer annotation scheme guided the entities, sentence, and interaction annotations. To the best of our knowledge, this is the first well annotated corpus for clinical pharmacodynamics DDI studies. It closes the knowledge gap cross all different types of DDI evidence, and provides the NLP community an unique opportunity to develop text mining tools to extract pharmaco-epidemiological DDI evidences.

Chapter 4. Named Entity Recognition Method for Drug Metabolite

4.1 Background

Drug metabolism, distribution, and excretion are the primary pharmacokinetics research areas. A drug's pharmacokinetics (PK) involves not only the parent compound, but also its metabolites (Malcolm Rowland et al., 2011). In some instances, an active drug metabolite can retain enough or even dominate its intrinsic activity at target receptor and contribute to the pharmacological effects. Certain drugs such as codeine and losartan have active metabolites (morphine and EXP3174 respectively) that are responsible for more therapeutic action than their parent drugs (Obach, 2013). In some other instances, pro-drugs, formulated in an inactive form, are designed to be metabolized inside the body to form the active drugs (Hacker et al., 2009). A salient example is tamoxifen, which itself is not an active compound to treat breast cancer. Instead, its metabolites, 4-OH-tamoxifen and endoxifen are potent inhibitors to estrogen alpha (Desta et al., 2004; Johnson et al., 2004; Lee et al., 2003; Stearns et al., 2003). Drug metabolites also play very interesting roles in drug interactions. A notable example is itraconazole. Itraconazole itself is a potent CYP3A inhibitor, so are its metabolites, such as hydroxy-itraconazole, keto-itraconazole, and N-desalkyl-itraconazole (Isoherranen et al., 2004). The metabolism of CYP3A substrates, such as the midazolam, are inhibited by itraconazole and its metabolites, if midazolam and itraconazole are taken together. Pharmacogenetics, another forefront of pharmacology research, also has a major impact on the drug metabolism products. Using the previous tamoxifen example, tamoxifen active metabolite, endoxifen, is generated through the CYP2D6 enzyme. Among breast cancer patients with CYP2D6 loss functional variants (e.g. *4, *5, and*10), the patients usually have very limited tamoxifen metabolite, endoxifen. Hence, these patients have much reduced endoxifen concentration such that the efficacy of tamoxifen treatment declined (Stearns et al., 2003). All these above examples demonstrate that drug metabolites and their parent drugs are equally important in pharmacokinetics research.

Although there are a number of well-established dictionaries for drug names, such as DrugBank, MeSH terms, Rx-Norm, NDC, PubChem, etc, there is very limited naming system for drug metabolites. In particular, we want to make a distinction between drug metabolites and metabolome, which is considered to be the collection of all metabolites in a biological cell, tissue, organ, or organism. Metabolome may include both endogenous metabolites that are naturally produced by an organism (such as amino acids, organic acids, nucleic acids, fatty acids, amines, sugars, vitamins, co-factors, pigments, antibiotics, etc.) as well as exogenous chemicals (such as drugs, environmental contaminants, food additives, toxins and other xenobiotics) that are not naturally produced by an organism (Nordstrom et al., 2006; Wishart, 2007). Therefore, ideally, metabolome shall include drug metabolites. However, due to the limitation of Mass-Spectrometry (MS) or Nuclear Magnetic Resonance (NMR) biotechnologies, metabolome studies and drug metabolisms studies are conducted using very different methodologies. Drug metabolites are usually measured with validated drug internal standard as the internal reference using MS technologies. Metabolome studies, on the other hand, rarely rely on drug internal standards. Therefore, drug metabolites rarely can be found from metabolome studies, because they are different metabolites. For instance, the highly populated Human Metabolome Database (HMDB) reports data on >29,000 endogenous metabolites, but there are only 2485 drugs, and 948 drug metabolites (Wishart et al., 2013). Other examples are DrugBank 4.0 (Law et al., 2014) and ChEBI (Degtyarenko et al., 2008), comprising of only 1,445 and 111 drug metabolites respectively, which are much less than the total number of generic drugs (8,184).

Although HMDB, Drugbank, and ChEBI provide the limited drug metabolite terminologies, most drug metabolites were reported and can be found through scientific literature, especially pharmacology-related articles. To capture these drug metabolite names from text, named entity recognition (NER) shall be utilized. Already there are many NER tools that enable to enrich text with semantic annotations for biomedical or biological terminologies (Alias-i, 2008; Björne, Kaewphan, & Salakoski, 2013; David, Sérgio, & José Luís, 2012; Eltyeb & Salim, 2014; Fukuda, Tamura, Tsunoda, & Takagi, 1998; Krauthammer, Rzhetsky, Morozov, & Friedman, 2000; Leaman & Gonzalez, 2008; McDonald & Pereira, 2005; Nadeau & Sekine, 2007; Neves & Leser, 2014; Nobata et al., 2011; Rebholz-Schuhmann, Arregui, Gaudan, Kirsch, & Jimeno, 2008; Rocktaschel, Weidlich, & Leser, 2012; Segura-Bedmar, Martinez P Fau - Segura-Bedmar, & Segura-Bedmar, 2008; Settles, 2005; Usie, Alves, Solsona, Vazquez, & Valencia, 2014; Vazquez, Krallinger, Leitner, & Valencia, 2011; G. Zhou, Zhang, Su, Shen, & Tan, 2004). Among those systems, most were designed to identify general biological terms such as proteins, DNA, RNA, cells, cell lines, etc (Alias-i, 2008; David et al., 2012; Leaman & Gonzalez, 2008; Settles, 2005; Tsuruoka & Tsujii, 2004), and some of those can annotate drugs, chemicals, or metabolome. Only a few use the dictionary lookup approach to annotate drug metabolites in ChEBI and HMDB (Björne et al., 2013; Eltyeb & Salim, 2014; Nobata et al., 2011; Rebholz-Schuhmann et al., 2008; Rocktaschel et al., 2012; Segura-Bedmar et al., 2008; Usie et al., 2014; Vazquez et al., 2011). For instance, Whatizit is a text mining system with a suite of modules that analyze text data based on TreeTagger and identify a set of selected annotation types based on publicly available resource (Rebholz-Schuhmann et al., 2008). Within the Whatizit, WhatizitChebiDict annotates ChEBI entities based on the dictionary search, and whatizitOSCAR3 identifies chemistry-specific terms using an approach that combines n-grams, regular expression, and heuristic rules (Corbett & Murray-Rust, 2006). Integrating both drug and chemical terms at the same time, whatizitChemical module annotates drug metabolite as it contains the annotations from both whatizitOSCAR3 and whatizitChebiDict. A research work by Chikashi et al. created a manually annotated golden standard corpus for yeast metabolome and proposed a NER tool to extract yeast metabolites using ChEBI and HMDB data as one of features (Nobata et al., 2011). This article demonstrated that whatizitChemical achieved lower precision and F-measure compared to their NER tool. Nevertheless, we think whatizitChemical's drug metabolites NER performance is restrained by the limited drug metabolite terminologies in ChEBI and HMDB databases.

There are several aspects that we shall increase the performance of the drug metabolite NER. Firstly, many drug metabolites are related with their parent drugs through chemical reactions via drug metabolism enzymes, such as CYP450 enzyme family. A drug metabolite is often times named after their drug names accompanying with a prefix or suffix substring that is related with enzyme catalysation. For example, 4OH-midazolam and 4OH-tamoxifen are metabolites of midazolam and tamoxifen through the CYP3A oxidation (Desta et al., 2004; Sevrioukova & Poulos, 2017). Secondly, abbreviations are frequently used in the biomedical literature to cite drugs and metabolites. Using the midazolam and tamoxifen as examples, they were also reported as MDZ and TAM repeatedly, and their metabolites were also written as 4OH-MDZ and 4OH-TAM, respectively. If these abbreviations can be integrated in the NER algorithm, it shall have a much better performance in recognizing not only drug names, but also their metabolites. Thirdly, if a drug and its metabolite are two different terms in a publication, they are usually very close. Therefore, a drug metabolite NER algorithm shall recognize these phenomena in order to have an improved drug metabolite detection performance.

In this article, we make two major contributions in developing an innovative and better drug metabolite NER tool. First, four different drug metabolite presentation patterns are defined, and a golden-standard corpus is constructed. This annotated corpus facilitates the next step NER algorithm development. Second, our new drug metabolite NER tool is a hybrid approach. It combines both a lexicon-based mapping and a machine-learning algorithm. This system captures both drug metabolites and their abbreviations.

4.2 Material

4.2.1 Define Drug Metabolite and Reaction

There are four patters that the drug metabolites are presented in the published literature. Tamoxifen metabolites are used as the primary example to illustrate these four patterns in Figure 4.1. The first two categories (single word drug metabolite Type I and Type II) are drug metabolite names in a single entity. Type I clearly contains a substring of a drug name as well as a chemical prefix or suffix (e.g. 4-OH-N-desmethyltamoxifen in MEDLINE: 15685451). Type II, however, does not contain either a substring of its parent drug or chemical reaction. Type II has two types of instances. The first one is an abbreviation of the drug metabolite (e.g. DFO abbreviates for dimemorfan oxidation in MEDLINE:19593786). The second one is a unique term given for its drug metabolite, and this term is unrelated to its parent drug name. For instance, endoxifen (MEDLINE: 20400308) is the primary active metabolite of tamoxifen via CYP2D6 enyzme), which has an alternative name of 4-OH-N-desmethyltamoxifen. The other two patterns are represented with the form of multi-word entities containing a preposition or conjunction for describing the chemical reaction. The examples of multi-word drug metabolite Type I and Type II are tamoxifen N-demethylation in MEDLINE: 24737844 and N-demethylation of tamoxifen in MEDLINE: 8104124, respectively.

Drug metabolite type I	4-OH-N-desmethyl-tamoxifen
Drug metabolite type II	endoxifen
Drug reaction type I	tamoxifen N-demethylation
Drug reaction type II	N-demethylation of tamoxifen

Figure 4.1 Patterns of Drug Metabolites

4.2.2 Corpus Construction

Drug metabolite corpus was constructed using 210 MEDLINE abstracts from in vitro PK corpus (H. Y. Wu, Karnik, et al., 2013). The corpus construction is a manual process (Figure 4.2). Three annotators with different training backgrounds, including informatics, biochemistry, and pharmacology, conduct the annotation tasks independently. The disagreed annotations are discussed among three annotators for consensus. If the consensus is not achieved, the disagreed annotations are further judged by pharmacological research experts (Professors in the Department of Pharmacology) for the final decision.

The annotations were restricted to those names that are involved in drug metabolism or are clearly mentioned as drug metabolites. Based on this criterion, we present few examples that describe four patters for drug metabolites accordingly. In Figure 4.3 (a), Drug metabolite type I is annotated in a sentence in PMID: 10383922. "Azelastine has been reported to be metabolized mainly to desmethylazelastine and 6-hydroxyazelastine in mammals" describes that both "desmethylazelastine" and "6-hydroxyazelastine" can be referred to as azelastine's metabolites because of the involvement of azelastine's metabolism. Another sentence in PMID: 11259331, "We have identified CYP2C19 and CYP3A4 as the principal cytochrome P450s involved in the metabolism of flunitrazepam to its major metabolites desmethylflunitrazepam and 3-hydroxyflunitrazepam", clearly mentions that "desmethylflunitrazepam" and "3-hydroxyflunitrazepam" are the metabolites of flunitrazepam. Figure 4.3 (b) shows one example of Drug metabolite type II. In this case, both dihydroqinghaosu and its abbreviation (DQHS) are the metabolite of

Artelinic acid (AL). In Figure 4.3 (c) and (d), ATRA 4-hydroxylation and hydroxylation of midazolam are examples of Drug reaction type I and type II, respectively.

To evaluate the quality of the annotation task, the measurement of inter-annotator agreement between two annotators was quantified using “Pairwise Percent Agreement”. Since any disagreements among three annotators are resolved by two supervisor annotators, the gold-standard corpus for drug metabolite is constructed. In addition, the results from three annotators are compared to the gold-standard corpus. Precision, recall, and F-measure are adopted to assess the performance of an individual annotator.

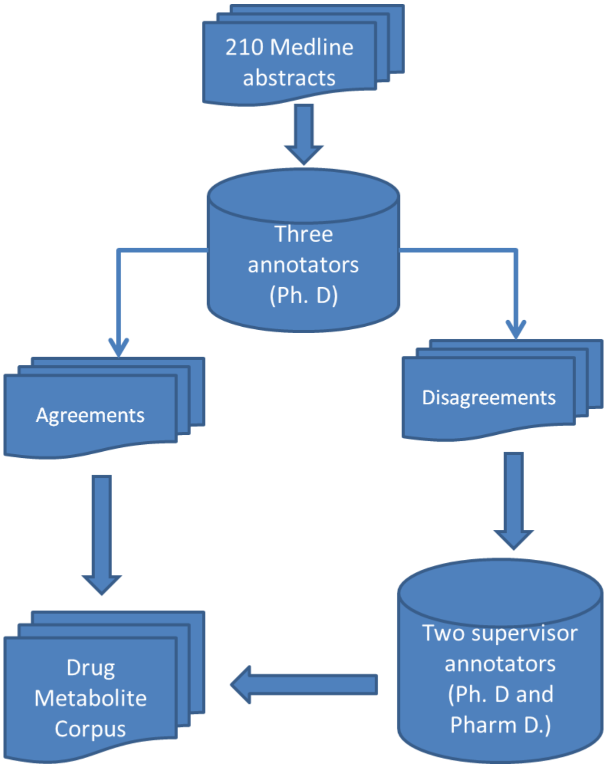
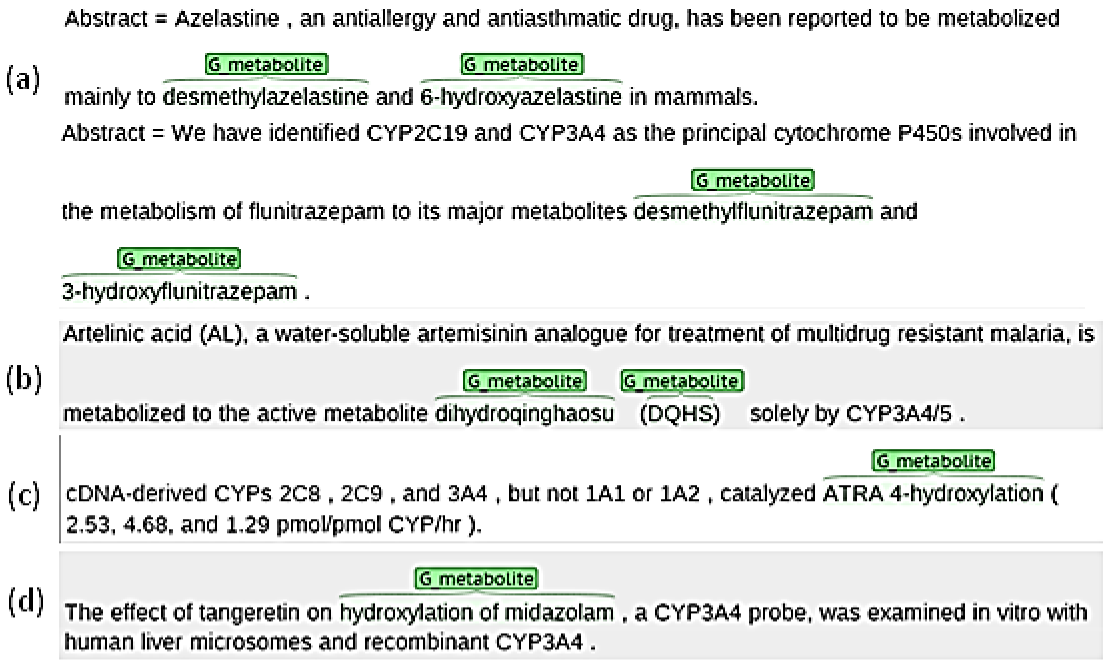


Figure 4.2 Drug Metabolite Annotation Flow Chart



Abstract = Azelastine , an antiallergy and antiasthmatic drug, has been reported to be metabolized mainly to desmethylazelastine and 6-hydroxyazelastine in mammals.

Abstract = We have identified CYP2C19 and CYP3A4 as the principal cytochrome P450s involved in the metabolism of flunitrazepam to its major metabolites desmethylflunitrazepam and 3-hydroxyflunitrazepam .

Artelinic acid (AL), a water-soluble artemisinin analogue for treatment of multidrug resistant malaria, is metabolized to the active metabolite dihydroqinghaosu (DQHS) solely by CYP3A4/5 .

cDNA-derived CYPs 2C8 , 2C9 , and 3A4 , but not 1A1 or 1A2 , catalyzed ATRA 4-hydroxylation (2.53, 4.68, and 1.29 pmol/pmol CYP/hr).

The effect of tangeretin on hydroxylation of midazolam , a CYP3A4 probe, was examined in vitro with human liver microsomes and recombinant CYP3A4 .

Figure 4.3 The Annotation of Drug Metabolite in Brat Annotation Tool

4.2.3 Drug and Drug Metabolism Reaction Lexicon Construction for Named Entity Recognition

In this NER task, two lexica are built, including Drug name lexicon and drug metabolism reaction lexicon. Drug name lexicon is built upon the drug names in Drugbank 4.0 (Knox et al., 2011) and MeSH term (MeSH). Drugbank has three types of drug names, generic (8,184), brand (17,336) and synonym (7,382). MeSH has 74,619 unique drug names. Among them, 14,946 MeSH terms are mapped to the drugbank names. In total, there are 70,712 unique drug names in the drug name lexicon.

The drug metabolism reaction lexicon are composed of 65 metabolites' prefix and suffix terms collected from the literature (Golan, 2012; Knollmann, 2011) and our previous work (H. Y. Wu, Karnik, et al., 2013). They are further evaluated by two domain experts. Within the lexicon, drug metabolism reactions are categorized into two groups: modification (phase I) and conjugation (phase II) reactions. The phase I metabolism includes oxidation, reduction, and hydrolysis. Phase I metabolism causes a small structural change to a drug, which is called phase I derivatives. Then the phase I derivatives go through further phase II metabolism. Phase II metabolism includes glucuronidation, sulphation, glutathione conjugation, amino acid conjugation, acetylation, and methylation. The drug metabolism reaction lexicon is available in Table S5.

4.3 Method

4.3.1 Overview of An Integrated Drug Metabolite Named Entity Recognition Algorithm

Our drug metabolite named entity recognition algorithm has three phases. The first phase is to create features for entities in text. Part-of-Speech (PoS) feature for each entity is assigned using PoS Parser in OpenNLP (Albright et al., 2013). In parallel, a dictionary-based tagging is applied to identify weather an entity is a drug name and find whether drug name and metabolite's reaction terms are the substring of that entity. Once drug names are recognized, their abbreviations in the same abstract, if available, will be also recognized using our proposed abbreviation detection method and then tagged as drug names (see Drug Abbreviation Detection). In this phase, the outcome for each entity will be PoS feature, Drug Index, and Pre/suffix Index. In the second phase, a searching window is created centering with a drug name entity. The window is adjusted according to predefined conditions of surrounding entities (see Window Construction and Adjustment). Within a window, the entity containing a reaction term will be recognized as candidate entities for drug metabolites. In the final phase, a supervised ML algorithm learns from the linguistic cues using POS information and the Boolean variables for drug names and reaction terms (introduced in Phase I) of entities within a searching window (adjusted in Phase II) to determine whether the candidate entities in the searching window are belonging to drug metabolites or not. The workflow of the proposed NER system is depicted in Figure 4.4.

To evaluate the performance of the NER algorithm, the identified term strings that match the start and end positions of the term strings in golden standard corpus constitute true positive (TP) predictions. If the identified terms that cannot be matched fully are false positives (FP) and terms in the corpus that were not be retrieved were false negatives (FN). Finally, the standard Information-Retrieval (IR) metrics: Precision (P), Recall (R), and F-measure (F1) were computed to measure the performance.

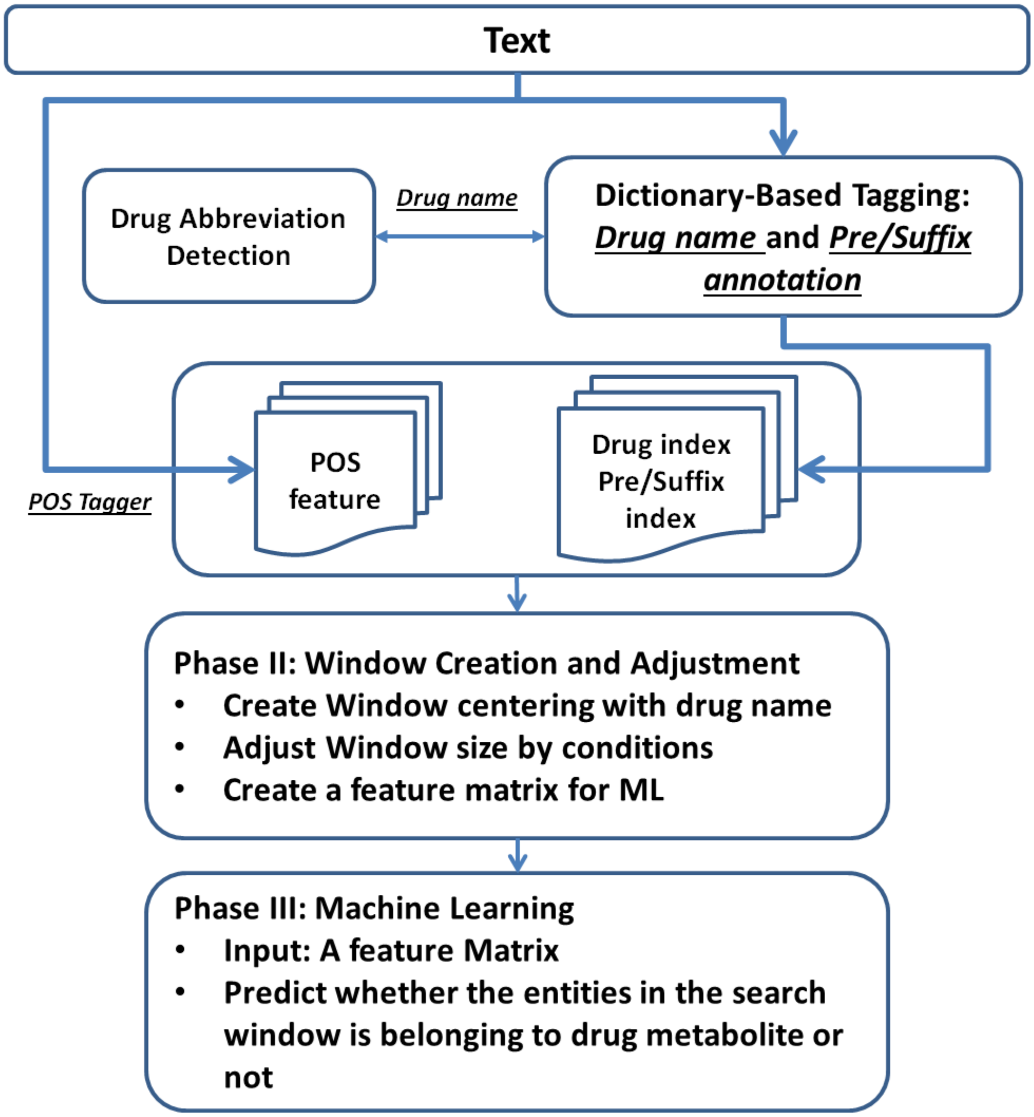

Text
Drug Abbreviation Detection
Drug name
Dictionary-Based Tagging: Drug name and Pre/Suffix annotation
POS Tagger
POS feature
Drug index Pre/Suffix index
Phase II: Window Creation and Adjustment
• Create Window centering with drug name
• Adjust Window size by conditions
• Create a feature matrix for ML
Phase III: Machine Learning
• Input: A feature Matrix
• Predict whether the entities in the search window is belonging to drug metabolite or not

Figure 4.4 Workflow of Drug Metabolite Annotation

4.3.2 Dictionary-based Drug Name and Pre/Suffix Term Tagging

Within the text, entities are tagged against both drug names and pre/suffix terms in the dictionary. Technically, drug names in dictionary are sorted based the length of string in the hash table. Then, if an entity can be partially mapped against a drug name or a reaction term in the table, drug index or Pre/Suffix index for that entity is given a one value to represent its availability. However, some entities might be erroneously tagged because of some special brand names. For instance, “Control” which is the brand name of chlordiazepoxide might cause the erroneous tagging to the verb “control” in text. To eliminate such a false positive, the tag will be removed if the term was recognized as a verb with PoS tagger.

4.3.3 Drug Abbreviation Detection

In this task, we proposed a drug abbreviation detection algorithm, which is depicted in Figure 4.5. In pharmacokinetic studies, a drug abbreviation is usually presented in a parenthesis after its full name that is first written in an abstract. This algorithm explores the existence of parentheses after the tagged drug names in a range of five words. However, in this way, we observed that not every term within parentheses are drug abbreviations. They might be an enzyme name (e.g. CYP3A4) that catalyzes its substrate, drug dosage or drug serum concentration measured in a clinical trial or PK experiment (e.g. 10 microM), PK parameters measured in a PK experiment (e.g. IC50), and sometimes statistical data analysis results (i.e. p-value or Conference Interval). These terms are removed if they are recognized as enzyme name, dosage information, PK parameters, or statistical result in the predefined list using regular expression.

Step 1: Find a drug name in a sentence.

Step 2: Check if a parenthesis is captured in the range of five words after a tagged drug name.

Step 3: Extract the whole term within the parentheses.

Step 4: Determine if the term is abbreviation or not. We exclude enzyme (e.g. CYP3A4), experimental parameters (e.g. 10 microM, r = 0.900, P <.001, or 30%), or company name (e.g. Zeneca, Ltd.)

Figure 4.5 The Procedure of Drug Abbreviation Detection

4.3.4 Window Construction and Adjustment

The procedure of window creation and adjustment is showed in Figure 4.6. First, based on the results of dictionary-based tagging, a window of a span-size of 5 is placed centering on the tagged drug name. Second, the window is further trimmed according to the following rules: if the window meets the end of a sentence; if the window overlaps with another drug name; or if the window meets the entity ending with a comma.

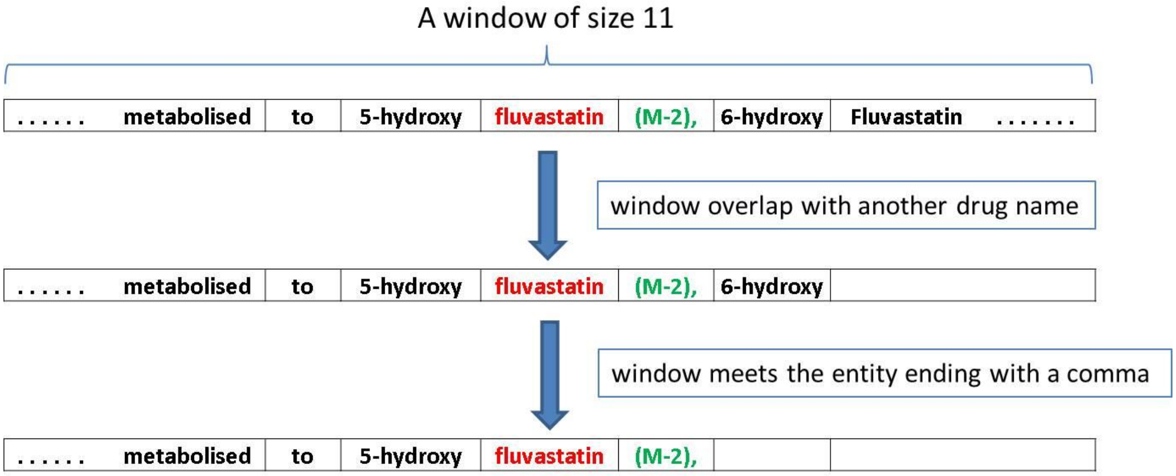
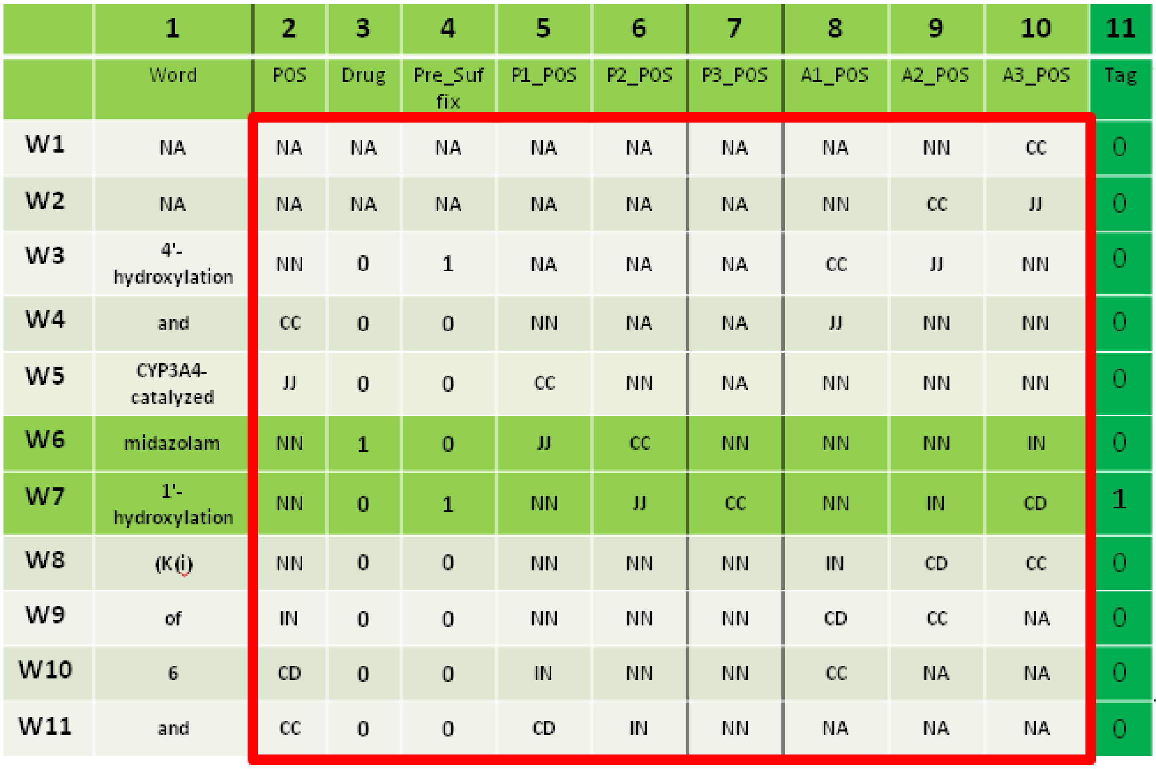

A window of size 11
...... metabolised | to | 5-hydroxy | fluvastatin | (M-2), | 6-hydroxy | Fluvastatin ....
window overlap with another drug name
...... metabolised | to | 5-hydroxy | fluvastatin | (M-2), | 6-hydroxy
window meets the entity ending with a comma
...... metabolised | to | 5-hydroxy | fluvastatin | (M-2),

Figure 4.6 Window Creation and Adjustment

4.3.5 Feature Matrix for Machine Learning

Three types of input features (PoS tags, drug index, and reaction index) are used to create feature matrix (input data) for machine learning algorithms. As shown in Figure 4.7, column 2 provides Part-of-Speech tags for each entity. In column 3, Drug index for midazolam (W6) located in the center of the searching window is indexed as 1. In column 4, Pre_Suffix index for both 4'-hydroxylation (W3) and 1'-hydroxylation (W7) are indexed as 1 because they contain a reaction term (hydroxyl). In addition, Column 5 to 10 provides the PoS information for its surrounding entities (± 3 entities) to represent sentence structure around the target entity. If this example is in the training dataset, Tag (column 11) for W7 is assigned as one. On the other hand, if it is in the testing dataset, all elements in Tag column are zero.

Once a feature matrix for entities in the window are obtained, machine learning algorithms are applied to predict whether the entities containing prefix or suffix terms in the window are the reaction elements for a drug metabolite name or not.



	1	2	3	4	5	6	7	8	9	10	11
	Word	POS	Drug	Pre_Suf fix	P1_POS	P2_POS	P3_POS	A1_POS	A2_POS	A3_POS	Tag
W1	NA	NA	NA	NA	NA	NA	NA	NA	NN	CC	0
W2	NA	NA	NA	NA	NA	NA	NA	NN	CC	JJ	0
W3	4'-hydroxylation	NN	0	1	NA	NA	NA	CC	JJ	NN	0
W4	and	CC	0	0	NN	NA	NA	JJ	NN	NN	0
W5	CYP3A4-catalyzed	JJ	0	0	CC	NN	NA	NN	NN	NN	0
W6	midazolam	NN	1	0	JJ	CC	NN	NN	NN	IN	0
W7	1'-hydroxylation	NN	0	1	NN	JJ	CC	NN	IN	CD	1
W8	(K(i)	NN	0	0	NN	NN	NN	IN	CD	CC	0
W9	of	IN	0	0	NN	NN	NN	CD	CC	NA	0
W10	6	CD	0	0	IN	NN	NN	CC	NA	NA	0
W11	and	CC	0	0	CD	IN	NN	NA	NA	NA	0

Figure 4.7 Feature Matrix of Entities in A Searching Window

4.4 Results

4.4.1 Performance of Corpus Construction

To measure the quality of corpus construction, the comparison between the result from each annotator and gold-standard data was measured pairwise using precision, recall, and F-measure. The result of analysis is shown in Table 4.1. The whole corpus evaluation suggested that expert comparable performance on this corpus and found that some of disagreements are introduced due to some extra information being communicated during the development of annotation guideline. For example, if an abbreviation was mentioned right behind its drug metabolite name, one annotator annotated both drug metabolite and its abbreviation as a tag. Another one only annotated drug metabolite but ignored the abbreviation part. As shown in the categorization of drug metabolite, the abbreviation of drug metabolite is fall in the category of Drug Metabolite Type II. This phenomenon is clearly shown in Table 4.2 and most of disagreements for Annotator 1 and Annotator 2 occurred in this category. For instances related to the issue of abbreviation, in PMID: 10859153, both annotator 1 and annotator 2 omitted the annotation of NORCIS, which is the abbreviation of norcisapride and is the metabolite of cisapride. In addition, many drug metabolites that are written with the mixture form of drug abbreviation and a reaction term were missed. An example of 3-hydroxyNVP (the metabolite of Nevirapine) can be found in PMID: 10570031. For the instance related to unique drug metabolite names, dihydroqinqhaosu in PMID: 10456689 is an active metabolite of artelinic acid. From this example, it is hard to find clues from its parent drug for recognizing its drug metabolite.

	Annotator 1	Annotator 2	Annotator 3
Precision	0.989	0.994	0.986
Recall	0.913	0.9	0.97
F-measure	0.950	0.945	0.978

Table 4.1 Comparison between Golden-Standard Corpus and The Result of Each Annotator

		Annotator 1	Annotator 2	Annotator 3
FN or missing annotation	Drug Metabolite Type I	21	11	3
	Drug Metabolite Type II	70	67	15
	Drug reaction Type I	18	3	13
	Drug reaction Type II	5	3	9
FP annotation		13	7	18

Table 4.2 Error Analysis for The Result of Each Annotator (FN)

	Pairwise Percent Agreement
Annotator 1-2	87.6%
Annotator 1-3	88.8%
Annotator 2-3	89.8%

Table 4.3 Inter-Annotator Agreement Result

In addition, the measurement of inter-annotator agreement between two annotators was quantified using pairwise percent agreement in Table 4.3. The pairwise percent agreement suggests that high levels of agreement (89.8%) are achieved.

4.4.2 Performance of Entities Tagging

To create the features of drug index or pre/suffix index for entity, the entities containing drug names or pre/suffix terms are identified using our proposed dictionaries. The statistics for drug metabolite corpus is shown in Table 4.4. There are 3789 drug entities in our corpus. Three drug entities (2 unique drugs) were not identified because of their drug names were not availability in our dictionary. These two drug names are RPR-106541 and cholantene. There are 1582 entities containing the reaction terms. Only 7 were erroneously tagged because some drug name or their synonyms comprise of our proposed pre/suffix terms. In tagging drug abbreviation, 452 terms are identified within the parentheses after the tagged drug names within a range of five words. Among these 452 terms, 138 are true drug abbreviations. Among the remaining 314 false positive terms, 161 enzyme names and 139 PK parameters and 5 other pharmacology terms were systemically removed. This process results in only 9 false positive identifications. They are all abbreviation of their reaction terms, such as 5-hydroxythalidomide (5-OH). Overall, the performance of identifying drug entities and pre/suffix terms is very good in this corpus. It minimizes the effect of error propagation due to the erroneous entity tagging.

	Frequency		
	Training	Testing	Total
Abstract	168	42	210
Drug	2966	823	3789
Term with Pre/suffix	1287	295	1582
Drug metabolite	1008	289	1297

Table 4.4 Statistics of Named Entity Recognition

4.4.3 Performance with Different Window

A window size=2*n+1 (n is one-sided word span) is placed centering on an entity containing a drug name. The drug metabolite identification depends on the window size. In order to optimize the drug metabolite identification performance, different window size is evaluated using our golden-standard corpus. Here, we have investigated the span size n = 2, 3, 4, 5, and 6. As shown in Figure 4.8, the best performance (optimal recall rate ~ 100%) was obtained using the window of size 11 (span = 5) via the analysis of pre/suffix distribution surrounding with drug name.

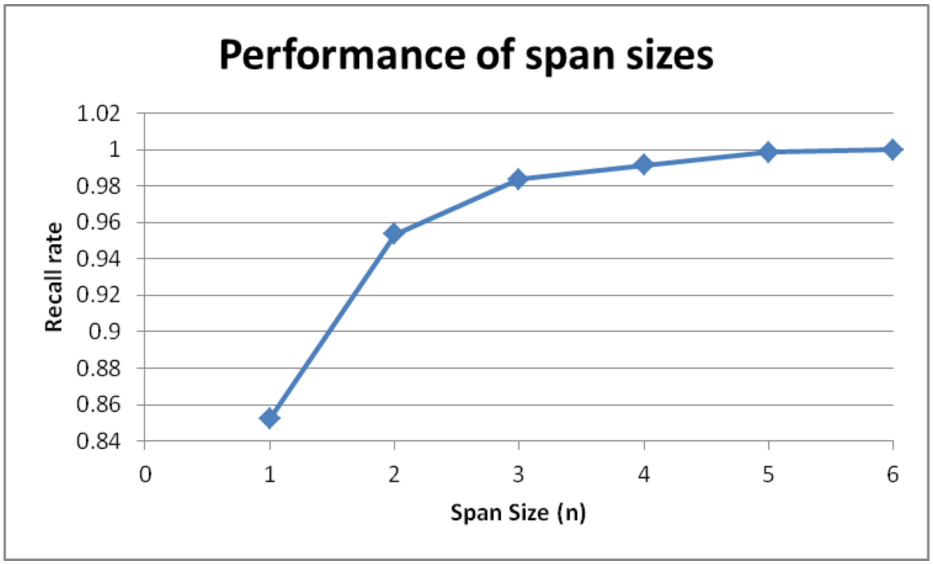


Figure 4.8 Performance of Different Span Sizes for A Searching Window

4.4.4 Performance of Drug Metabolite NER

In this task, from 210 abstracts, 168 abstracts are selected as the training dataset, and the remaining 42 abstracts are used as a testing dataset. With 10-fold cross validation on manually annotated corpora of training data, three different machine learning (ML) algorithms in Weka pipeline (M. Hall et al., 2009), LIBSVM (a Library of Support Vector Machines) (Chang & Lin, 2011), J48 (C4.5 algorithm) (Quinlan, 1993), and LMT (Logistic Model Tree) (Landwehr, Hall, & Frank, 2005), were implemented to predict the drug metabolites in the testing dataset.

Table 4.5 shows the overall evaluation results for our proposed NER system by using three different ML algorithms. The best precision (0.8884), recall (0.7716) and F-measure (0.8259) are achieved by SVM algorithm. While the second is the J48 tree, which has comparable precision (0.8818) but much lower recall (0.6194) and F-measure (0.7276).

	J48 Tree	LMT Tree	SVM
Precision	0.8818	0.8461	0.8884
Recall	0.6194	0.5709	0.7716
F-measure	0.7276	0.6818	0.8259

Table 4.5 Evaluation of Drug Metabolite NER System with The Test Corpus

4.5 Discussion

4.5.1 Error Analysis

For error analysis, a manual check was performed to know the major reasons to cause errors and further investigate how the performance of the proposed NER system is in identifying four different types of drug metabolite representations.

From the analysis, we perceive five reasons that incur errors during the annotation task, including unidentified drug abbreviation, incorrect decisions from machine learning, metabolite-like names, unique drug metabolite names, and drug names are not in dictionary. Via the analysis using the prediction result of SVM algorithm, Table 4.6 shows that a barrier to identify abbreviations of drugs and drug metabolites accounts for about 44 percent of error annotations. Such an omission because of abbreviation is still a challenge to all kinds of NER system. Incorrect decision by machine learning algorithms is the second most reason for error annotations (31.87%), which leads to both FP and FN errors. Those FP errors occurred when their POS patterns are similar to that of true drug metabolite. For example, "hydroxylation in vitro by nelfinavir" in PMID: 11159797 has a similar POS pattern (NN_reaction + IN_by + NN_drug) to that of drug reaction type II (NN_reaction + IN_of + NN_drug). On the other hand, FN errors occurred when the complicated and long phrase is used to represent drug reaction type II. The third reason is metabolite-like names, which accounts for 9.89% of error detection. For instance, "dihydroergotamine" is recognized as the metabolite of a drug name ("ergotamine"). But it is actually a generic drug name in Drugbank. The forth reason is the unique drug metabolite name, which accounts for 7.69% of errors. For example, UK-103 320 in PMID: 11298070 (the main metabolite of sildenafil) and cycloguanil in PMID: 9923577 (the metabolite of proguanil) are not identified because their denominations are not based on their parents drug and their names do not exist in our dictionary. Finally, few errors are from the unidentified drug name in our dictionary. This challenge might be still an issue if there is a comprehensive dictionary. Since the error analysis was manually analyzed, we also realized that most of reasons lead to false negative detection in our system. All those issues majorly lower the value of recall for all three ML algorithms. From Table 4.5, all three algorithms have decent performance in precisions but not their recall rates. This result can fully reflect this phenomenon.

Overall, we realized the superior performance using SVM algorithm from Table 4.5 and its main reasons of error detection from Table 4.6. In this analysis, we further inspected how the performance is in four different representations of drug metabolites. Table 4.7 shows the recall and precision rates for the proposed system using SVM algorithm. In this result, SVM can handle best in drug reaction type I, which capture 93.75% of positive instances with 96.33% precision rate. Compared to drug reaction type I, a comparatively lower performance in drug metabolite type I (R:92.65%/P:91.3%) and drug reaction type II (R:77.27%/P:89.47%) are acquired. However, for drug metabolite type II, the system only can reach a poor recall rate of 32.3%.

Reasons	Major Error types	Error percentage (%)
Unidentified abbreviation	FN	43.96%
Incorrect decision from ML	Both FP and FN	31.87%
Metabolite-like names	FP	9.89%
Unique Drug metabolite name	FN	7.69%
Drug name is not in dictionary	FN	6.59%
Total		100%

Table 4.6 Error Analysis of Drug Metabolite Annotation for SVM Algorithm

Recall /Precision	SVM
Drug Metabolite Type I	92.65%/91.3%
Drug Metabolite Type II	32.3%/87.5%
Drug Reaction Type I	93.75%/96.33%
Drug Reaction Type II	77.27%/89.47 %

Table 4.7 The Recall and Precision Rate for Four Different Types of Drug Metabolite Representations

From our observation, we ascribe the success of annotating drug reaction type I to its simpler structure. When both drug entity and reaction entity in this category are assigned a grammatical category of noun (NN) and laid side by side (NN_drug + NN_reaction), over 90% of instances were correctly identified. For those false negative cases in this category, many are referred to the reason of unidentified drug names. For instance, RPR 106541 sulfoxidation in PMID: 10411567 is one of false negative cases. For drug reaction type II, its performance is lower than that of drug reaction Type I. Except for the issue of unidentified drug names or abbreviations, complicated and long phrase lead to "incorrect decision from ML algorithm". For drug metabolite type I, most of error detection occurred due to the wrong detection of metabolite-like names.

To face the unfavorable result of drug metabolite type II, we observe that false negatives incident to unidentified abbreviations of drug names or unique drug metabolite names account for most erroneous detections. This issue is still the most challenging task in identifying drug metabolite because there is no cue in perspective of naming convention and no existing dictionary that contains comprehensive terminologies. It can be probably solved by manual curation or identification using sophisticated natural language process from context.

4.5.2 Performance Drug Metabolite NER without Drug Metabolite Type II

Via the error analysis, we realized that even though those true instances of Drug metabolite Type II are included in both training and testing dataset, it seems not to help for recognition and may deteriorate the prediction capability of the training model. Therefore, an exercise was implemented to investigate the model without the inclusion of Drug Metabolite Type II instances. Table 4.8 shows that this model can improve the recall for Drug Reaction Type I but impair that of Drug Metabolite Type I. Interestingly, the precision rates for all three representations are all boosted.

Recall /Precision	SVM
Drug Metabolite Type I	89.71%/98.39%
Drug Reaction Type I	95.54%/98.17%
Drug Reaction Type II	77.27%/97.14%

Table 4.8 The Recall and Precision Rates of The Model without Drug Metabolite Type II

4.5.3 Performance Comparisons with WhatizitChemical

In this study, we try to explore existing methodologies for comparing performance. Unfortunately, there is no existing NER system developed for the same purpose as ours. (Nobata et al., 2011) is one of limited NERs designed to extract one type of metabolites, yeast metabolites, but not drug metabolite. In this task, the performance comparison to those available through Whatizit pipelines was implemented. From the result, it showed that whatizitChemical (Rebholz-Schuhmann et al., 2008), which is a literature search tool for chemical metabolites, achieved higher recall but lower precision and F-measure than their system.

To our understanding, whatizitChemical is designed to identify chemical entities, drugs and protein names for EBIMed individually, but it was not fully designed to annotate the full term of a drug metabolite. Using an example in PMID: 10460803, "dextromethorphan o-demethylation" is a CYP2D6 reaction for dextromethorphan. Using whatizitChemical, dextromethorphan and o-demethylation were tagged as a drug and a chemical, respectively. Thus, it is difficult to compare the performance of two systems that have different annotation criteria. Here, we assume that whatizitChemical recognizes and pairs both drug name and its reaction term. Whenever it correctly annotates both drug term and its reaction term from the annotation in gold-standard corpus, we call it as a true positive annotation. Otherwise, it is a false negative. To make a fair comparison, we only count the number of TPs and FNs of annotated terms on the manually curated gold-standard corpus and calculate their recall. Table 4.9 shows that, our NER recall is 0.77, while whatizitChemical has a recall of 0.65. Our NER outperforms whatizitChemical by 12%.

	Recall
Our NER system with SVM	0.77
whatizitChemical	0.65

Table 4.9 Performance Comparison with WhatizitChemical

4.6 Conclusion

The characteristics of drug metabolite are recognized to be an important feature to investigate drug-drug interactions, adverse effects of chemical compounds and their associations to toxicological endpoints or the extraction of pathway and metabolic reaction relations. However, there is no existing dictionary containing comprehensive drug metabolite terminologies but also no named entity recognition (NER) system focusing on the identification for drug metabolite and reaction. Here, we developed a novel NER system to annotate drug metabolites and reactions in scientific text, utilizing an integrated dictionary and machine learning algorithms (including SVM, J48, and LML). This system utilizes the information of Part-of-Speech, drug index and pre/suffix feature to determine whether the entities containing reaction terms belong to the drug metabolite or not. To evaluate performance, a golden-standard corpus is created by three annotators. With 10-fold cross validation on the corpora, SVM outperformance J48 and LMT with the precision (0.8884), recall (0.7710), F-measure (0.8259). In this work, we compared our performance with an existing NER system, whatizitChemical, which is designed for recognizing small molecules or chemical entities. Our system with SVM algorithm outperforms whatizitChemical by 12%.

Chapter 5. Translational Drug Interaction Evidence Gap Discovery

5.1 Background

Drug-Drug Interaction (DDI) is one of the major causes of adverse drug reaction (ADR) and has been demonstrated to threat public health (M. J. Hall et al., 2010; Niska, Bhuiya, & Xu, 2010). It causes an estimated 195,000 hospitalizations and 74,000 emergency room visits each year in the USA alone. With increasing rates of poly-pharmacy, the incidence of DDIs is most likely to increase such that drug interaction research remains essential (Hajjar ER, 2007). Current DDI research aims to investigate different scopes of drug interactions: molecular level of pharmacokinetics interaction (PG), pharmacokinetics interaction (PK), and clinical pharmacodynamics consequences (PD) (R. Boyce et al., 2009a, 2009b; Hennessy & Flockhart, 2012). All types of experiments are important, but they are playing different roles for DDI research. For instance, In vitro PK studies investigate molecular interactions within tissue cells and uncover protein activity and genetic underpinnings of distinct molecular responses to drugs making them highly relevant to current pharmacogenetics research (Crews et al., 2012; Wilke et al., 2012). Follow-up of positive in vitro finding helps develop an in vivo assessment while its negative findings alleviate the need for further in vivo studies. Further studies via In vivo PK experiments are used to evaluate whether the molecular interactions impact drug exposure in human body. Once potential DDIs are identified based on in vitro and/or in vivo studies, researchers can further design studies or collect data for determining whether the effects of the experimental drug on a range of substrates. Finally, clinical PD studies test whether drug interactions can change the actual response to drugs, including drug efficacy and the induction of ADR (Prueksaritanont et al., 2013).

As diverse disciplines and varied studies are involved, interaction evidence is often not available cross all three types of evidence, which create knowledge gaps and these gaps hinder both DDI and pharmacogenetics research. For example, clinical evidence alone does not convey sufficient information about underlying molecular or pharmacokinetics mechanism. On the other hand, in vitro experiments alone cannot determine how a given drug interaction influence drug efficacy or lead to ADR. Therefore, simultaneously considering all three types of DDI evidence had been suggested as an effective way to reduce false DDI predictions and develop safer treatment for clinical usage (R. Boyce et al., 2009a, 2009b).

Despite the existence of several services or databases, either the methods used to obtain the knowledge, its reliability, extent or coverage is not always available. First Databank provides a drug-drug interaction module that classifies DDIs in terms of severity levels and identifies the type of evidence for each drug interaction (e.g. human clinical trial or animal studies) ("First Databank," 2014)). Drugbank database provides a list of interacting candidates for a queried drug and describes the basis of interaction with a short and simple sentence, but it is not enough for distinguishing evidence types (Law et al., 2014)). Thus, lacking of knowledge cross all three types of evidences introduces the gap for translational drug interaction study. To investigate different types of experimental evidences, more recent research aims to identify DDI evidence from biomedical literature by using in silico technologies (Computer-based technologies such as text mining) (B. Percha, Garten, Y., and Altman, R.B., 2012; I. Segura-Bedmar, Crespo, M., de Pablo-Sanchez C., and Martinez, P., 2011; I. Segura-Bedmar, P. Martinez, et al., 2011a; Segura-Bedmar, Martinez, & de Pablo-Sanchez, 2011c; L. Tari, Anwar, S., Liang, S., Cai, J., Baral, C., 2010; H. Y. Wu, Karnik, et al., 2013). However, most of those studies focus on single type of evidence and no one addressed knowledge gaps to distinguish the different types of experimental evidence, which impedes the translation of information about molecular mechanism into clinical understanding. Thus, we would like to close such gaps in DDI evidence by using informatics methods to integrate and tap into our collective scientific knowledge.

5.2 Material

5.2.1 Lexica Construction

A lexica comprising of drug name, enzyme/transporter, and action terms were collected from the result of AIM 1 (in Chapter 2). For drug name, the proposed drug name dictionary is an integrated drug name database with the connection of multiple oriented drug resources. Fundamentally, this database is built based on the drug names in Drugbank (Knox et al., 2011), which includes three types of drug names, generic (8194), brand (17337) and synonym (7383). In addition, the unique drug names (74619) from MeSH are supplied to expand the coverage. Within those from MeSH, 14946 terms are connected to the generic names from DrugBank. In total, there are 70712 drug names in the dictionary. In this exercise, we focused only on FDA approved and withdrawn drugs, which left 2403 unique drug generic names for the mining purpose. For enzyme's terms, 94 generic names and their synonyms (350 terms in total) are collected from Gene ontology ("Gene Ontology Consortium: going forward," 2015), HUGO Gene Nomenclature Committee (HGNC) ("HUGO Gene Nomenclature Committee at the European Bioinformatics Institute,"), and The Human Cytochrome P450 (CYP) Allele Nomenclature Database (Sim & Ingelman-Sundberg, 2010). For transporter's terms, 624 generic names and their synonyms (1993 terms in total) are collected from Gene ontology ("Gene Ontology Consortium: going forward," 2015), HGNC ("HUGO Gene Nomenclature Committee at the European Bioinformatics Institute,"), and Transporter Classification

Database (Saier et al., 2016). The action terms are collected from our previous work, PK ontology (H. Y. Wu, Karnik, et al., 2013).

5.2.2 DDI Corpus Construction

Two different corpora were prepared (the result of AIM 1). The first corpus is constructed to be the golden standard (GS) corpus, which is used in IR and IE exercise. DDI information in entity and sentence level and entity-relationship are indicated and annotated with the types of evidence. This corpus comprises of in vitro PK DDI abstracts (n = 210), clinical PK DDI abstracts (n = 218), and clinical PD abstracts (n=140). The second corpus is prepared to be the training and testing data for building the model of Information Retrieval (IR) task. In IR corpus, there are 300 DDI relevant abstracts for each evidence type as positive data, 800 DDI irrelevant abstracts (200 non-DDI abstracts filtered from the phase of PubMed search, 300 drug-related abstracts such as single drug or drug-nutrition, and 300 drug-irrelevant abstracts) and 10,000 randomly selected abstracts from Medline database as negative data.

Data collection: For building both IR and GS corpora, candidate articles were obtained from a PubMed query first. However, many false results may be retrieved using the search via the MeSH (Medical Subject Headings) controlled vocabulary used to index Medline articles. To improve the quality, a strategy of manual screening was proposed to filter out false positives and to determine whether an article satisfies the inclusion-exclusion criteria (IECs). In this task, a validation process was executed via manually reading each article by more than two curators. If each article is not agreed by all curators, supervisors will involve and make the final decision.

Text Annotation: The annotation guideline was proposed to implement the annotation process. This guideline provides the standards or rules of annotating entities, sentences, and entities relationships. In GS corpora, a hierarchical three-level annotation scheme (H. Y. Wu, Karnik, et al., 2013) for gold-standard corpus was implemented to annotate three layers of DDI information: key terms, DDI sentences, and DDI pairs. Such a golden standard corpus can provide the semantic information for specific entities, offer a rule or criteria to determine DDI in different types of studies, and be the standard for evaluating the text mining result.

Within each abstract, our annotators will manually tag each sentence with a label indicating whether it has evidence for interaction or for no-interaction. Each sentence will also be annotated with the specific relevant components namely: drug names, DDI relations, interaction-denoting verbs and action words (such as inhibits, metabolizes etc.), ADR, experimental parameters and their associated values. Annotations will be done using XML format similar to those used in the GENIA corpus (J. D. Kim, Ohta, T., Tateisi, Y., and Tsujii, J., 2003).

Annotation Evaluation: In this task, three senior PhD students were recruited as annotators for corpora construction. To guarantee the quality of the annotations, two supervised annotators who are professionals in pharmacology and biomedicine will ensure that all annotators possess the background necessary for performing the task. If needed, they will further train the annotators on pharmacology and epidemiology of the DDI research. The three annotators will go over the retrieved abstracts. Since discrepancies arise, they will be resolved through supervised annotators. Krippendorff's alpha (K. Krippendorff, 2004) will be used to assess agreement among the annotators.

For the detail of data collection, annotation procedure and annotation evaluation, please see all the content in Chapter 2 and Chapter 3.

5.3 Method

5.3.1 Overview of Evidence-based Text Mining Tools for Drug-Drug Interaction

In this task, we aim to develop a suite of text mining tools to explicitly identify each type of DDI evidence, namely in vitro PK, clinical PK and clinical PD. The workflow is shown in Figure 5.1. Considering the importance of the different types of DDI evidence, we developed Document-level classifiers – to distinguish PubMed abstracts likely to contain evidences, DDI-level classifiers – to identify interacting drug within those retrieved PubMed abstracts. Given the separate corpora, we were able to train distinct, highly-focused classifiers for each type of evidence, which our preliminary studies (Kolchinsky, Lourenco, Li, & Rocha, 2013) had demonstrated that it can lead to high performance. Notably, unique drug pairs from distinct abstracts can be extracted using proposed DDI-level classifier without using syntactic analysis, which improves the efficiency of data extraction. Making use of such evidence, drug pairs from three types of studies can integrate the collective scientific knowledge of DDIs, identify gaps therein, and inform future drug studies thus forming a basis for improved clinical support.

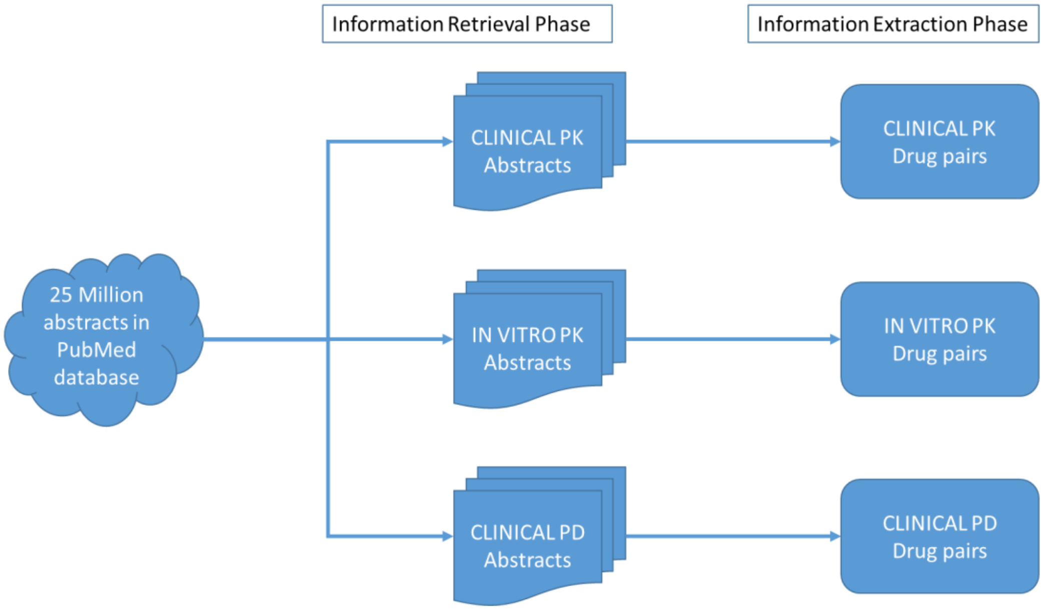


Figure 5.1 Workflow of Large-Scale Mining for Drug Interaction Evidence

As shown in Figure 5.1, the text mining task for each type of study was accomplished with two steps: Information Retrieval phase and Information Extraction phase. For the experiment setting, the detail is shown in Figure 5.2.

In IR phase, IR corpora were utilized to construct the optimal model, which can maximize recall rate, for abstract categorization. The IR model was first built based on 150 DDI abstracts for each type of studies (positive) and 10,000 randomly selected abstracts (negative). It was tuned for reaching the optimal recall on the testing dataset (another 150 DDI abstracts for each study and 500 single-drug or nutrition-related abstracts plus 300 random articles. With the optimal settings learnt from IR exercise, the large-scale screening task for 25 million abstracts were implemented using the full IR corpora (300 DDI abstracts for each study type as positive data and 10,000 random abstracts as negative data). In addition, the retrieved abstracts were validated using GS corpora to evaluate the recall rate in the large-scale screening task.

In IE phase, GS corpora (210, 218, and 140 abstracts for in vitro PK, clinical PK, and clinical PD abstracts) are applied to build the optimal model, which can maximize F-measure, for relation-level classification. Similar to IR exercise, 60% of true entity relation pairs were used to build the optimal model and test on the rest of relation pairs. With the optimal settings, the IE task for extracting DDI pairs from the result of IR phase were constituted using full GS corpora.

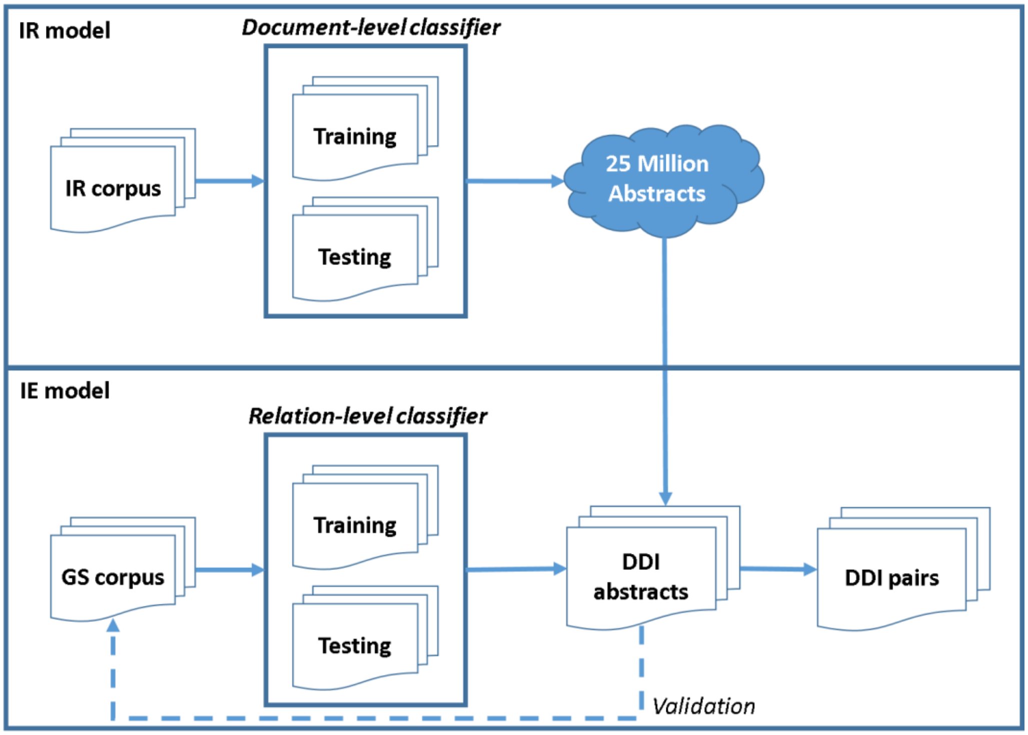
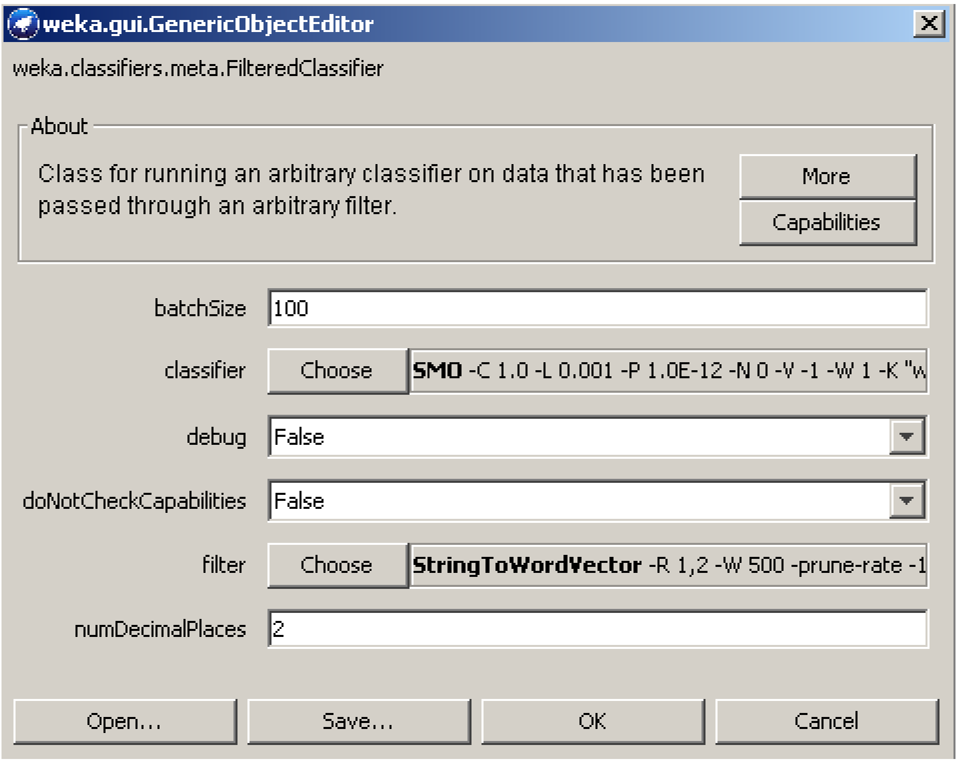


Figure 5.2 Text Mining Experiment Setting

5.3.2 Information Retrieval (IR) Exercise

To identify abstracts containing DDI evidence of three study types from whole Medline database, three distinct corpora (in vitro PK, clinical PK, and clinical PD) collected from the result of AIM 1 were utilized to build the training model for abstract categorization. To accomplish this task, classification architectures in IR exercise were examined and trained using finely curated training corpus (150 DDI-relevant abstracts for each study type and 10,000 DDI-irrelevant abstracts) and testing corpus (150 DDI-relevant abstracts for each study type and 800 DDI-irrelevant abstracts including single drug study, drug-nutrition study, PD related and randomly selected abstracts).

For experimental setting, this exercise was implemented with Support Vector Machine (SVM) in Weka pipeline (Witten, Frank, Hall, & Pal, 2016) (Figure 5.3). In Figure 5.4, string attributes in each abstract are converted into a set of attributes representing word occurrence information from the text contained in the strings using "StringToWordVector" module. IteratedLovinsStemmer, stopwordsHandler, NGramTokenizer (1-3), lowerCaseTokens and wordsToKeep (1000) are used to create word features while IDFTransfrom, TFTransform, and outputWordCounts are placed to create statistical feature for text classification. With those features, LibSVM in Weka is implemented for text classification subject to the optimization of recall rate.



weka.gui.GenericObjectEditor

weka.classifiers.meta.FilteredClassifier

About

Class for running an arbitrary classifier on data that has been passed through an arbitrary filter.

batchSize 100

classifier **SMO** -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "w

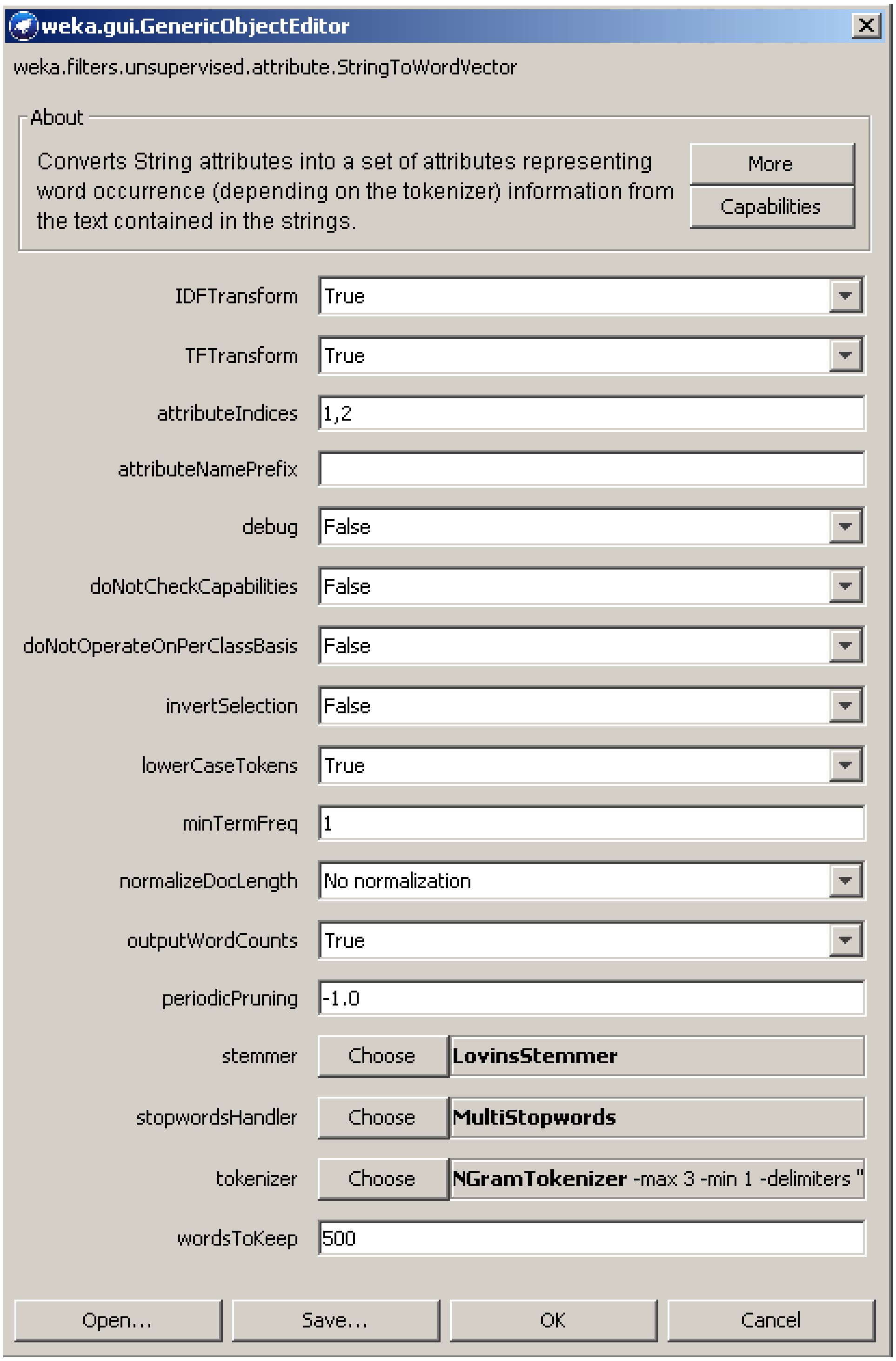
debug False

doNotCheckCapabilities False

filter **StringToWordVector** -R 1,2 -W 500 -prune-rate -1

numDecimalPlaces 2

Figure 5.3 Classifier Selection and Bag-of-Word Creation



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.StringToWordVector

About

Converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings.

IDFTransform: True
TFTransform: True
attributeIndices: 1,2
attributeNamePrefix:
debug: False
doNotCheckCapabilities: False
doNotOperateOnPerClassBasis: False
invertSelection: False
lowerCaseTokens: True
minTermFreq: 1
normalizeDocLength: No normalization
outputWordCounts: True
periodicPruning: -1.0
stemmer: **LovinsStemmer**
stopwordsHandler: **MultiStopwords**
tokenizer: **NGramTokenizer** -max 3 -min 1 -delimiters "
wordsToKeep: 500

Figure 5.4 The Creation of String Attributes using StringToWordVector

5.3.3 Information Extraction (IE) Exercise

After identifying abstracts that are likely to contain DDI evidence, classified by evidence type, the next step is to identify actual DDI-related information, including drug names, potential interacting drug pairs and interaction indicators, within these abstracts. It was achieved with two steps:

Entity recognition and normalization: We shall tag relevant entities, including drug names and interaction verb, using name-entity recognition method (NER): string-matching against the developed lexica. Extracted drug names are normalized by mapping them to "DrugBank identifiers" or MeSH identifiers and interaction verbs are normalized to stemmed forms. In this IE exercise, we focus on FDA approved drug (2202) and withdraw drugs (201).

Drug pair extraction: Relation extraction process is not a trivial task. Most of research works recognize a piece of text having a semantic property of interest and extract semantic relations between entities using natural language process technology (B. Percha, Garten, Y., and Altman, R.B., 2012; I. Segura-Bedmar, P. Martinez, et al., 2011a, 2011c; Segura-Bedmar, Martinez, & Herrero-Zazo, 2014; Segura-Bedmar, Martınez, & Sánchez-Cisneros, 2011b; L. Tari, Anwar, S., Liang, S., Cai, J., Baral, C., 2010). However, biomedical concept relationships should be considered as properties of biomedical entities. It is insufficient to define a relation between entities using individual sentences (B. Percha & Altman, 2015). For instance, an experimental finding in one sentence found that "Drug A prolonged the half-life of Drug B". It is not sufficient to constitute an inhibitory relationship if there exists another statement of “Drug A did not significantly change Drug B’s AUC” in a different sentence. Therefore, in this work, we implement a machine learning approach to inferring DDI pairs from whole context in an abstract. It allows determining interacting drug pairs based on multiple concept relationships from different sentences at once.

<u>Feature creation</u>: We assume that drug i ($\mathrm{D_i}$) and an action term ($\mathrm{V_k}$) are mentioned starting at $\mathcal{DL}_{\mathrm{i},\mathcal{S}_{\mathrm{D_i}}}\mathrm{th}$ or $\mathcal{VL}_{\mathrm{k},\mathcal{S}_{\mathrm{V_k}}}\mathrm{th}$ characters in a sentence $\mathcal{S}_{\mathrm{D_i}}$ or $\mathcal{S}_{\mathrm{V_k}}$and, where $\mathcal{S}_{\mathrm{D_i}} \in [1, 2, \dots, \mathrm{N_{max}}]$ is a variable set that represents the sentence numbers containing $\mathrm{D_i}$ and $\mathrm{N_{max}}$ is the number of sentence in each abstract. If both $\mathrm{D_i}$ and $\mathrm{V_k}$ co-occur in more than one of sentences in each abstract, the value of $\left|\mathcal{S}_{\mathrm{D_i}} \cap \mathcal{S}_{\mathrm{V_k}}\right|$ will greater than 1. To determine whether drug i and drug j are investigated for drug interaction in each abstract, the following 15 measurements are utilized to be the inputs of machine learning algorithms.

<u>Numerical measurements</u>:

1. Minimum sentence difference between $\mathrm{D_i}$ and $\mathrm{D_j}$ ($\mathrm{SD_{i,j}}$):

$$SD_{i,j} = min(|\mathcal{S}_{D_i} - \mathcal{S}_{D_j}|), \qquad (1)$$

2. Number of sentences containing $\mathrm{D_i}$ and $\mathrm{D_j}$ that can be divided by an action term ($\mathrm{NDV_{i,j}}$):

$$\mathrm{NDV_{i,j}} = \sum\nolimits_{\mathrm{k,S}} \mathrm{I}_{\mathrm{NDV}_{\mathrm{i,j,k},\mathcal{S}}} \qquad (2)$$

where $\mathrm{I}_{\mathrm{NDV}_{\mathrm{i,j,k},\mathcal{S}}} = \begin{cases} 1, & \text{if } \mathcal{DL}_{\mathrm{i},\mathcal{S}} < \mathcal{VL}_{\mathrm{k},\mathcal{S}} < \mathcal{DL}_{\mathrm{j},\mathcal{S}} \\ 0, & \text{otherwise} \end{cases}$ is an index function to determine if $\mathrm{D_i}$ and $\mathrm{D_j}$ can be separated with $\mathrm{V_k}$ in sentence $\mathcal{S}$ and $\mathcal{S} \in [\mathcal{S}_{\mathrm{D_i}} \cap \mathcal{S}_{\mathrm{D_j}} \cap \mathcal{S}_{\mathrm{V_k}}]$.

3. % of sentences containing $\mathrm{D_i}$ and $\mathrm{D_j}$ that can be divided by an action term ($\mathrm{ANDV_{i,j}}$)

$$\mathrm{ANDV_{i,j}} = \frac{\mathrm{NDV_{i,j}}}{|\mathcal{S}|} \qquad (3)$$

where $|\mathcal{S}|$ is the number of sentences containing $\mathrm{D_i}$, $\mathrm{D_j}$ and $\mathrm{V_k}$.

4. Average angle of each drug pair to interaction verb ($\measuredangle_{\mathrm{i,j}}$):

$$\measuredangle_{\mathrm{i,j}} = \frac{\sum_{\mathrm{k,s}} \angle_{\mathrm{i,j,k,s}}}{|\#\ \mathrm{of}\ \mathrm{V_k}\ \mathrm{in}\ \mathcal{S}\ |} \qquad (4)$$

where $\angle_{\mathrm{i,j,k,s}} = \cos^{-1}\left(\frac{\mathrm{dis}(\mathrm{D_i},\mathrm{V_k})^2 + \mathrm{dis}(\mathrm{D_j},\mathrm{V_k})^2 - \mathrm{dis}(\mathrm{D_i},\mathrm{D_j})^2}{2 \times \mathrm{dis}(\mathrm{D_i},\mathrm{V_k}) \times \mathrm{dis}(\mathrm{D_j},\mathrm{V_k})}\right)$, $\mathrm{dis}(\mathrm{D_i},\mathrm{V_k}) = \sqrt{|\mathcal{DL}_{\mathrm{i,s}} - \mathrm{V}\mathcal{L}_{\mathrm{k,s}}|}$, $\mathrm{dis}(\mathrm{D_j},\mathrm{V_k}) = \sqrt{|\mathcal{DL}_{\mathrm{j,s}} - \mathrm{V}\mathcal{L}_{\mathrm{k,s}}|}$, and $\mathrm{dis}(\mathrm{D_i},\mathrm{D_j}) = \sqrt{|\mathcal{DL}_{\mathrm{i,s}} - \mathcal{DL}_{\mathrm{j,s}}|}$

5. The frequency of $\mathrm{D_i}$ in an abstract ($\mathrm{F}_{\mathrm{D_i}}$)

6. The frequency of $\mathrm{D_j}$ in an abstract ($\mathrm{F}_{\mathrm{D_j}}$)

7. The frequency of $\mathrm{D_i}$-$\mathrm{D_j}$ mentioned in the same sentence ($\mathrm{F}_{\mathrm{D_i,D_j}}$)

8. The frequency of $\mathrm{D_i}$ tagged in the abstract/Total # of tagged drugs ($\overline{\mathrm{F}}_{\mathrm{D_i}}$)

9. The frequency of $\mathrm{D_j}$ tagged in the abstract/Total # of tagged drugs ($\overline{\mathrm{F}}_{\mathrm{D_j}}$)

10. The frequency of $\mathrm{D_i}$ - $\mathrm{D_j}$ mentioned in the same sentence/ Total # of drug pair combinations ($\overline{\mathrm{F}}_{\mathrm{D_i,D_j}}$)

Categorical measurement:

11. Whether both $\mathrm{D_i}$ and $\mathrm{D_j}$ mentioned in title sentence (YES/NO)

12. Whether both $\mathrm{D_i}$ mentioned in title sentence (YES/NO)

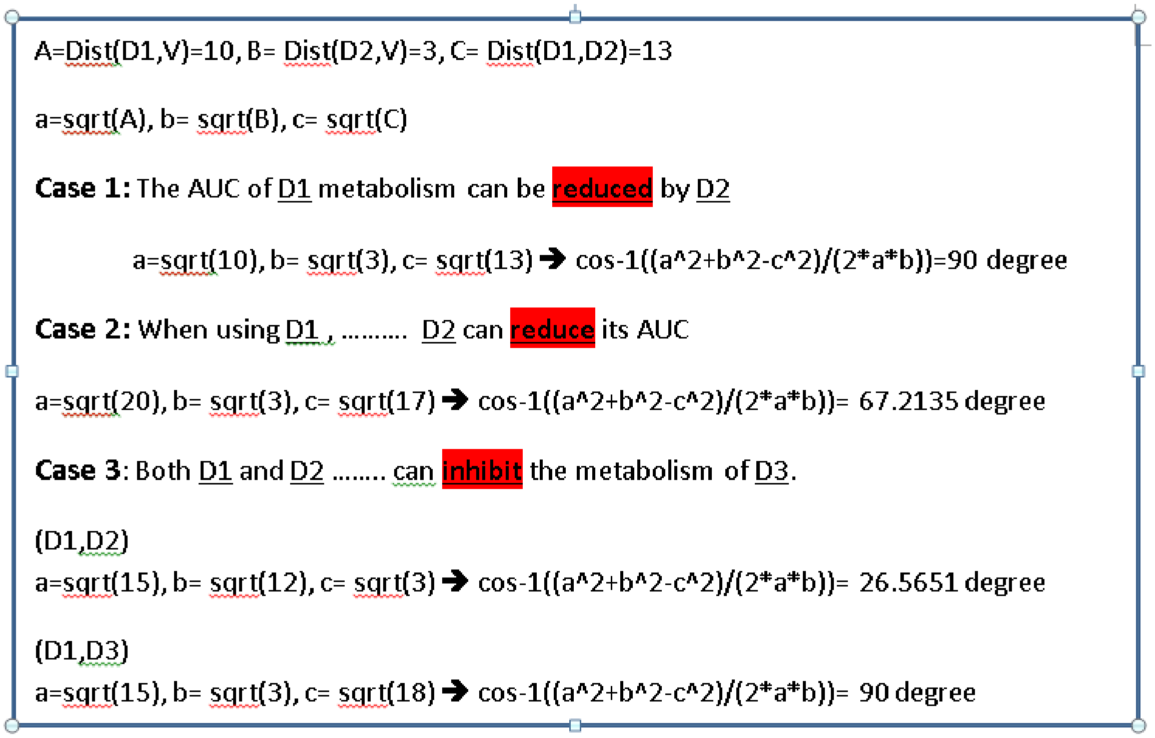
13. Whether both $\mathrm{D_j}$ mentioned in title sentence (YES/NO)

14. Have action verbs in one of the sentence with either drugs (YES/NO)

15. FDA probe information (5 Categories)

16. Whether both $\mathrm{D_i}$ and $\mathrm{D_j}$ have shared ATC code at level 4 (YES/NO)

Different measurements can introduce different characteristics for helping distinguish whether an interaction of drug pairs is mentioned in an abstract or not. Since we do not utilize deep parsing techniques in this task, instead Feature 1-4 are used to represent the characteristics of syntactic structure for text. Feature 1 ($\mathrm{SD_{i,j}}$) is minimum sentence difference between $\mathrm{D_i}$ and $\mathrm{D_j}$, which provides a weight to adjust the possibility of having a relationship between two drugs. In the perspective of natural language, if the value of $\mathrm{SD_{i,j}}$ is high, which means two drugs locate apart from one another, we can assume that one of drugs are only mentioned once in the beginning or a part of abstracts, not an investigated drug in that experiment. On the other hand, if the value of $\mathrm{SD_{i,j}}$ is lower, which means two drugs locate closely in term of sentence level, it is likely to deem that they might involve in the same biomedical event. Feature 2 and Feature 3 are the number and percentage of sentences containing $\mathrm{D_i}$ and $\mathrm{D_j}$ that are divided by an action term ($\mathrm{NDV_{i,j}}$ and $\mathrm{ANDV_{i,j}}$), respectively. These two features provide a clue to their frequency and possibility of drug interactions mentioned in an abstract. For instance, when the values of both $\mathrm{NDV_{i,j}}$ and $\mathrm{ANDV_{i,j}}$ are great, this drug pair might have a higher chance of having an interaction. But when $\mathrm{ANDV_{i,j}}$ is low, this article is to compare their efficacy on the other drug. Feature 4, average angle of a drug pair to interaction verbs ($\measuredangle_{\mathrm{i,j}}$), is a virtual index to measure the relative location of a drug pair to action verbs. By taking square root of 2 in the absolute distance among three nodes in a sentence, it converts the relative location among $\mathrm{D_i}$ and $\mathrm{D_j}$ and $\mathrm{V_k}$ from one dimension into two dimensions and creates a triangle for three nodes. In equation (4), if the value of $\angle_{\mathrm{i,j,k,s}}$ is 90 degree, it means $\mathrm{D_i}$ and $\mathrm{D_j}$ can be separated by $\mathrm{V_k}$ in the sentence s and $\mathrm{dis}(\mathrm{D_i}, \mathrm{V_k}) + \mathrm{dis}(\mathrm{D_j}, \mathrm{V_k}) = \mathrm{dis}(\mathrm{D_i}, \mathrm{D_j})$. Besides, when the value of $\angle_{\mathrm{i,j,k,s}}$ is less than 90 degree, two drugs will be in the same side from an action verb. In terms of the value of $\angle_{\mathrm{i,j,k,s}}$, when it is close to 0 degree, two drugs locate closely but apart from an action verb. But when it is closer to 90 degree (e.g. 75 degree), two drugs are apart but one of the drug is more close to an action verb. As shown in Figure 5.5, unlike Feature 2 and 3, Feature 4 can further differentiate the relationship between two drugs even though they locate in the same side from an action verb. In addition, Feature 5-10 are the numerical features to offer the statistics information for drug pairs and Feature 11-14 offer the categorical features to determine the availability of drugs found in the title sentence for each abstracts. Differed from Feature 1-14 collected from the text of an abstract itself, Feature 15-16 are the features assigned based on a priori knowledge (FDA probe information and ATC code). For Feature 15, one of categories (5 categories in total) is assigned for each drug pair according to their drug probe information. If both drugs have same genes involved (enzyme or transporter) and act as different roles, e.g. one drug is a CYP3A4 inhibitor and the other is a CYP3A4 substrate, Category 1 will be assigned, which indicates these two drugs are likely to interact with each other because of the property of drug metabolism and inhibition. If both drugs have the same gene involved, and act as the same role, Category 2 is assigned, which indicates these two drugs are less likely to have interactions. For Category 3, only one drug has FDA probe information, but the other does not. For Category 4, both drugs have FDA probe information, but no common enzyme/transporter gene involved. For Category 5, FDA probe information for both drugs is unknown. Similar to Feature 15, Feature 16 is a binary category to define whether two drugs have shared ATC codes at level 4. If the value of Feature 16 is positive, it means that they belong to the same chemical/pharmacological/therapeutic subgroup and obtain similar pharmacology property and chemical structure, which hints are less likely to have interaction.



A=Dist(D1,V)=10, B= Dist(D2,V)=3, C= Dist(D1,D2)=13

a=sqrt(A), b= sqrt(B), c= sqrt(C)

Case 1: The AUC of D1 metabolism can be reduced by D2

a=sqrt(10), b= sqrt(3), c= sqrt(13) → cos-1((a^2+b^2-c^2)/(2*a*b))=90 degree

Case 2: When using D1 , ……… D2 can reduce its AUC

a=sqrt(20), b= sqrt(3), c= sqrt(17) → cos-1((a^2+b^2-c^2)/(2*a*b))= 67.2135 degree

Case 3: Both D1 and D2 …….. can inhibit the metabolism of D3.

(D1,D2)
a=sqrt(15), b= sqrt(12), c= sqrt(3) → cos-1((a^2+b^2-c^2)/(2*a*b))= 26.5651 degree

(D1,D3)
a=sqrt(15), b= sqrt(3), c= sqrt(18) → cos-1((a^2+b^2-c^2)/(2*a*b))= 90 degree

Figure 5.5 Examples of ∠_(i,j,k,s) Calculation

Features and Classifier selection: Drug interaction evidence from different types of studies might be described in different ways. In vitro PK studies characterize extensively its major drug-metabolizing human CYP isoforms; Clinical PK studies investigate how a drug behave in metabolism with LC-MS analysis of drug concentration and other pharmacokinetic matrices; Clinical PD studies majorly emphasize on the risk or adverse reaction of drug interaction in population level. Due to such diversity of interest, text for drug interaction evidence is written or organized differently. Therefore, applying all 15 features as inputs to a single machine learning algorithm to determine drug interactions for all three study types is not feasible. In this task, we examine features and classifiers in terms of the performance of information extraction in a particular study. In Table 5.1, the feature sets for G1, G2, and G5 are manually selected. G1 and G2 describe features differently in statistics information (Feature 5-10) and G5 utilizes all features. For G3 and G4, their features are selected in statistical manners. Using the measurement of Akaike information criterion (AIC), optimal features for distinct classifiers are determined using stepwise regression model. The only difference is that G3 does not consider 2-way interaction terms, but G4 does.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Interaction terms
G1	X		X	X	X	X	X				X				X	X	NA
G2	X		X	X				X	X	X	X				X	X	NA
G3		V		V				V	V			V			V	V	NA
	T	T		T	T			T	T			T			T	T	
	C	C			C			C	C				C	C			
G4		V		V				V	V	V						V	V: 15*9, 2*15, 10*8, 4*8, 2*8, 2*16, 9*12
	T	T															T: 15*2, 4*2, 9*10, 4*10, 9*15, 8*12, 2*12, 2*10, 4*12
	C			C	C			C	C				C	C			C: 9*13, 8*4, 13*2, 8*13
G5	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	NA
NOTE: X,V,T, and C symbols represent that the features are selected by all three studies, Clinical PK, In Vitro PK, and Clinical PD studies, respectively																	

Table 5.1 Feature selections for three study types

With 5 groups of features, 7 popular classifiers (J48, Naïve Bayes, SMO, Logistic Regression, Random Forest, LMT, and Iterative Classifier Optimizer in Weka pipeline) are used to examine the performance in terms of F-measure.

5.4 Results

5.4.1 Performance Evaluation and Result of Information Retrieval Exercise

To evaluate the performance of the proposed IR model, 150 DDI related abstracts for each type of studies (positive dataset) and 10,000 randomly selected abstracts (negative dataset) are built for training model. For testing, another 150 positive abstracts for each study are chosen; 500 drug- or nutrition-related abstracts plus 300 random articles are used to be negative dataset. In this work, the performance using Support Vector Machine (SVM) classifier in Weka is shown in Table 5.2. In this table, the desirable F-measures for in vitro PK, clinical PK, and clinical PD (0.94, 0.84, and 0.7, respectively) are obtained. Also, the recall rates for all three studies are high, which means most of DDI relevant abstracts in GS corpus can be captured.

SVM Classifier	Precision	Sensitivity	Specificity	F-Measure
In vitro PK	0.907407407	0.98	0.98125	0.942307692
Clinical PK	0.726829268	0.993333333	0.937777778	0.83943662
Clinical PD	0.589041096	0.86	0.8875	0.699186992

Table 5.2 Performance of IR Exercise

Using the same experiment setting with 300 positive abstracts and 10,000 randomly selected abstracts, a large-scale IR exercise from around 25 million abstracts in PubMed (1975-2015) is implemented. From the result of this exercise, the numbers of retrieved abstracts for in vitro PK, clinical PK, and clinical PD are 7,924, 20676, and 93653, respectively. However, among those abstracts, some research works are animal-related studies. To remove those abstracts, MeSH terms under tree number “B01.050” (Animal) are utilized. After eliminating animal-related abstracts, the numbers of retrieved abstracts for in vitro PK, clinical PK, and clinical PD are 5,199, 17,048, and 80,246.

To investigate the sensitivity of information retrieval exercise, golden standard corpora (218, 210, and 140 abstracts for Clinical PK, in vitro PK, and clinical PD studies) are used for evaluation. This exercise demonstrated that the proposed IR model can be able to capture more than 96% of relevant abstracts (Table 5.3).

SVM Classifier	Sensitivity
In vitro PK	210/210=100%
Clinical PK	210/218=96.3%
Clinical PD	138/140=98.6%

Table 5.3 Performance of Large-Scale IR Exercise

5.4.2 Performance Evaluation and Result of Information Extraction Exercise

To evaluate the performance for information extraction, golden standard corpora comprising of 210, 218, and 140 abstracts for in vitro PK, Clinical PK, and clinical PD studies are used. Utilizing various features calculated using the measurement of drugs' and interaction terms' location and existing knowledge for drugs, proposed classifiers were implemented to distinguish whether unique drug pairs in each abstract are interacting or not. In identifying drug interaction pairs, 5 feature groups and 7 classifiers are tested. The performance evaluation for in vitro PK, Clinical PK, and clinical PD are shown in Table 5.4, Table 5.5, and Table 5.6, respectively. For in vitro PK study, the optimal precision and recall rates are 0.76 and 0.91using feature group 5 (G5) with NaiveBays classifier. For clinical PK study, the optimal precision and recall rates are 0.87 and 0.82 using Feature Group 1 (G1) with IterativeClassifierOptimizer. For clinical PD study, the optimal precision and recall rates are 0.80 and 0.67 using feature group 1 (G1) with NaiveBays classifier.

	J48	NB	SMO	LogiR	RandF	LMT	ICO
Feature Group 1 (G1)							
P	0.778	0.555	0.754	0.750	0.683	0.618	0.737
R	0.525	0.950	0.575	0.563	0.513	0.588	0.525
F-1	0.627	0.700	0.652	0.643	0.586	0.603	0.613
Feature Group 2 (G2)							
P	0.724	0.664	0.737	0.790	0.741	0.423	0.736
R	0.525	0.938	0.525	0.613	0.500	0.513	0.663
F-1	0.609	0.777	0.613	0.690	0.597	0.463	0.697
Feature Group 3 Selected features by using stepwise regression (G3)							
P	0.825	0.672	0.909	0.900	0.833	0.825	0.880
R	0.598	0.943	0.460	0.724	0.460	0.598	0.506
F-1	0.693	0.785	0.611	0.803	0.593	0.693	0.642
Feature Group 4 (stepwise regression with interaction terms) (G4)							
P	0.765	0.784	0.952	0.783	0.865	0.729	0.836
R	0.747	0.874	0.230	0.540	0.517	0.494	0.701
F-1	0.756	0.826	0.370	0.639	0.647	0.589	0.763
Feature Group 5 All features (G5)							
P	0.836	0.760	0.882	0.887	0.880	0.558	0.880
R	0.529	0.908	0.517	0.724	0.506	0.494	0.506
F-1	0.648	0.827	0.652	0.797	0.642	0.524	0.642

Table 5.4 Performance Evaluation of IE for In Vitro PK Study

	J48	NB	SMO	LogiR	RandF	LMT	ICO
Feature Group 1 (G1)							
P	0.802	0.796	0.831	0.824	0.875	0.837	0.872
R	0.77	0.82	0.74	0.75	0.7	0.77	0.820
F-1	0.786	0.808	0.783	0.785	0.778	0.802	0.845
Feature Group 2 (G2)							
P	0.903	0.819	0.874	0.878	0.878	0.871	0.809
R	0.650	0.680	0.760	0.720	0.650	0.740	0.720
F-1	0.756	0.743	0.813	0.791	0.747	0.800	0.762
Feature Group 3 Selected features by using stepwise regression (G3)							
P	0.889	0.872	0.911	0.911	0.907	0.900	0.847
R	0.720	0.680	0.720	0.720	0.680	0.720	0.720
F-1	0.796	0.764	0.804	0.804	0.777	0.800	0.778
Feature Group 4 (stepwise regression with interaction terms) (G4)							
P	0.795	0.840	0.922	0.918	0.854	0.914	0.830
R	0.700	0.630	0.710	0.780	0.700	0.740	0.730
F-1	0.745	0.720	0.802	0.843	0.769	0.818	0.777
Feature Group 5 All features (G5)							
P	0.845	0.795	0.877	0.899	0.886	0.886	0.855
R	0.710	0.660	0.710	0.710	0.700	0.700	0.710
F-1	0.772	0.721	0.785	0.793	0.782	0.782	0.776

Table 5.5 Performance Evaluation of IE for Clinical PK Study

	J48	NB	SMO	LogiR	RandF	LMT	ICO
Feature Group 1 (G1)							
P	0.815	0.795	0.938	0.833	0.774	0.842	0.724
R	0.478	0.674	0.326	0.435	0.522	0.348	0.457
F-1	0.603	0.729	0.484	0.571	0.623	0.492	0.560
Feature Group 2 (G2)							
P	0.909	0.800	0.938	1.000	0.741	0.871	0.731
R	0.435	0.609	0.326	0.370	0.435	0.587	0.413
F-1	0.588	0.691	0.484	0.540	0.548	0.701	0.528
Feature Group 3 Selected features by using stepwise regression (G3)							
P	0.645	0.667	0.875	0.828	0.722	0.821	0.739
R	0.435	0.696	0.304	0.522	0.565	0.500	0.370
F-1	0.519	0.681	0.452	0.640	0.634	0.622	0.493
Feature Group 4 (stepwise regression with interaction terms) (G4)							
P	0.537	0.784	1.000	0.839	0.727	0.885	0.933
R	0.478	0.630	0.261	0.565	0.522	0.500	0.304
F-1	0.506	0.699	0.414	0.675	0.608	0.639	0.459
Feature Group 5 All features (G5)							
P	0.875	0.750	1.000	0.793	0.806	0.792	0.680
R	0.457	0.652	0.326	0.500	0.543	0.413	0.370
F-1	0.600	0.698	0.492	0.613	0.649	0.543	0.479

Table 5.6 Performance Evaluation of IE for Clinical PD Study

Using the optimal settings obtained from IE exercises, 3,894, 3,920, and 17,315 interacting drug pairs are extracted from 5,199 in vitro PK abstracts, 17,315 clinical PK abstracts, and 80,246 clinical PD abstracts, respectively. With those retrieved drug pairs, the Venn Diagram shown in Figure 5.6 was constructed to represent the overlapping among the drug combinations cross three study types. In this figure, the highlight must spot on those 986 unique drug pairs. That is because those 986 drug pairs are highly possible to be well-investigated in all three types of studies and can be potential for clinical utilities. Also, the genetics hypothesis for the relationship among drug-gene-ADR can be generated based on the validated result of those 986 drug pairs. Otherwise, another important number (2157) represents the overlapping between clinical PK and clinical PD studies, which means they have clinical PD/PK DDI evidence but their DDI mechanism in molecular level is unknown. 13,012 DDI pairs with only clinical PD evidence will have enormous research potential for pharmacology communities. Table 5.7 shows some existing examples of knowledge gaps from literature articles. With the cases of Theophylline & Ciprofloxacin, "Fexofenadine & Itraconazole" and "Theophylline & Propranolol", there are few publications, describing their mechanism in Clinical PK and PD levels, but in vitro PK mechanisms are still not very clear. With the example of "Tamoxifen & Midazolam", tamoxifen reversely inhibited midazolam in PK level but no pharmacodynamics evidence can be found. Another example of "Clopidogrel & Acetylsalicylic acid" shows 200 records only in Clinical PD result. Dual therapy with Acetylsalicylic acid and clopidogrel may result in an antiplatelet effect with fewer side effects. However, its mechanismcannot be found in clinical and in vitro PK studies but also not be predicted through the result of drug-gene interaction.

Drug pair	Evidence	PMID
Theophylline & Ciprofloxacin	Clinical PK and PD	MEDLINE:2328197,3571046
Fexofenadine & Itraconazole	Clinical PK and PD	MEDLINE:16669847,16796706
Theophylline & Propranolol	Clinical PK and PD	MEDLINE:4041342,7408406, 2888791
Tamoxifen & Midazolam	Clinical PK and In vitro PK	MEDLINE:12419016
Clopidogrel & Acetylsalicylic acid	Clinical PD	MEDLINE:16421012

Table 5.7 Examples of Knowledge Gaps for Drug Interacting Pairs

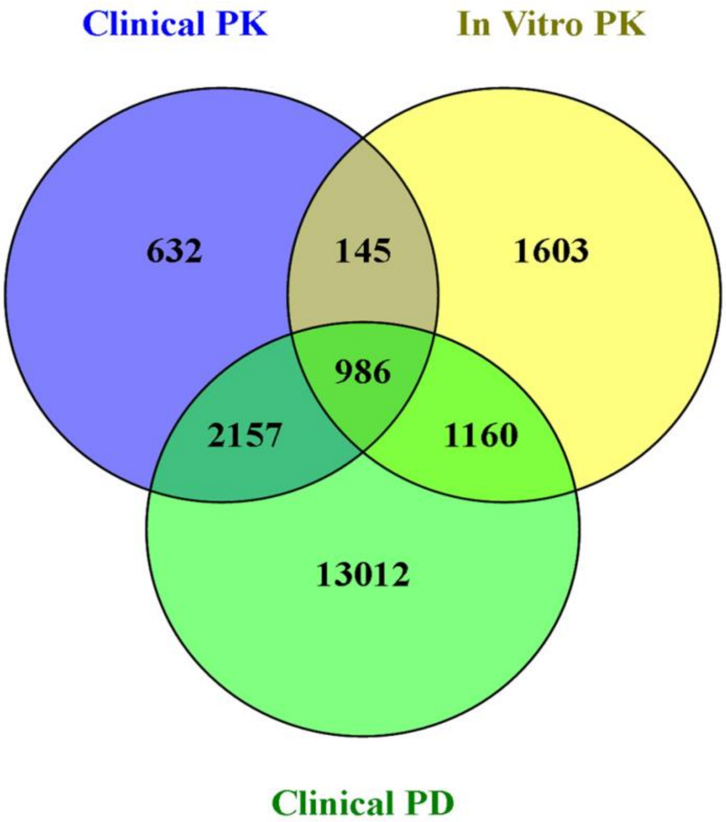


Figure 5.6 Venn Diagram of Drug Interaction Evidence for In Vitro PK, Clinical PK, and Clinical PD Studies

5.5 Validation

5.5.1 Validation with Drugbank Database

To evaluate the validity of our findings, the information of drug interaction in Drugbank is utilized. In Drugbank, drug interaction information were majorly collected from several sources, including Physician's Desk Reference ("Physicians' Desk Reference,"), e-Therapeutics ("e-Therapeutics,"), Medicines Complete ("Medicines Complete,"), Epocrates RX ("Epocrates RX,"), and Drugs.com ("Drugs.com,"). In this validation task, focusing on FDA approved and withdrawn drugs, we retrieved 46,244 drug interactions from Drugbank database. Unfortunately, Drugbank does not distinguish in vitro PK, clinical PK, and clinical PD evidence. To make a reasonable comparison, the overlapping between those of Drugbank (46,244) and all unique drug pairs from three groups (19,695) are used. This comparison task introduced 9,588 overlapping drug pairs. Otherwise, 10,107 drug pairs can be found in our result but not in Drugbank; 36,656 drug pairs can be found in Drugbank but not in our result. To discuss these three numbers, for 9,588 overlapping drug pairs, those drug combinations can be considered as reliable information. On the other hand, 36,656 drug pairs may come from the reports of studies or clinical trials in pharmacology companies, such as Physician's Desk Reference, Medicines Complete, or Epocrates RX. They were not always reported or published in published literature. For those 10,107 drug pairs found in our result but not in Drugbank, they can be used for improving the comprehensiveness of existing databases.

In addition, to demonstrate the reliability of our result, we manually validate the top 20 DDIs that are retrieved from all three evidences types. From the result of evaluation (Table 5.8), 19 out of 20 DDI pairs are validated to have interaction information from our discovery, but only 13 DDI pairs exist in Drugbank database. For these two false positive predictions in our result, both drug pairs, including “Simvastatin & Atorvastatin” (C10AA; C10BA; C10BX) belong to the same drug groups according to ATC classification. This means that they are frequently co-administrated in clinical trial or are compared in terms of efficacy.

Drug Pair	Frequency	Availability in Drugbank	Availability in our result
Fluorouracil & Leucovorin	256	YES	YES
Clopidogrel & Acetylsalicylic acid	200	YES	YES
Carboplatin & Paclitaxel	147	YES	YES
Gemcitabine & Cisplatin	119	NO	YES
Warfarin & Acetylsalicylic acid	113	YES	YES
Ritonavir & Lopinavir	111	YES	YES
Cisplatin & Fluorouracil	108	NO	YES
Cisplatin & Paclitaxel	101	YES	YES
Oxaliplatin & Fluorouracil	96	NO	YES
Cisplatin & Etoposide	95	YES	YES
Hydrocortisone & Corticotropin	95	NO	YES
Warfarin & Phylloquinone	90	YES	YES
Simvastatin & Ezetimibe	90	YES	YES
Fluorouracil & Irinotecan	77	NO	YES
Midazolam & Propofol	76	YES	YES
Clopidogrel & Prasugrel	76	YES	NO
Simvastatin & Atorvastatin	70	NO	NO
Ritonavir & Atazanavir	69	YES	YES
Cisplatin & Docetaxel	69	YES	YES
Oxaliplatin & Capecitabine	67	NO	YES

Table 5.8 Availability of Top 20 DDI Pairs (Retrieved from All Three Evidence Types) in Drugbank and Our Results

To further study the top 20 DDIs of each study type (by reporting frequency), manually validation was implemented. Of the top 20 drug pairs for in vitro PK, clinical PK, and clinical PD, 13, 9, and 9 were reported in Drugbank, but 20, 17, and 19) can be manually validated. From these two analyses, it can prove that our result not only can provide solid drug interaction information with experimental evidence from publication but also can further explore extra DDI information other than those in Drugbank.

5.5.2 Validation with Genetic Hypothesis

Clinical decisions typically stem from in vivo and clinical evidence. However, studying molecular interaction mechanisms in vitro is essential for understanding the hazards of specific drugs given certain genetic polymorphisms and for exploring potential alternative treatments. Since translational DDI research aims to link between knowledge of molecular mechanisms underlying DDI and their clinical consequences, it is of paramount importance to identify knowledge gaps that prevent such translation.

To fulfill such a translational research work, understanding how the relationship between genetics and DDI induced adverse drug events will be critical. In this task, 986 drug interaction pairs found from all three studies are valuable candidates for further investigation. Since we realized that CYP3A and CYP2D6 are two major cytochrome P450 systems that are responsible for 55% of drug metabolism; as such, drug interactions mediated through inhibition or induction of those CYPs might induce specific adverse drug events. With this concept, we are looking for all ADEs in those abstracts containing drug interaction via CYP3A and CYP2D6 and exploring their potential relationships. Among those 986 drug combinations, there are 71 sensitive substrates of CYP3A and 26 sensitive substrates of CYP2D6 co-occurring with 552 ADEs and 192 ADEs, respectively. The top 20 ADEs for both CYP3A and CYP2D6 substrates are shown in Table 5.9. Those ADEs provides us a starting point to explore potential candidates for the relationship of DDI-enzyme-ADE.

To validate the assumption of DDI-enzyme-ADE relationship, the exploration of ADEs caused by the interaction between sensitive substrates and strong inhibitors of CYP3A and CYP2D6 was manually extracted from published articles. Table 5.10 and Table 5.11 show the sentences that describe DDI-induced ADEs via the pathways of CYP3A and CYP2D6, respectively. Taking one example from Table 5.10, in PMID: 8623953, the psychomotor effect is enhanced by the interaction of Midazolam and Itraconazole and the interaction is significant in statistics. This result can be supported using the combined knowledge collected from PMID: 17655375 and PMID: 20739919. In PMID: 17655375, this article mentions that Midazolam is the sensitive substrate of CYP3A. If its CYP3A activity is inhibited, it might result in prolonged drowsiness and inhibition of psychomotor performance, which means CYP3A is dependent on psychomotor performance. In addition, in PMID: 20739919, Itraconazole is studied to be a strong CYP3A inhibitor. Therefore, with such collective information, the evidence we found in PMID: 8623953 can be validated using the integrated information obtained from PMID: 17655375 and PMID: 20739919. With this exploration, those DDI-induced ADE instances further support by the prediction using our genetic hypothesis. Therefore, such a comprehensive drug-interaction evidence of all three types combining with the prediction of genetic hypothesis can be an essential and fundamental step toward developing reliable clinical decision systems.

CYP3A sensitive substrate	Report Freq.	CYP2D6 sensitive substrate	Report Freq.
IMMUNODEFICIENCY	322	DEPRESSION	65
BLOOD CHOLESTEROL	281	HEART RATE	33
VIRAL LOAD	180	SENSATION OF PRESSURE	25
TRANSPLANT	157	TACHYCARDIA	24
RENAL TRANSPLANT	154	NOREPINEPHRINE	19
DIARRHOEA	153	NAUSEA	17
BLOOD TRIGLYCERIDES	141	ANXIETY	16
LOW DENSITY LIPOPROTEIN	111	NERVOUSNESS	14
RHABDOMYOLYSIS	105	DIABETES MELLITUS	13
HEPATITIS C VIRUS TEST	98	ELECTROCARDIOGRAM QT INTERVAL	13
NAUSEA	97	EOSINOPHILIA MYALGIA SYNDROME	13
SENSATION OF PRESSURE	95	PREMENSTRUAL SYNDROME	13
HEADACHE	92	INFUSION	12
HYPERCHOLESTEROLAEMIA	87	PAIN	12
IMMUNOSUPPRESSION	87	SCHIZOPHRENIA	12
ASTHENIA	73	DIZZINESS	10
HYPERTENSION	73	HEADACHE	10
RENAL FAILURE	70	HYPERHIDROSIS	10
CORONARY ARTERY DISEASE	69	ARRHYTHMIA	9
MYOPATHY	69	BLOOD PRESSURE DIASTOLIC	9
ELECTROCARDIOGRAM QT INTERVAL	64	DIARRHOEA	8
HYPERLIPIDAEMIA	64	HOT FLUSH	8
PROPHYLAXIS	62	VENTRICULAR EXTRASYSTOLES	8
HIGH DENSITY LIPOPROTEIN	61	AFFECT LABILITY	7
RASH	61	FATIGUE	7
HEART RATE	58	BIPOLAR DISORDER	6
NEPHROPATHY TOXIC	56	CARDIAC FIBRILLATION	6
DRUG TOLERANCE	50	CONSTIPATION	6
LIVER TRANSPLANT	50	GASTRIC PH	6
VOMITING	47	INSOMNIA	6

Table 5.9 The Top 20 ADEs for Both CYP3A and CYP2D6 Substrates

Substrate	Inhibitor	Clinical Effect	PMID	Genetic Effect	PMID
Midazolam	Itraconazole	Increased psychomotor effects	8623953	psychomotor performance	17655375
				Itraconazole contribute to CYP3A4 inhibition	20739919
Triazolam	Itraconazole	Increased PD effects	8841155	Significant benzodiazepine agonist-like pharmacodynamic effects	10773013
				Itraconazole is a potent inhibitor of cytochrome P450 (CYP) 3A	20497744
Midazolam	Ritonavir	Prolonged sedation effect	19792991	Sedative effect	12402721
				ritonavir is an inhibitor of CYP3A	20002087
Buspirone	Itraconazole	Increased PD effect	9333111	Pharmacodynamic effect	18220561
				Itraconazole is a potent inhibitor of CYP3A4	9333111
Midazolam	Saquinavir	Profound sedative effects.	10430107	Sedative effect	12402721
				Saquinavir is an inhibitor of CYP3A4	19792991
Tacrolimus	Itraconazole	Developed renal dysfunction	16503502	Decline in renal function	22205779; 28280692
				Itraconazole is a potent inhibitors of cytochrome P450 (CYP) 3A	22971159
Midazolam	Voriconazole	Increased the psychomotor effects	16580904	Prolonged drowsiness and inhibition of psychomotor performance.	17655375
				Voriconazole is an inhibitor of CYP3A4	16205037
Simvastatin	Clarithromycin	Increased risk of death or hospitalisation	26497728	hospitalisation	25571290
				Clarithromycin is a potent inhibitor of CYP3A4.	
Everolimus	Voriconazole	Increased kidney transplant recipient	25417855	kidney transplant recipient	19499965
				Voriconazole inhibits P450-3A4 actitivy	25417855
Triazolam	Ritonavir	Increased benzodiazepine agonist properties	10935688	Significant benzodiazepine agonist-like pharmacodynamic effects	16513448; 10773013
				Ritonavir inhibits both enteric and hepatic CYP3A	16513448
Triazolam	Clarithromycin	Enhanced Benzodiazepine agonist effects	9757151	Significant benzodiazepine agonist-like pharmacodynamic effects	9757151; 10773013
				Clarithromycin, a potent inhibitor of CYP3A	19897389
Simvastatin	Itraconazole	Enhanced myotoxicity risk	18563955	risk of myotoxicity	17178259
				Itraconazole is a strong inhibitor of CYP3A4	17178259
Atorvastatin	Itraconazole	Experienced dry skin and vomiting	11061579	Adverse events (abdominal distention, nausea, vomiting, and hunger)	11061579; 28207527
				Itraconazole, a potent inhibitor of CYP3A4	9695720
Alprazolam	Ritonavir	Enhanced PD effects consistent with its benzodiazepine agonist properties	10801241	PD effects	10801241
				ritonavir is both an inhibit and an inducer of CYP3A	10801241
Tacrolimus	Telaprevir	Adverse events of mild pruritusand mild excoriation	21618566	Pruritusand	22205779
				Telaprevir is an inhibitor of the enzyme cytochrome P450 3A	21618566

Table 5.10 DDI Pairs via CYP3A and Their ADR Found Three Types of Studies

Substrate	Inhibitor	Effect when coadministration	PMID	Genetic Effect	PMID
Desipramine	Paroxetine	Increased tiredness	14730412	Side effects, including nausea, tiredness, dizziness.	20840444; 2271367
				Paroxetine is a potent inhibitor of CYP2D6	12584155
Dextromethorphan	Quinidine	Increased QTc intervals or headedness	20373255	Side effects were light-headedness, slurred speech	28290770; 7998781
				Quinidine (Q), a potent cytochrome P450 2D6 inhibitor	20839238
Dextromethorphan	Paroxetine	The incidence of AEs was higher	22283559	Insomnia was reported	28290770; 15231978
				Paroxetine is a potent inhibitor of CYP2D6	15903129
Metoprolol	Paroxetine	The heart rate and systolic blood pressure decreased	21923449	Increased HRV indexes	12891223; 8607401
				Paroxetine is a very potent inhibitor of CYP2D6	15903129
Propafenone	Quinidine	Reduction of premature ventricular contraction	3630896	70--80% reduction of total number of PVCs	23585605; 7201833
				Quinidine (Q), a potent cytochrome P450 2D6 inhibitor	20839238
Propranolol	Quinidine	A significant inhibition of exercise-induced tachycardia	2093126	cardiac failure in 2 patients	9399616; 5922889
				Quinidine (Q), a potent cytochrome P450 2D6 inhibitor	20839238
Imipramine	Fluoxetine	QT interval prolongation	15687478	Prolong QTc interval	9205822; 11830802
				Potent inhibition of cytochrome P450 2D6 (CYP2D6) by fluoxetine	8477556
Oxycodone	Paroxetine	VA scores for subjective drug effects, drowsiness and deterioration of performance were slightly increased	20642550	ADEs: Nausea and pruritus	20590588; 20857093
				Paroxetine is a potent inhibitor of CYP2D6	12584155
Mexiletine	Quinidine	Increased ERP, thereby producing greater postrepolarization refractoriness than either drug alone	2481766	Increased the ERP/APD ratio	9690950; 2795468
				Quinidine (Q), a potent cytochrome P450 2D6 inhibitor	20839238

Table 5.11 DDI Pairs via CYP2D6 and Their ADR Found Three Types of Studies

5.6 Conclusion

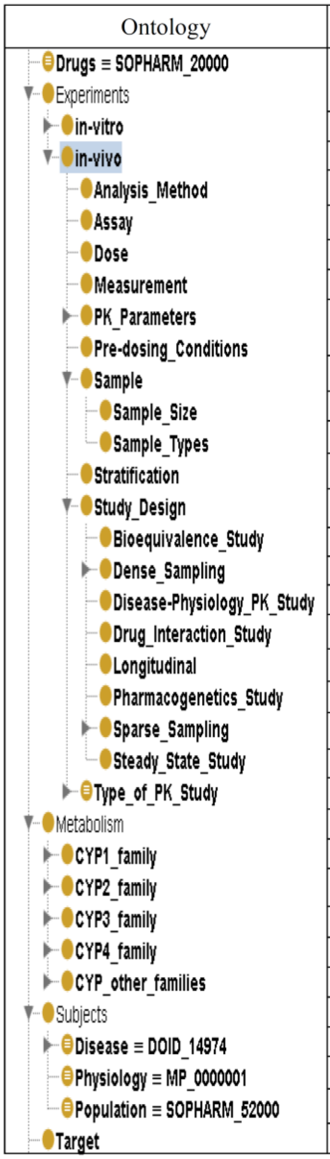
The successful completion of Aim 3 will result in three sets of DDI relevant abstracts (from IR exercise) and their corresponding drug interacting pairs (from IE exercise) for in vitro PK, clinical PK and clinical PD studies. The aggregative information will introduce a Venn Diagram and show the overlapping cross three different evidences, which intuitively reveals the practical status of drug interaction research work. Notably, this task need not be completed using complicated natural language process technology for creating features for machine learning and further facilitate the downstream development of more effective clinical decision systems with the use of collective knowledge. Indeed, the outcome of the proposed work provides an exciting opportunity to promote translation of molecular to clinical research.

Chapter 6. Conclusion and Future Work

The contribution of this work can be listed in the following: 1) Provide a comprehensive lexicon for DDI related terminologies; 2) Provide finely curated corpus with semantic information for specific entities and drug interactions and introduce a DDI annotation guidance and; 3) Propose a NER tool for identifying drug metabolite; 4) Construct a text mining pipeline to retrieve, extract, and explore the knowledge gap for drug interactions. 5) Utilize a hypothesis based on genetic studies to generate a dataset that contains the strong evidences of DDI-induced ADEs cross all three studies.

To further uncover the unknown components and close such a knowledge gap in clinical PD and PK evidence on DDI, *DDI prediction using Drug-Gene Interaction* (DGI) is a potential method to identify the unknown components of potential DDI in in vitro PK studies. To identify the drug-gene relationship within Medline abstracts for predicting DDI evidence in molecular level, the methodology is similar to the work of DDI retrieval and extraction. Based on the discovery, this task will be able to explore those drug pairs that have clinical PK or PD DDI evidences but no existing in vitro investigation. Even though we cannot provide granular information about interaction types for specific drug pairs, the asset of this work is to reflect the big picture of drug interaction for pharmacology research.

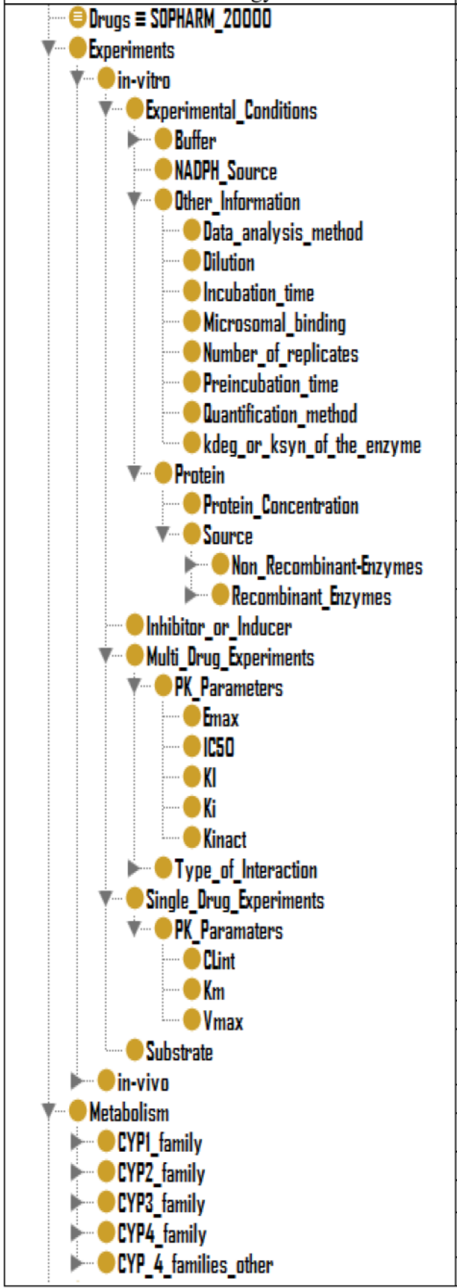
Supplementary



Ontology	Pharmacogenetics Trial	Drug Interaction Trail
Drugs = SOPHARM_20000	Tamoxifen (TAM)	Midazolam (MDZ, PO 4mg; IV 0.05mg/kg), Ketoconazole (KTZ, PO, 200, 400 mg)
Experiments		
in-vitro		
in-vivo	*in-vivo*	*in-vivo*
Analysis_Method		
Assay	HPLC/MS	HPLC/MS
Dose	SOLTAMOX™, 20mg/day	MDZ PO, IV; KTZ PO
Measurement	month 1, 4, 8, 12	before and 0.5, 0.75, 1, 2, 4, 6, 9 hrs
PK_Parameters	TAM and its metabolites conc	MDZ and KTZ: AUC, AUCR, $t_{1/2}$, and Cmax
Pre-dosing_Conditions		
Sample		
Sample_Size	298	24
Sample_Types	Blood	blood
Stratification	prior chemo, menopausal	
Study_Design		
Bioequivalence_Study		
Dense_Sampling		
Disease-Physiology_PK_Study		
Drug_Interaction_Study		inhibition
Longitudinal	Longitudinal	three-phase crossover
Pharmacogenetics_Study	prospective, single arm	prospective, single arm
Sparse_Sampling		
Steady_State_Study	steady state	
Type_of_PK_Study		
Metabolism		
CYP1_family		
CYP2_family	CYP2D6, 2C9, 2B6	
CYP3_family	CYP3A4/5	CYP3A4/5
CYP4_family		
CYP_other_families		
Subjects		
Disease = DOID_14974	breast cancer	healthy volunteers
Physiology = MP_0000001		
Population = SOPHARM_52000	Caucasian/African American	
Target	ESR1/ESR2	

Table S1 Clinical PK Studies

Note: The annotations are aligned for each row. The left column is the ontology tree presentation. The central and right columns display their corresponding annotations from the paper.



Ontology	in-vitro study
Drugs ≡ SOPHARM_20000	MDZ, APZ, TZ, CLAR, TAM, DTZ, NIF, BFC, HFC, TEST, E2
Experiments	
in-vitro	Compare metabolic capabilities of CYP3A4, 3A5, 3A7
Experimental_Conditions	
Buffer	
NADPH_Source	sodium phosphate, NADPH, methanol.
Other_Information	
Data_analysis_method	
Dilution	WinNonlin
Incubation_time	4 fold, 10% methanol (TZ)
Microsomal_binding	5 min
Number_of_replicates	insect cell (CYP3A)
Preincubation_time	N/A
Quantification_method	3min; 6 min
kdeg_or_ksyn_of_the_enzyme	HPLC, MS, Fluorimetry
Protein	CYP3A4/5/7, P450 reductase, b5
Protein_Concentration	1mol, 6.6mol, 9mol
Source	BD Gentest, PanVera, PanVera
Non_Recombinant-Enzymes	
Recombinant_Enzymes	CYP3A
Inhibitor_or_Inducer	
Multi_Drug_Experiments	
PK_Parameters	
Emax	
IC50	
KI	
Ki	
Kinact	
Type_of_Interaction	
Single_Drug_Experiments	
PK_Paramaters	
CLint	CL for individual substrates
Km	Km for individual substrates
Vmax	Vmax for individual substrates
Substrate	MDZ, APZ, TZ, CLAR, TAM, DTZ, NIF, BFC, HFC, TEST, E2
in-vivo	
Metabolism	
CYP1_family	
CYP2_family	
CYP3_family	CYP3A4, 3A5, 3A7
CYP4_family	
CYP_4_families_other	

Table S2 In Vitro PK Study

Note: The annotations are aligned for each row. The left column is the ontology tree presentation. The central and right columns display their corresponding annotations from the paper.

Journal Name	Frequency	Journal Category
Arch Intern Med	7	Epidemiology
Am J Cardiol	6	Special Clinical Domains
Pharmacoepidemiol Drug Saf	6	Epidemiology
Am J Med	5	Epidemiology
Clin Pharmacol Ther	5	Pharmacology
Drug Saf	4	Epidemiology
Am J Health Syst Pharm	3	Pharmacology
Br J Clin Pharmacol	3	Pharmacology
J Manag Care Pharm	3	Pharmacology
JAMA	3	Epidemiology
Am J Gastroenterol	2	Special Clinical Domains
Ann Pharmacother	2	Pharmacology
Arthritis Rheum	2	Special Clinical Domains
Clin Ther	2	Pharmacology
J Clin Epidemiol	2	Epidemiology
J Clin Pharm Ther	2	Pharmacology
Am J Cardiovasc Drugs	1	Special Clinical Domains
Am J Geriatr Pharmacother	1	Pharmacology
Ann Intern Med	1	Special Clinical Domains
Ann Med	1	Special Clinical Domains
Arch Gen Psychiatry	1	Special Clinical Domains
Arthritis Res Ther	1	Special Clinical Domains
BMJ	1	Special Clinical Domains
CMAJ	1	Special Clinical Domains
Clin J Am Soc Nephrol	1	Special Clinical Domains
Gastroenterology	1	Special Clinical Domains
J Am Coll Cardiol	1	Special Clinical Domains
J Am Geriatr Soc	1	Special Clinical Domains
J Med Assoc Thai	1	Special Clinical Domains
Med Care	1	Special Clinical Domains
N Engl J Med	1	Epidemiology
PLoS One	1	Special Clinical Domains
Pediatr Allergy Immunol	1	Special Clinical Domains
Pharmacotherapy	1	Pharmacology
Res Social Adm Pharm	1	Pharmacology
Rheumatology (Oxford)	1	Special Clinical Domains
Thromb Haemost	1	Special Clinical Domains
Pubmedhealth	1	Special Clinical Domains
PLoS Comput Biol	1	Pharmacology

Table S3 Pharmaco-epidemiology Drug Interaction Associated Journal Names, Abstract Frequencies and Journal Categories

Note: The selected journals are highlighted in bold

Category	PMID and Example Abstract
Pharmacoepidemiology drug interactions	Medline 16581331
Epidemiological study on the frequency of the known drug interactions in the health databases	Medline 12071783
Implementations of drug interaction software	Medline 16622155
Population drug safety study, but no DDI information	Medline: 8876849
Reviews	Medline 1312320

Table S4 Categories and Examples of Related and no-Related Clinical PD DDI Abstracts

Group	Reaction	Phase	
acetyl	Acetylation	Phase II	Acetyltransferase
acyl	Amino Acid Conjugation	Phase II	UDP-glucuronosyltransferase
adenosyl	Methylation	Phase II	Methyltransferases
alkyl	Oxidation, dealkylation	Phase I	Cytochrome P450
amide	Hydrolysis	Phase I	Amidases
amino and aryl	Acetylation	Phase II	Acetyltransferases
aroma	Oxidation	Phase I	
ascorbyl			
azo (aryl or alkyl nitro)	Reduction	Phase I	This hydrolysis can occur with no enzyme involved,
butylat			
carbamoyl			
carbonyl	Reduction	Phase I	carbonyl reductases
carboxyl	Conjugation	Phase II	UDP-glucuronosyltransferase to form aceyl-glucuronides
carboxylate	Glucuronidation	Phase II	UDP-glucuronosyltransferase to form aceyl-glucuronides
carboxylic	Glucuronidation	Phase II	UDP-glucuronosyltransferase to form aceyl-glucuronides
decarboxyl	Hydrolysis	Phase I	decarboxylase
deacetyl	Hydrolysis	Phase I	deactylation
dealkyl	Oxidation	Phase I	Cytochrome P450
debutyl	Oxidation	Phase I	Cytochrome P450
dechloroethyl	Oxidation	Phase I	Cytochrome P450
deethyl	Oxidation	Phase I	Cytochrome P450
dehydroly	Oxidation	Phase I	Cytochrome P450
demethyl	Oxidation	Phase I	Cytochrome P450
deoxy			
desalkyl	Oxidation	Phase I	Cytochrome P450
desisopropyl	Oxidation	Phase I	Cytochrome P450
desmethyl	Oxidation	Phase I	Cytochrome P450
desulfur	Oxidation	Phase I	Cytochrome P450
dehalo	Reduction	Phase I	Cytochrome P450
didemethyl	Oxidation	Phase I	Cytochrome P450
didesmethyl	Oxidation	Phase I	Cytochrome P450
diethyl	Oxidation	Phase I	Cytochrome P450
dihydro	Reduction	Phase I	Cytochrome P450
esters	Hydrolysis	Phase I	
epoxide	Hydrolysis	Phase I	
ethyl	Oxidation (dethylation)	Phase I	
fluoro	Oxidation (defluorination)	Phase I	
glucuronide	Glucuronidation	Phase II	UDP-glucuronosyltransferase to form aceyl-glucuronides
glutathione	Glutathione Conjugation	Phase II	Glutathione S-transferases
hydroxy	Oxidation	Phase I	Cytochrome P450
hydroxyl	Oxidation	Phase I	Cytochrome P450
OH	Oxidation	Phase I	Cytochrome P450
oxid	Oxidation	Phase I	Cytochrome P450
hydroxylamine	Glucuronidation	Phase II	
keto	Hydrolysis	Phase I	Aldo-keto reductase
lactone	Hydrolysis	nonenzymatic	
laurate			
methoxy			
methyl	Methylation	Phase II	
methylhydroxy	Oxidation	Phase I	Cytochrome P450
nitro	Reduction	Phase I	
nor	Oxidation	Phase I	Cytochrome P450
phenyl	Oxidation	Phase I	
phosphosulfate	Sulphation	Phase II	
sulfate	Sulfate conjugation	Phase II	Sulfotransferases
sulfhydryl	Glucuronidation	Phase II	
sulfide	Hydrolysis	nonenzymatic	
sulfoxide	Oxidation	Phase I	
tetrahydro	Reduction	Phase I	
threo	Reduction	Phase I	Aldoketoreductase
Sulfur	Oxidation	Phase I	Cytochrome P450
N-oxide	Oxidation	Phase I	Cytochrome P450 or FMOs
alcohol	Oxidation	Phase I	Dehydrogenases
aldehyde	Oxidation	Phase I	Dehydrogenases

Table S5 The List of Prefix and Suffix for Drug Metabolite Reaction

Reference

Airola, A., Pyysalo, S., Bjorne, J., Pahikkala, T., Ginter, F., & Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinformatics, 9 Suppl 11, S2. doi:10.1186/1471-2105-9-S11-S2

Ajayi, F. O., Sun, H., & Perry, J. (2000). Adverse drug reactions: a review of relevant factors. J Clin Pharmacol, 40(10), 1093-1101.

Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W. F. t., Warner, C., Hwang, J. D., . . . Savova, G. K. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. J Am Med Inform Assoc, 20(5), 922-930. doi:10.1136/amiajnl-2012-001317

Alias-i. (2008). LingPipe 4.1.0. Retrieved from http://alias-i.com/lingpipe

Becker, L. B., Kallewaard, M., Caspers, P.W., Visser, L.E., Leufkens, H.G., et al. . (2007). Hospitalisations and emergency department visits due to drug–drug interactions: a literature review. Pharmacoepidemiol. Drug Saf., 16(6), 641-651.

Becker, M. L., Kallewaard, M., Caspers, P. W., Visser, L. E., Leufkens, H. G., & Stricker, B. H. (2007). Hospitalisations and emergency department visits due to drug–drug interactions: a literature review. pharmacoepidemiology and drug safety, 16, 641–651.

Björne, J., Kaewphan, S., & Salakoski, T. (2013, 2013//). UTurku: Drug Named Entity Recognition and Drug-Drug Interaction Extraction Using SVM Classification and Domain Knowledge. Paper presented at the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).

Borges, S., Desta, Z., Jin, Y., Faouzi, A., Robarge, J. D., Philips, S., . . . Li, L. (2010). Composite functional genetic and comedication CYP2D6 activity score in predicting tamoxifen drug exposure among breast cancer patients. J Clin Pharmacol, 50(4), 450-458. doi:10.1177/0091270009359182

Boyce, R., Collins, C., Horn, J., & Kalet, I. (2009a). Computing with evidence Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment. J Biomed Inform, 42(6), 979-989. doi:10.1016/j.jbi.2009.05.001

Boyce, R., Collins, C., Horn, J., & Kalet, I. (2009b). Computing with evidence Part II: An evidential approach to predicting metabolic drug-drug interactions. J Biomed Inform, 42(6), 990-1003. doi:10.1016/j.jbi.2009.05.010

Boyce, R., Collins, C., Horn, J., and Kale, I. (2009). Computing with evidence Part II: An evidential approach to predicting metabolic drug-drug interactions. J Biomed Inform, 42(6), 990-1003.

Boyce, R., Collins, C., Horn, J., Kalet, I. (2009). Computing with evidence Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment. J Biomed Inform., 42(6), 979-989.

Boyce, R. D., Collins, C., Clayton, M., Kloke, J., & Horn, J. R. (2012). Inhibitory metabolic drug interactions with newer psychotropic drugs: inclusion in package inserts and influences of concurrence in drug interaction screening software. Ann Pharmacother, 46(10), 1287-1298. doi:10.1345/aph.1R150

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol., 2(3), 1-27. doi:10.1145/1961189.1961199

Chen, Y., Liu, F., & Manderick, B. (2009). Normalizing Interactor Proteins and Extracting Interaction Protein Pairs using Support Vector Machines. Paper presented at the BioCreative II. 5 Workshop 2009 on Digital Annotations.

Chien, J. Y., Lucksiri, A., Ernest, C. S., 2nd, Gorski, J. C., Wrighton, S. A., & Hall, S. D. (2006). Stochastic prediction of CYP3A-mediated inhibition of midazolam clearance by ketoconazole. Drug Metab Dispos, 34(7), 1208-1219. doi:10.1124/dmd.105.008730

Chou, T. C. (2006). Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. Pharmacol Rev, 58(3), 621-681. doi:10.1124/pr.58.3.10

Corbett, P., & Murray-Rust, P. (2006). High-Throughput Identification of Chemistry in Life Science Texts. In M. R. Berthold, R. Glen, & I. Fischer (Eds.), Computational Life Sciences II (Vol. 4216, pp. 107-118): Springer Berlin Heidelberg.

Crews, K. R., Gaedigk, A., Dunnenberger, H. M., Klein, T. E., Shen, D. D., Callaghan, J. T., . . . Clinical Pharmacogenetics Implementation, C. (2012). Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for codeine therapy in the context of cytochrome P450 2D6 (CYP2D6) genotype. Clin Pharmacol Ther, 91(2), 321-326. doi:10.1038/clpt.2011.287

David, C., Sérgio, M., & José Luís, O. (2012). Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools.

De Marneffe, M., MacCartney, B., Manning, C. (2006). Generating typed dependency parses from phrase structure parses. Paper presented at the LREC.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., . . . Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res, 36(Database issue), D344-350. doi:10.1093/nar/gkm791

Desta, Z., Ward, B. A., Soukhova, N. V., & Flockhart, D. A. (2004). Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system in vitro: prominent roles for CYP3A and CYP2D6. J Pharmacol Exp Ther, 310(3), 1062-1075. doi:10.1124/jpet.104.065607

DiMasi, J. A., & Grabowski, H. G. (2007). The Cost of Biopharmaceutical R&D: Is Biotech Different? Managerial and Decision Economics, 28, 469–479.

Drugs.com. Retrieved from Drugs.com

Duke, J. D., Han X Fau - Wang, Z., Wang Z Fau - Subhadarshini, A., Subhadarshini A Fau - Karnik, S. D., Karnik Sd Fau - Li, X., Li X Fau - Hall, S. D., . . . Li, L. (2012). Literature based drug interaction prediction with clinical assessment using. PLoS Comput Biol, 8(8), e1002614 LID - 1002610.1001371/journal.pcbi.1002614 [doi].

e-Therapeutics. Retrieved from http://www.e-therapeutics.ca

Eltyeb, S., & Salim, N. (2014). Chemical named entities recognition: a review on approaches and applications. J Cheminform, 6, 17. doi:10.1186/1758-2946-6-17

Epocrates RX. Retrieved from http://www.epocrates.com/products/features

Feldman, R., Regev, Y., Finkelstein-Landau, M., Hurvitz, E., & Kogan, B. (2002). Mining biomedical literature using information extraction. Current Drug Discovery, 19-23.

First Databank. (2014). Retrieved from http://www.fdbhealth.com/

FP, G. (2008). Cytochrome p450 and chemical toxicology. Chem Res Toxicol, 21(1), 70-83.

Fukuda, K., Tamura, A., Tsunoda, T., & Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput, 707-718.

Gene Ontology Consortium: going forward. (2015). Nucleic Acids Res, 43(Database issue), D1049-1056. doi:10.1093/nar/gku1179

Gibaldi M, P. D. (1982). Pharmacokinetics (2nd ed.). New York: Marcel Dekker.

Golan, D. E. T. A. H. (2012). Principles of pharmacology : the pathophysiologic basis of drug therapy. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Gottlieb, A., Stein, G. Y., Oron, Y., Ruppin, E., & Sharan, R. (2012). INDI: a computational framework for inferring drug interactions and their associated recommendations. Mol Syst Biol, 8, 592. doi:10.1038/msb.2012.26

Hachad, H., Ragueneau-Majlessi, I., & Levy, R. H. (2010). A useful tool for drug interaction evaluation: the University of Washington Metabolism and Transport Drug Interaction Database. Hum Genomics, 5(1), 61-72.

Hacker, M., Messer II, W. S., & Bachmann, K. A. (2009). Pharmacology: principles and practice: Academic Press.

Hajjar ER, C. A., Hanlon JT. (2007). Polypharmacy in elderly patients. Am J Geriatr Pharmacother.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. SIGKDD Explor. Newsl., 11(1), 10-18. doi:10.1145/1656274.1656278

Hall, M. J., DeFrances, C. J., Williams, S. N., Golosinskiy, A., & Schwartzman, A. (2010). National Hospital Discharge Survey: 2007 summary. Natl Health Stat Report(29), 1-20, 24.

Hennessy, S., & Flockhart, D. A. (2012). The need for translational research on drug-drug interactions. Clin Pharmacol Ther, 91(5), 771-773. doi:10.1038/clpt.2012.39

Herrero-Zazo, M., Segura-Bedmar I Fau - Martinez, P., Martinez P Fau - Declerck, T., & Declerck, T. (2013). The DDI corpus: an annotated corpus with pharmacological substances and drug-drug. J Biomed Inform, 46(5), 914-920 LID - 910.1016/j.jbi.2013.1007.1011 [doi] LID - S1532-0464(1013)00112-00113 [pii].

Hewett, M., Oliver, D. E., Rubin, D. L., Easton, K. L., Stuart, J. M., Altman, R. B., & Klein, T. E. (2002). PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res, 30(1), 163-165.

Huang, S. M., Temple, R., Throckmorton, D. C., & Lesko, L. J. (2007). Drug interaction studies: study design, data analysis, and implications for dosing and labeling. Clin Pharmacol Ther, 81(2), 298-304. doi:10.1038/sj.clpt.6100054

HUGO Gene Nomenclature Committee at the European Bioinformatics Institute. Retrieved from http://www.genenames.org

International Transporter, C., Giacomini, K. M., Huang, S. M., Tweedie, D. J., Benet, L. Z., Brouwer, K. L., . . . Zhang, L. (2010). Membrane transporters in drug development. Nat Rev Drug Discov, 9(3), 215-236.

Isoherranen, N., Kunze, K. L., Allen, K. E., Nelson, W. L., & Thummel, K. E. (2004). Role of itraconazole metabolites in CYP3A4 inhibition. Drug Metab Dispos, 32(10), 1121-1131. doi:10.1124/dmd.104.000315

Jia, J., Zhu, F., Ma, X., Cao, Z., Cao, Z. W., Li, Y., . . . Chen, Y. Z. (2009). Mechanisms of drug combinations: interaction and network perspectives. Nat Rev Drug Discov, 8(2), 111-128. doi:10.1038/nrd2683

Johansson, I., & Ingelman-Sundberg, M. (2011). Genetic polymorphism and toxicology--with emphasis on cytochrome p450. Toxicol Sci, 120(1), 1-13. doi:10.1093/toxsci/kfq374

Johnson, M. D., Zuo, H., Lee, K. H., Trebley, J. P., Rae, J. M., Weatherman, R. V., . . . Skaar, T. C. (2004). Pharmacological characterization of 4-hydroxy-N-desmethyl tamoxifen, a novel active metabolite of tamoxifen. Breast Cancer Res Treat, 85(2), 151-159. doi:10.1023/B:BREA.0000025406.31193.e8

Juurlink, D. N., Mamdani, M., Kopp, A., Laupacis, A., & Redelmeier, D. A. (2003). Drug-drug interactions among elderly patients hospitalized for drug toxicity. Jama, 289(13), 1652-1658. doi:10.1001/jama.289.13.1652

K., F., R., K. f., & R., Z. (2007). RelEx—relation extraction using dependency parse trees. Bioinformatics, 23, 365–371.

Karnik, S., Subhadarshini, A., Wang, Z., Rocha, L. M., & Li, L. (2011). Extraction of drug-drug interactions using all paths graph kernel. Paper presented at the the 1st Challenge task on Drug Drug Interaction Extraction, Huelva, Spain.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus--semantically annotated corpus for bio-textmining. Bioinformatics, 19 Suppl 1, i180-182.

Kim, J. D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics, 19(Supp 1), i180-182.

Klotz, U. (2007). The role of pharmacogenetics in the metabolism of antiepileptic drugs: pharmacokinetic and therapeutic implications. Clin Pharmacokinet, 46(4), 271-279. doi:10.2165/00003088-200746040-00001

Knapp, D., & Tomita, D. (1987). Second Annual Adverse Drug/Biologic Reaction Report.

Knollmann, L. L. B. B. A. C. B. C. (2011). Goodman and Gilman"S the Pharmacological Basis of Therapeutics: McGraw-Hill Professional.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., . . . Wishart, D. S. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res, 39(Database issue), D1035-1041. doi:10.1093/nar/gkq1126

Kolchinsky, A., Lourenco A Fau - Li, L., Li L Fau - Rocha, L. M., & Rocha, L. M. (2013). Evaluation of linear classifiers on articles containing pharmacokinetic evidence. Pac Symp Biocomput, 409-420.

Kolchinsky, A., Lourenco, A., Li, L., & Rocha, L. M. (2013). Evaluation of linear classifiers on articles containing pharmacokinetic evidence of drug-drug interactions. Pac Symp Biocomput, 18, 409-420.

Krauthammer, M., Rzhetsky, A., Morozov, P., & Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. Gene, 259(1-2), 245-252. doi:Doi 10.1016/S0378-1119(00)00431-5

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage.

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Thousand Oaks, CA: SAGE Publications, Inc.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic Model Trees. Mach. Learn., 59(1-2), 161-205. doi:10.1007/s10994-005-0466-3

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., . . . Wishart, D. S. (2014). DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res, 42(Database issue), D1091-1097. doi:10.1093/nar/gkt1068

Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput, 652-663.

Lee, K. H., Ward, B. A., Desta, Z., Flockhart, D. A., & Jones, D. R. (2003). Quantification of tamoxifen and three metabolites in plasma by high-performance liquid chromatography with fluorescence detection: application to a clinical trial. J Chromatogr B Analyt Technol Biomed Life Sci, 791(1-2), 245-253.

Li, L. (2007). Discussion on parameter estimation for differential equations: a generalized smoothing approach. Journal of Royal Statistics Sociaty, Series B, 69, 787-788.

Li, L., Yu, M., Chin, R., Lucksiri, A., Flockhart, D. A., & Hall, S. D. (2007). Drug-drug interaction prediction: a Bayesian meta-analysis approach. Stat Med, 26(20), 3700-3721. doi:10.1002/sim.2837

LL, B., BA, C., & BC, K. (2011). Goodman & Gilman's The Pharmacological Basis Of Therapeutics (12nd ed.). New York: McGraw-Hill.

M., K., F., L., & A., V. (2009). The BioCreative II.5 challenge overview. Paper presented at the Proceedings of the BioCreative II. 5 Workshop 2009 on Digital Annotations.

Magro, L., Moretti, U., & Leone, R. (2012). Epidemiology and characteristics of adverse drug reactions caused by drug-drug interactions. Expert Opin Drug Saf, 11(1), 83-94. doi:10.1517/14740338.2012.631910

McDonald, R., & Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics, 6, S6. doi:Artn S6 10.1186/1471-2105-6-S1-S6

Medicines Complete. Retrieved from https://www.medicinescomplete.com/about/

Merle, L., Laroche, M. L., Dantoine, T., & Charmes, J. P. (2005). Predicting and preventing adverse drug reactions in the very old. Drugs Aging, 22(5), 375-392.

MeSH. Retrieved from http://www.nlm.nih.gov/mesh/meshhome.html

Muller, H. M., Kenny, E. E., & Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol, 2(11), e309. doi:10.1371/journal.pbio.0020309

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26. doi:doi: 10.1075/li.30.1.03nad

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. J Am Med Inform Assoc, 18(5), 544-551. doi:10.1136/amiajnl-2011-000464

Neves, M., & Leser, U. (2014). A survey on annotation tools for the biomedical literature. Brief Bioinform, 15(2), 327-340. doi:10.1093/bib/bbs084

Nisha, R., Bhuiya, F., Xu, J.. . (2010). National Hospital Ambulatory Medical Care Survey: 2007 Emergency Department Summary. National Health Statistics Reports, 26, 1-32.

Niska, R., Bhuiya, F., & Xu, J. (2010). National Hospital Ambulatory Medical Care Survey: 2007 emergency department summary. Natl Health Stat Report(26), 1-31.

Nobata, C., Dobson, P. D., Iqbal, S. A., Mendes, P., Tsujii, J., Kell, D. B., & Ananiadou, S. (2011). Mining metabolites: extracting the yeast metabolome from the literature. Metabolomics, 7(1), 94-101. doi:10.1007/s11306-010-0251-6

Nordstrom, A., O'Maille, G., Qin, C., & Siuzdak, G. (2006). Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. Anal Chem, 78(10), 3289-3295. doi:10.1021/ac060245f

Obach, R. S. (2013). Pharmacologically active drug metabolites: impact on drug discovery and pharmacotherapy. Pharmacol Rev, 65(2), 578-640. doi:10.1124/pr.111.005439

Pang, K. S., Rodrigues, A. D., & Peter, R. M. (2010). Enzyme- and Transporter-Based Drug-Drug Interactions. New York: Springer.

Percha, B., & Altman, R. B. (2015). Learning the Structure of Biomedical Relationships from Unstructured Text. PLoS Comput Biol, 11(7), e1004216. doi:10.1371/journal.pcbi.1004216

Percha, B., Garten, Y., & Altman, R. B. (2012). Discovery and explanation of drug-drug interactions via text mining. Pac Symp Biocomput, 410-421.

Percha, B., Garten, Y., and Altman, R.B. (2012). Discovery and explanation of drug-drug interactions via text mining. Pacific Symp. Biocomput., 410-421.

Physicians' Desk Reference. Retrieved from http://www.pdr.net/

Prueksaritanont, T., Chu, X., Gibson, C., Cui, D., Yee, K. L., Ballard, J., . . . Hochman, J. (2013). Drug-drug interaction studies: regulatory guidance and an industry perspective. Aaps j, 15(3), 629-645. doi:10.1208/s12248-013-9470-x

Pyysalo, S., Airola, A., Heimonen, J., Bjorne, J., Ginter, F., & Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. BMC Bioinformatics, 9 Suppl 3, S6. doi:10.1186/1471-2105-9-S3-S6

Qian, L., & Zhou, G. (2012). Tree kernel-based protein-protein interaction extraction from biomedical literature. J Biomed Inform, 45(3), 535-543. doi:10.1016/j.jbi.2012.02.004

Quinlan, J. R. (1993). C4.5: programs for machine learning: Morgan Kaufmann Publishers Inc.

Quinney, S. K., Zhang, X., Lucksiri, A., Gorski, J. C., Li, L., & Hall, S. D. (2010). Physiologically based pharmacokinetic model of mechanism-based inhibition of CYP3A by clarithromycin. Drug Metab Dispos, 38(2), 241-248. doi:10.1124/dmd.109.028746

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., & Jimeno, A. (2008). Text processing through Web services: calling Whatizit. Bioinformatics, 24(2), 296-298. doi:10.1093/bioinformatics/btm557

Rocktaschel, T., Weidlich, M., & Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), 1633-1640. doi:10.1093/bioinformatics/bts183

Rogers, F. B. (1963). Medical subject headings. Bull Med Libr Assoc, 51, 114-116.

Rostami-Hodjegan A, T. G. (2004). In silico simulations to assess the in vivo consequences of in vitro metabolic drug–drug interactions. Drug Disc Today Technol., 1, 441-448.

Rostami-Hodjegan, A., & Tucker, G. (2004). In silico simulations to assess the in vivo consequences of in vitro metabolic drug–drug interactions. Drug Discovery Today: Technologies, 1(4), 441-448.

Rowland, M., & Tozer, T. N. (1995). Clinical Pharmacokinetics: Concepts and Applications. London: Lippincott Williams & Wilkins.

Rowland, M., & Tozer, T. N. (1995). Clinical Pharmacokinetics: Concepts and Applications. London: Lippincott Williams & Wilkins.

Rowland, M., Tozer, T. N., & Rowland, M. (2011). Clinical pharmacokinetics and pharmacodynamics : concepts and applications. Philadelphia: Wolters Kluwer Health/Lippincott William & Wilkins.

Rubin, D. L., Noy, N. F., & Musen, M. A. (2007). Protege: a tool for managing and using terminology in radiology applications. J Digit Imaging, 20 Suppl 1, 34-46. doi:10.1007/s10278-007-9065-0

Saier, M. H., Jr., Reddy, V. S., Tsu, B. V., Ahmed, M. S., Li, C., & Moreno-Hagelsieb, G. (2016). The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res, 44(D1), D372-379. doi:10.1093/nar/gkv1103

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc, 17(5), 507-513. doi:10.1136/jamia.2009.001560

Segel, I. H. (1975). Enzyme kinetics: behavior and analysis of rapid equilibrium and steady state enzyme systems. New York: John Wiley & Sons, Inc.

Segura-Bedmar I, C. M., de Pablo-Sánchez C., Martínez P,. (2010). Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. BMC Bioinformatics, 11(suppl 2), S1.

Segura-Bedmar, I., Crespo, M., de Pablo-Sanchez, C., & Martinez, P. (2010). Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. BMC Bioinformatics, 11 Suppl 2(suppl 2), S1. doi:10.1186/1471-2105-11-S2-S1

Segura-Bedmar, I., Crespo, M., de Pablo-Sanchez C., and Martinez, P. (2011). Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. BMC Bioinformatics, 11(suppl 2), S1.

Segura-Bedmar, I., Martinez, P., & de Pablo-Sanchez, C. (2011a). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. BMC Bioinformatics, 12 Suppl 2(suppl 2), S1. doi:10.1186/1471-2105-12-S2-S1

Segura-Bedmar, I., Martinez, P., & de Pablo-Sanchez, C. (2011b). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. BMC Bioinformatics, 12 Suppl 2, S1. doi:10.1186/1471-2105-12-s2-s1

Segura-Bedmar, I., Martinez, P., & de Pablo-Sanchez, C. (2011c). Using a shallow linguistic kernel for drug-drug interaction extraction. J Biomed Inform, 44(5), 789-804. doi:10.1016/j.jbi.2011.04.005

Segura-Bedmar, I., Martinez, P., & de Pablo-Sanchez, C. (2011d). Using a shallow linguistic kernel for drug-drug interaction extraction. J Biomed Inform, 44(5), 789-804. doi:10.1016/j.jbi.2011.04.005

Segura-Bedmar, I., Martinez P Fau - de Pablo-Sanchez, C., & de Pablo-Sanchez, C. (2011). A linguistic rule-based approach to extract drug-drug interactions from. BMC Bioinformatics, 12 Suppl 2, S1 LID - 10.1186/1471-2105-1112-S1182-S1181 [doi].

Segura-Bedmar, I., Martinez P Fau - Segura-Bedmar, M., & Segura-Bedmar, M. (2008). Drug name recognition and classification in biomedical texts. A case study. Drug Discov Today, 13(17-18), 816-823 LID - 810.1016/j.drudis.2008.1006.1001 [doi].

Segura-Bedmar, I., Martinez, P., & Herrero-Zazo, M. (2014). Lessons learnt from the DDIExtraction-2013 Shared Task. J Biomed Inform, 51, 152-164. doi:10.1016/j.jbi.2014.05.007

Segura-Bedmar, I., Martınez, P., & Sánchez-Cisneros, D. (2011a). The 1st DDIExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts Paper presented at the Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011, Spain.

Segura-Bedmar, I., Martınez, P., & Sánchez-Cisneros, D. (2011b). The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. Challenge Task on Drug-Drug Interaction Extraction, 2011, 1-9.

Segura-Bedmar, I., Martínez, P., de Pablo-Sánchez, C. (2011). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. BMC Bioinformatics, 12(suppl 2), S1.

Segura-Bedmar, I., Martínez, P., de Pablo-Sánchez, C. (2011). Using a shallow linguistic kernel for drug-drug interaction extraction. J Biomed Inform, 44(5), 789-804.

Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics, 21(14), 3191-3192. doi:10.1093/bioinformatics/bti475

Sevrioukova, I. F., & Poulos, T. L. (2017). Structural basis for regiospecific midazolam oxidation by human cytochrome P450 3A4. Proc Natl Acad Sci U S A, 114(3), 486-491. doi:10.1073/pnas.1616198114

Sim, S. C., & Ingelman-Sundberg, M. (2010). The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. Hum Genomics, 4(4), 278-281.

Stearns, V., Johnson, M. D., Rae, J. M., Morocho, A., Novielli, A., Bhargava, P., . . . Flockhart, D. A. (2003). Active tamoxifen metabolite plasma concentrations after coadministration of tamoxifen and the selective serotonin reuptake inhibitor paroxetine. J Natl Cancer Inst, 95(23), 1758-1764.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. i. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. Paper presented at the In Proceedings of the Demonstrations Session at EACL.

Tari, L., Anwar, S., Liang, S., Cai, J., & Baral, C. (2010). Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics, 26(18), i547-553. doi:10.1093/bioinformatics/btq382

Tari, L., Anwar, S., Liang, S., Cai, J., Baral, C. (2010). Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics, 26(18), i547-553.

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., & Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. PLoS Comput Biol, 6, e1000837. doi:10.1371/journal.pcbi.1000837

Tsuruoka, Y., & Tsujii, J. (2004). Improving the performance of dictionary-based approaches in protein name recognition. J Biomed Inform, 37(6), 461-470. doi:10.1016/j.jbi.2004.08.003

Use, C. f. M. P. f. H. (2012). Guideline on the Investigation of Drug Interactions. London: European Medicines Agency.

Usie, A., Alves, R., Solsona, F., Vazquez, M., & Valencia, A. (2014). CheNER: chemical named entity recognizer. Bioinformatics, 30(7), 1039-1040. doi:10.1093/bioinformatics/btt639

Vazquez, M., Krallinger, M., Leitner, F., & Valencia, A. (2011). Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. Mol Inform, 30(6-7), 506-519. doi:10.1002/minf.201100005

Wang, Z., Kim, S., Quinney, S. K., Guo, Y., Hall, S. D., Rocha, L. M., & Li, L. (2009). Literature mining on pharmacokinetics numerical data: a feasibility study. J Biomed Inform, 42(4), 726-735. doi:10.1016/j.jbi.2009.03.010

Wang, Z., Kim, S., Quinney, S. K., Zhou, J., & Li, L. (2010). Non-compartment model to compartment model pharmacokinetics transformation meta-analysis--a multivariate nonlinear mixed model. BMC Syst Biol, 4 Suppl 1, S8. doi:10.1186/1752-0509-4-S1-S8

Wienkers, L. C., & Heath, T. G. (2005). Predicting in vivo drug interactions from in vitro drug discovery data. Nat Rev Drug Discov, 4(10), 825-833. doi:10.1038/nrd1851

Wilbur, W. J., Rzhetsky, A., & Shatkay, H. (2006). New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics, 7, 356. doi:10.1186/1471-2105-7-356

Wilke, R. A., Ramsey, L. B., Johnson, S. G., Maxwell, W. D., McLeod, H. L., Voora, D., . . . Clinical Pharmacogenomics Implementation, C. (2012). The clinical pharmacogenomics implementation consortium: CPIC guideline for SLCO1B1 and simvastatin-induced myopathy. Clin Pharmacol Ther, 92(1), 112-117. doi:10.1038/clpt.2012.57

Williams, J. A., Ring, B. J., Cantrell, V. E., Jones, D. R., Eckstein, J., Ruterbories, K., . . . Wrighton, S. A. (2002). Comparative metabolic capabilities of CYP3A4, CYP3A5, and CYP3A7. Drug Metab Dispos, 30(8), 883-891.

Wishart, D. S. (2007). Current progress in computational metabolomics. Brief Bioinform, 8(5), 279-293. doi:10.1093/bib/bbm030

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., . . . Scalbert, A. (2013). HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res, 41(Database issue), D801-807. doi:10.1093/nar/gks1065

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann.

Wong, C. M., Ko, Y., & Chan, A. (2008). Clinically significant drug-drug interactions between oral anticancer agents and nonanticancer agents: profiling and comparison of two drug compendia. Ann Pharmacother, 42(12), 1737-1748. doi:10.1345/aph.1L255

Wu, H.-Y., Chiang, C.-W., & Li, L. (2014). Text Mining for Drug–Drug Interaction. In V. D. Kumar & H. J. Tipney (Eds.), Biomedical Literature Mining (Vol. 1159, pp. 47-75): Springer New York.

Wu, H. Y., Karnik S Fau - Subhadarshini, A., Subhadarshini A Fau - Wang, Z., Wang Z Fau - Philips, S., Philips S Fau - Han, X., Han X Fau - Chiang, C., . . . Li, L. (2013). An integrated pharmacokinetics ontology and corpus for text mining. BMC Bioinformatics, 14, 35 LID - 10.1186/1471-2105-1114-1135 [doi].

Wu, H. Y., Karnik, S., Subhadarshini, A., Wang, Z., Philips, S., Han, X., . . . Li, L. (2013). An integrated pharmacokinetics ontology and corpus for text mining. BMC Bioinformatics, 14, 35. doi:10.1186/1471-2105-14-35

Yu, M., Kim, S., Wang, Z., Hall, S., & Li, L. (2008). A Bayesian meta-analysis on published sample mean and variance pharmacokinetic data with application to drug-drug interaction prediction. J Biopharm Stat, 18(6), 1063-1083. doi:10.1080/10543400802369004

Zhang, L., Reynolds, K. S., Zhao, P., Huang, S.M. (2010). Drug interactions evaluation: An integrated part of risk assessment of therapeutics. Toxicology and Applied Pharmacology, 243, 134-145.

Zhang, L., Zhang, Y., Zhao, P., Huang, S.M. (2009). Predicting Drug-Drug Interactions: An FDA Predictive. Aaps j, 11(2), 300-306.

Zhou, D., & He, Y. (2008). Extracting interactions between proteins from the literature. J Biomed Inform, 41(2), 393-407. doi:10.1016/j.jbi.2007.11.008

Zhou, G., Zhang, J., Su, J., Shen, D., & Tan, C. (2004). Recognizing names in biomedical texts: a machine learning approach. Bioinformatics, 20(7), 1178-1190. doi:10.1093/bioinformatics/bth060

Zhou, J., Qin, Z., Sara, Q. K., Kim, S., Wang, Z., Hall, S. D., & Li, L. (2009). Drug-drug interaction prediction assessment. J Biopharm Stat, 19(4), 641-657. doi:10.1080/10543400902964084

Zhou, J., Qin, Z., Yu, M., Lucksiri, A., Wang, Z., Kim, S., . . . Li, L. (2009). A new probabilistic rule for drug-drug interaction prediction. Journal of Pharmacokinetics and Pharmacodynamics, 36, 1-18.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., & Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. Brief Bioinform, 8(5), 358-375. doi:10.1093/bib/bbm045

Curriculum Vitae

Heng-Yi Wu

Research Interests:

- Text/Data-Mining for clinical data and biomedical literature
- Clinical Pharmacokinetics and Pharmacodynamics for Drug-Drug Interaction
- Network Modeling for Genetics Regulatory Network
- Genome-wide analysis and Next generation sequence
- Signal and image Processing

Education:

- Ph. D, School of Informatics & computing, Aug 2014 –Oct 2017

  Health Informatics

  Indiana University-Purdue University Indianapolis, USA
- M.S., School of Informatics & computing, Bioinformatics Sep 2011 –May 2014

  Indiana University-Purdue University Indianapolis, USA
- M.S., Electrical & Computer Engineering, Sep 2008 –Dec 2010

  The Ohio State University, USA
- M.S., Electrical Engineering, Sep 2005 – Jan 2007

  National Chung Hsing University, Taiwan
- B.S., Electrical Engineering, Sep 2000 – Jun 2005

  National Chung Hsing University, Taiwan

Research and Training Experience

- Research Assistant – Center for Computational Biology and Bioinformatics: Sep 2011 – Present
- Analytic Programmer – Department of Medical & Molecular Genetics, Indian University: School of Medicine: Dec 2010 – Sep 2011
- Research Assistant – Department of Biomedical Informatics, The Ohio State University: Sep 2009-Dec 2010

Computer Languages and Tools:

- Languages: C, R, Perl, Python, SQL, SAS, Matlab, HTML, and UNIX.

Conference Attended

Oral Presentation

1. Heng-Yi Wu, Using Machine learning algorithms to identify genes essential for cell survival, ICIBM 2017 in Houston
2. Heng-Yi Wu, Translational drug interaction study using text mining technologies, 6th Annual Indiana CTSI Symposium on Disease and Therapeutic Response Modeling, at Indiana University School of Medicine (Nov 9-10, 2016)
3. Yaoyun Zhang*, Heng-Yi Wu*, Xu Hua, and Lang Li, Leveraging Syntactic and Semantic Graph Kernels to Extract PK Drug Drug Interactions from Biomedical Literature, ICIBM2015
4. Heng-Yi Wu, Yu Wang, Zheng, P., Jiang, G., Yunlong Liu, Huang, T.H.M., Nephew, K.P.,

Lang Li, "An ERα/modulator regulatory network in the breast cancer cells" *in Proc. IEEE 2011 Workshop on Genomic Signal Processing and Statistics* (GENSIPS).

Poster Presentation

1. Heng-Yi Wu, Deshun Lu, Mustafa Hyder, and Lang Li, Name Entity Recognition for Drug Metabolite by Using Text-Mining Technology, PSB2016
2. Heng-Yi Wu, Shijun Zhang, Luis M. Rocha, Hagit Shatkay, Desta Zeruesenay, Sara K. Quinney, Lang Li, Translational Drug Interaction Evidence Gap Discovery Using Text Mining, ASCPT 2017, AMIA 2017, and Regenstrief conference 2016

Publications:

Journal paper

Under Preparation or Submitted

1. Pengyue Zhang*, Heng-Yi Wu*, Chien-Wei Chiang, Lei Wang, Samar Binkheder, Xueying Wang, Sara K. Quinney, and Lang Li, Translational Biomedical Informatics and Pharmacometrics Approaches in the Drug Interaction Research, CPT: Pharmacometrics & Systems Pharmacology (Under Preparation)
2. Pengyue Zhang, Meng Li, Wang Lei, Yang Xiang, Lijun Cheng, Weixing Feng, Heng-Yi Wu, Donglin Zeng, Lang Li, A Three-Component Mixture Model Based Adverse Drug Event Signal Detection for the Adverse Event Reporting System, (Submitted to CPT: Pharmacometrics & Systems Pharmacology IF=3.24)

3. Xueying Wang, Pengyue Zhang, Chien-Wei Chiang, Heng-Yi Wu, Li Shen, Xia Ning, Donglin Zeng, Lei Wang, Sara K. Quinney, Weixing Feng, Lang Li, Mixture Drug-Count Response Model for the High Dimensional Drug Combinatory Effect on Myopathy (Submitted to Biometrics IF= 1.827)
4. Heng-Yi Wu, Shijun Lee, Luis M. Rocha; Hagit Shatkay; Lang Li, Biomedical text annotation definitions, guidelines and corpus construction for drug interaction, (Under Preparation)
5. Heng-Yi Wu, Shijun Zhang, Desta Zeruesenay, Sara K. Quinney, Lang Li, Translational Drug Interaction Evidence Gap Discovery Using Text Mining, (Under Preparation)
6. Heng-Yi Wu, Deshun Lu, Mustafa Hyder, and Lang Li, Name Entity Recognition for Drug Metabolite by using text mining method, (Under Preparation)

Published

1. Santosh Philips, Heng-Yi Wu, Lang Li, Using Machine learning algorithms to identify genes essential for cell survival, *BMC Genomics*, 2017 (IF=3.87)
2. Yaoyun Zhang*, Heng-Yi Wu*, Xu Hua, and Lang Li, Leveraging Syntactic and Semantic Graph Kernels to Extract PK Drug Drug Interactions from Biomedical Literature, BMC Systems Biology 2016 (IF=3.24)
3. Yaoyun Zhang*, Heng-Yi Wu*, Jingchen Du, Jingqi Wang, Cui Tao, Lang Li , Xu Hua, Extracting Drug-Enzyme Relation from Literature as Evidence for Drug Drug Interaction, Journal Of Biomedical Semantics 2016 Mar 7 (IF=1.62)
4. Lei Du, Arindom Chakraborty, Chien-Wei Chiang, Lijun Cheng, Sara K. Quinney, Heng-

Yi Wu, Pengyue Zhang, Lang Li, and Li Shen, Graphic Mining of High-Order Drug Interactions and Their Directional Effects on Myopathy Using Electronic Medical Records, *Pharmacometrics & Systems Pharmacology*, 2015 (IF=3.24)

5. Pengyue Zhang, Lei Du, Lei Wang, Lijun Cheng, Chien-Wei Chiang, Heng-Yi Wu, Sara K. Quinney, Li Shen, and Lang Li, A mixture Does-Response Model for Identifying High-Dimensional Drug Interaction Effects on Myopathy Using Electronic Medical Record Databases, *Pharmacometrics & Systems Pharmacology*, 2015 (IF=3.24)
6. Lei Wang, ChienWei Chiang, Hong Liang, Heng-Yi Wu, Weixing Feng, Sara K. Quinney, Jin Li and Lang Li, How to Choose In vitro Systems to Predict In Vivo Drug Clearance: A System Pharmacology Perspective, *BioMed Research Internatioal*, 2015 (IF=1.58)
7. Artemy Kolchinsky, Anália Lourenço, Heng-Yi Wu, Lang Li, Luis M. Rocha, Extraction of Pharmacokinetic Evidence of Drug-drug Interactions from the Literature, *PLOS One*, 2015 (IF=3.23)
8. Heng-Yi Wu, Shreyas Karnik, Abhinita Subhadarshini, Zhiping Wang, Santosh Philips, Xu Han, Chienwei Chiang, Lei Liu, Malaz Boustani, Luis M Rocha, Sara K Quinney, David Flockhart, Lang Li, An integrated pharmacokinetics ontology and corpus for text mining. *BMC Bioinformatics*, 2013. 14:p.14-35 (IF=2.44)
9. Heng-Yi Wu, Zheng, P.Jiang, G., et al.: A modulator based regulatory network for ERalpha signaling pathway. *BMC Genomics*, 2012. 13 Suppl 6: p. S6. (IF=3.87)

Book Chapter (1)

- Heng-Yi Wu, Chien-Wei Chiang and Lang Li. “Text Mining for Drug-Drug Interaction”, *Methods in molecular biology* (Clifton, N.J.) 01/2014, 1159:47-75

Conference paper

1. Heng-Yi Wu, Shijun Zhang, Desta Zeruesenay, Sara K. Quinney, Lang Li, Translational Drug Interaction Evidence Gap Discovery Using Text Mining, ASCPT 2017
2. Santosh Philis*, Heng-Yi Wu*, and Lang Li, Using Machine learning algorithms to identify genes essential for cell survival, ICIBM2016
3. Yaoyun Zhang*, Heng-Yi Wu*, Xu Hua, and Lang Li, Leveraging Syntactic and Semantic Graph Kernels to Extract PK Drug Drug Interactions from Biomedical Literature, ICIBM2015
4. Heng-Yi Wu, Yu Wang, Zheng, P., Jiang, G., Yunlong Liu, Huang, T.H.M., Nephew, K.P., Lang Li, "An ERα/modulator regulatory network in the breast cancer cells" *in Proc. IEEE 2011 Workshop on Genomic Signal Processing and Statistics* (GENSIPS).
5. Heng-Yi Wu, Jie Zhang and Kun Huang, "Peak Detection on ChIP-Seq data using Wavelet transformation" *in Proc. IEEE BIBM 2010 Workshop on data-mining of Next Generation Sequencing Data*.
6. Shien-Tang Chiu, Guo-Shiang Lin, Heng-Yi Wu and Min- Kuan Chang, "An Effective Shot Boundary Detection Algorithm for Movies and Sports," *in Proc. IEEE ICICIC2008*
7. Shi-Yong Lee, Heng-Yi Wu, and Min-Kuan Chang, "New Lifetime-aware Bit and Power Allocation in OFDM Systems," *in Proc. IEEE APWCS2007*
8. Heng-Yi Wu, Min-Kuan Chang, and Chia-Chung Chang, "The Switch-based Subcarrier Allocation Policies in Multi-service OFDM Systems," *in Proc. IEEE VTC06 Spring*

9. Heng-Yi Wu, and Min-Kuan Chang, " A Novel subchannel Allocation in Multi-service OFDM Systems based on Weighted Round Robin" *in Proc. IEEE APWCS2006*

Note: * means Co-First author