BIOMEDICAL LITERATURE MINING FOR PHARMACOKINETICS NUMERICAL PARAMETER COLLECTION

Zhiping Wang

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 and Computing,

Indiana University

December, 2012

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

Doctoral Committee

Luis Rocha, Ph.D.

Lang Li, Ph.D.

Michael Gasser, Ph.D.

Haixu Tang, Ph.D.

December 18th, 2012

Acknowledgements

I would like to thank the members of my committee, Dr. Luis Rocha, Dr. Lang Li, Dr. Michael Gasser and Dr. Haixu Tang for their support during my PhD and the preparation of this dissertation. In particular, I am indebted to Dr. Luis Rocha who guided me through PhD study during the past years, and Dr. Lang Li for his support on my thesis project. I am also grateful to Dr. Sun Kim, who helped me on the study of Bioinformatics as a minor degree.

Most importantly, none of this would have been feasible or worthy without the patient and love of my family, to whom I dedicated this dissertation. Without the encouragements and prayers of my dear wife Hailan and two lovely sons, Samuel and Simon, I can't imagine to ever finishing this manuscript.

Zhiping Wang

ABSTRACT

BIOMEDICAL LITERATURE MINING FOR PHARMACOKINETICS NUMERICAL PARAMETER COLLECTION

Model-based drug studies have been developing very fast recently. They require high quality pharmacokinetics (PK) parameter numerical data. However, most parameter measurements are still buried in the scientific literature. Traditional manual data extraction is too expensive to handle the exponentially growing number of publications. This thesis focuses on the application of text mining (TM) and machine learning (ML) for drug pharmacokinetics parameter data collection from the published literature.

First, we explore the feasibility of TM on the extraction of drug PK parameter data from PubMed abstracts. Our method achieves higher precision and obtains rich information content. For the test drug *Midazolam*, it extracts 10 times more PK clearance data than the manually constructed commercial Drug Interaction Database (DiDB). Similar performance is obtained on additional test drugs.

Following the success of TM on abstracts; we extended the methodology to full text articles and developed a literature mining pipeline for PK parameter data extraction. It combines machine learning, automatic information processing, and manual curation. It compromises four main components: (1) information retrieval, which applies both ontology-based name entity recognition (NER) and ML methods to classify PubMed search results; (2) article downloading of full PDF articles through PubMed external links; (3) information extraction of PK data from

both tables and free text of articles; and (4) transformation and storage of mined information, so that it can be reachable in a drug-modeling-friendly manner. This literature mining pipeline and methodology is the first working approach to extract numerical data from full text articles, capable of processing both plain text and tabular data.

The specific contributions of this thesis include:

- A new PK ontology for entity template construction
- Comparison of NLP and machine learning algorithms for PK information retrieval
- Tabular data extraction
- PK information extraction from full text literature
- A complete pipeline of numerical data extraction from both abstracts and full-text articles for pharmacokinetics

Contents

List of Tables

LIST OF FIGURES

Chapter 1.

INTRODUCTION AND REVIEW

1.1 Background

In recent decades, a new drug requires an average of 15 years to be released, costing almost a billion dollars in research and development [1]. In 2004, the FDA released a report entitled: “Innovation or Stagnation, Challenge and Opportunity on the Critical Path to New Medical Products” [2]. Among its six general topic areas, three of them emphasized the importance of computational modeling and bioinformatics in biomarker development and streamlining clinical trials. In multiple follow-up papers, clinical researchers, experimental biologists, computational biologists, and biostatisticians from both academia and industry started to discuss the challenges and opportunities of the pharmacokinetic-pharmacodynamic (PK/PD) model based approach in drug development [3-6]. PK/PD modeling simulates the pharmaceutical effects of a drug using mathematical equations by integrating both of its pharmacokinetical and pharmacodynamical characteristics. Today, drug discovery is considered impossible without sophisticated modeling and computation [7], which can substantially reduce the cost of drug development by constructing effective simulations, identifying therapeutic strategies, and making novel predictions.

This thesis focuses on text mining techniques to assist PK modeling, which is centered on the absorption and distribution mechanism of an administered drug, its metabolism, excretion and duration of effects. To fulfill the PK modeling potential in drug development, there is an enormous need for databases of PK parameters. For example, to specify the first human dose of a new compound, one needs its in-vitro and in-vivo PK parameters based on animal studies [3, 5]. However, current pharmacology databases provide little PK data. DiDB [8] is an on-going project which manually accumulates published PK data for each drug. DrugBank [9] is a comprehensive pharmacology database which has rich annotations on the structure, mechanism, pathway, and targets of drugs, but offers very sparse PK data. The other pharmacology resources, e.g. PharmGKB [10], dailyMed [11], PubPK [12]and PK/DB [13], have varying emphasis on drug properties but neither has complete drug PK parameter data available. One common feature of the above databases is the reliance on manual data curation of PubMed search results. The widely used online literature search service, PubMed, contains about 20 million abstracts from MEDLINE and additional life sciences journals, which has been growing exponentially [14]. Therefore, it is almost impossible for biomedical researchers to discover and keep track of all relevant publications in their own discipline manually. In addition to the lack of scalability for the publication growth, manual curation also suffers from inter-annotator disagreement [15] (though mining algorithms also don’t agree with each other, it is convenient to include validation and meta-analysis mechanisms in mining to minimize this effect as we did in this thesis). The complexity of information from PK studies makes the manual curation even harder when various drug doses, administration routes, patients, sample collection intervals, and the like are target information for collection. Moreover, the drug PK information that needs to be collected depends heavily on the mechanics of PK models and their simulations, which is in turn driven by the science. Thus the required literature information becomes a moving target, which adds more difficulty for manual as well as automated knowledge integration.

Though currently, most PK databases still rely on manual curation to collect accurate data, this is inefficient to keep up with the exponentially growing scientific publications, and to handle the large amount of varying information needs of users. One study argued that it will take years, or even decades, for biomedical database construction, if we just rely on manual curation [16]. Meanwhile, computer-aided curation has been proven to be effective in maintaining the Medline database [17, 18]. Therefore we investigated *literature and text mining* (TM) as an alternative solution to manual curation for PK parameter data collection which targets to efficiently handle large scale of information, automatically extract PK data with good quality, and timely data update, analysis and storage.

1.2 Review of Literature Mining

1.2.1 Information Retrieval / Text Classification

It has not been very long that TM was applied in the biomedical domain but it has already shown enormous progress [19]. PubMed possesses *information retrieval* (IR) capabilities to find articles relevant to a certain topical query. It is such an important tool for biomedical researchers that many PubMed based IR tools have been developed to make the search itself easy and precise. Rooted in Pubmed, alternative search engines have been implemented such as Relemed [20] which enables sentence-level searches, PubMed Reader , and HubMed [22]. Some tools allow PubMed searches to be filtered according to the similarity to given input articles such as eTBlast [23] and MScanner [24], which makes the query more personalized.

The most obvious IR approach is to look for keywords that characterize the articles, for instance, using the vector model [25]. PubMed is an example of this type of similarity query. More complicated retrieval systems were also developed to improve performance, which originate from different algorithms such as latent semantic indexing (LSI) [26], statistical models [27], combination of statistical and LSI models [28], and other machine learning methods such as support vector machines [29], and hidden markov models [30]. The performance of most of these algorithms depends on appropriate feature selection (single terms, term combinations, or strings) for the literature of interest. Common feature selection methods include *term frequency–inverse document frequency* (tf-idf) [31], *principle component analysis* (PCA) [32] and *information gain* [33]. Feature selection for the biomedical domain, from both linguistic and application points of view, is still under improvement [34].

1.2.2 Entity Recognition

Another fast growing area in biomedical TM is entity recognition, the process of automatic recognizing key terms and concepts of a certain area such as gene/protein names. There are several challenges in dealing with biomedical entities. One is the polysemy (multiple meanings of a word). Medical terms or gene names have been found to carry greater ambiguity than general English words, up to 14.20% [35]. For example, 2.4% of gene names in FlyBase are common English words which make dictionary-based gene name recognition uncertain. Another problem is synonymy [36], e.g. P53, TP53 and TRP53 all refer to the same gene. To deal with the variability of biological nomenclatures, the HUGO Gene nomenclature committee (HGNC) was created to assign every gene a unique symbol, however, not all genes have been covered and it will be difficult to cover all gene names used in the past and those to be used in future. Several word-tokenizers were developed to account for entity name variants [37, 38] with varying degrees of success.

The biomedical literature also relies heavily on the use of abbreviations [39] or acronyms, for example, *Caenorhabditis elegans* is commonly abbreviated to *C. elegans*; however, 49 other species have a name that can be abbreviated to this short form. Available tools to deal with this problem include ADAM [40], Abbreviation Sever [41], and AcroMine [42]. NCBI Taxonomy [43, 44] also contributes by assigning an accession number for phylogenetic and taxonomic terms from a variety of sources. Another problem in entity recognition is anaphora, an instance of an expression referring to another, which can be processed by linguistic rules and semantic analysis [45].

Studies have shown that biomedical researchers tend not to adhere to naming standards which makes the dictionary-based entity recognition less effective [46]. Existing terminology and ontologies are still widely used in TM (e.g. LINNAEUS [47]) as a quick reference of domain knowledge. Such ontologies include UMLS , GO [49], and BioThesaurus [50]. The ontologies can also be combined with general English synonym databases, e.g. WordNet [51], for better entity recognition [52]. Besides these existing resources, some terminology development systems, e.g. TerMine [53, 54], were also developed for term recognition using linguistic and statistical analysis. However, the employment of machine learning methods for biomedical ontology construction did not achieve a very good performance, with f-score below 50% [55, 56], because of the complex architecture, relevant term and relationship selection in an ontology. As a result, the combination of human-engineered ontologies and automatically deduced ontologies may be a better choice, because the results of manual curation provide validation data , while machine learning results provide guidance for curations [57].

Most of the machine learning algorithms applied in information retrieval can also be ported to entity recognition [58]. In addition, Conditional Random Fields [59] also gained popularity in ER. However there was also disagreement regarding the question of the relative importance of the selection of features and specific ER algorithm [60]. In spite of all the challenges, entity recognition in TM has actually reached a very high performance in some problems. In BioCreative II, the gene name recognition problem reached an f-score up to 87% [58]. BioCreative III extended the identification of biological entities to full text articles. This time, cross-species data was used for gene name normalization which increased ER complexity, thus the performance decreased accordingly [61]. However, The BioCreative competition has represented the trend of ER being applied in real life problems.

1.2.3 Information Extraction

The success on entity recognition has set a foundation for text mining to go further, leading to *information extraction*. Similar to discovering related articles in IR, Information extraction is concerned with the discovery of relevant facts. Until now, most information extraction research focused on entity relationship extraction such as protein-protein interactions (PPIs) [62], but there also has been many interests in event detection [63-65].

The success of IE depends on a better understanding and automatic manipulation of biomedical language. Techniques from natural language processing (NLP) can be used to decode information in human language. NLP occurs at multiple levels: words, syntax, semantics and even pragmatics. For example, a solution to the polysemy problem requires a machine to discern the semantic features of the text at various levels. In the tokenization process of NLP, the morphological analysis of words is usually based on stemming [66, 67] because the meaning of a word is carried predominantly by the stem or root. Popular NLP software include: Stanford NLP (http://nlp.stanford.edu/), GATE (http://gate.ac.uk), NLTK (http://nltk.sourceforge.net), MALLET (http://mallet.cs.umass.edu/). A comprehensive summary of the available NLP software is available in [68]. As biomedical literature shows apparent difference compared with general English texts, specific NLP tools working for the biomedical domain have been developed, e.g. MedPost [69], JULIE [38], dTagger [70]. In the BioCreative gene normalization task, many teams [71] integrated external resources, e.g. gene mention taggers ABNER [72] or LingPipe [73], into NLP to improve performance.

Several methods are used for information extraction. The simplest way to detect relations is to collect texts or sentences in which the biomedical entities co-occur [74], assuming that if two entities are mentioned together in the literature, they should have some biological interaction relationship. This method usually requires analyzing the context of entity co-occurrence or finding interaction-relevant textual patterns. Rule based methods [75, 76] analyze the semantic and syntactic features of the text to generate extraction patterns and rules [77-79]. Syntactic information can also be converted into sentence parse trees to derive relationships between co-occurred entities [80, 81]. The rule based method works naturally with NLP for semantic and syntactic processing. Expert knowledge can be easily incorporated into rules that are easy to understand, however facts not covered by rules are missed. The rule-based method can also work on negation relationships [82]. Compared with rule construction via manual curation, most machine learning methods train a model based on features extracted from corpora from information classification or extraction [83]. A method can be easily ported to different data sets. This leads to higher scalability but the learned features and models are not usually portable to different problems. Most of the top ranked methods in BioCreative gene mention task used a machine-learning algorithm, similar to methods discussed for IR [33]. Other techniques from complex network theory have also been used successfully for this problem [84-86].

1.2.4 Other Text Mining Areas

In addition to the most popular IR, ER and IE studies, TM is also used in other biomedical relevant areas such as hypothesis discovery, and making novel discoveries based on known facts. The most representative hypothesis discovery study is Swanson's serendipitous discovery of the connection between Raynaud's disease and fish oil [87] which was clinically invalidated three years later [88]. This led to the Swanson's ABC model or so called Swanson Linking, which is a method to generate hypotheses of undiscovered knowledge based on known facts [89-91]. Another example is the study trends in the biological research topics based on publications [92], which showed the emergence of biological domains in time.

1.2.5 Biomedical Text Mining Applications

Many of the methods discussed in different areas of TM have been implemented on various biomedical applications. Term co-occurrence has been used in multiple gene and protein relationship studies [93, 94]. Some systems also provide online access to co-occurrence association analysis, such as PubGene [95], CoPub Mapper [96], and MedEvi [97].

In the BioCreative challenges, several teams [58, 98, 99] made very impressive contributions to entity recognition and information extraction, especially relationship extraction (e.g. PPI), thus concluding that TM can be used to solve practical biomedical IE problems. Nonetheless, the performance of machine learning methods for relationship extraction is still under improvement [100, 101] for specific biomedical fields.

In addition relationship extraction, IE has also been used in other biomedical domains, such as subcellular location information search [102], amino acid mutation [103], nucleotide polymorphisms [104], physicochemical information [105], microRNA-gene association [106], gene-protein network [107], medication information extraction [108], knowledge database construction for cancer studies [109-111], biomedical images [112], diagnostic aids [113], and even Clinical question answering [114] etc.

TM is becoming more attractive to automate biomedical information processing after being used to solve real biological problems specified by researchers as mentioned above or to provide support to expedite the biomedical database curation, e.g. in Flybase [33]. Research has shown that the potential of TM in database curation is very encouraging [115, 116]. For example, the curation of FlyBase records can get 20% faster with the application of an interactive mining tool [117]. Some systems integrate TM with manual annotation, such as Textpresso [118], GOAnnotator [119], and PreBIND [120]. However, such systems were not developed to replace manual curation but to speed up and standardize the curation process. Many think that manual curation will always be necessary [121] for precise database annotation.

A collection of popular biomedical text mining tools is available from a paper presented at BioNLP [122]. In addition to the advances on TM for entity recognition and relationship extraction, BioNLP focused on modeling more complex regulatory pathways [63-65]. This indicates the trend to expand TM from entity-level to system-level. The recent BioNLP conferences were basically a relay of its previous focus, mostly PubMed based biomedical TM. However, in several papers published in the last conference, TM has applied on more extensive biomedical fields including drug studies (e.g. drug-drug interaction detection) and entity recognition for Electronic Medical Records (EMR) [123].

The fast growth of biomedical domains where TM has been applied to indicates its importance. Though machine learning algorithms in TM are highly portable, the domain dependent feature selection and specific external resource reference makes it very difficult to port a mining system from one domain to another. However, It is still possible that some general-purpose functions of the existing mining tools can be applied to novel mining tasks [72] to gain efficiency and superior performance. Furthermore the combination of multiple existing TM systems has proved able to improve performance [58]. Another high-performing system [124] combined four independent IR systems and found that the fusion significantly outperformed individual systems. Finally the integration of data from multiple resources has also been proven to improve the performance of mining [33, 125-127].

1.2.6 Text Mining on Full Text

Most of the mining progress above is mostly based on PubMed abstracts. Although abstracts contain short descriptions that highlight the most relevant aspects of a given article, they only cover a small fraction of the information contained in full-text articles [128]. One statistical study claims that only 30% of curated PPIs can be found in the abstracts rather than the full text [129]. Some end users, e.g. researchers from pharmacology, need more detailed information which is usually presented in the full text, such as specific subject information of a clinical trial. These issues emphasize the need for full-text TM, which is drawing more interests. For example, a PubMed search using key terms “full text mining” in July, 2009 only returned 59 hits, with 7 relevant. However, the same search gave back 369 hits in June, 2012. However, the comparably slower progress was mostly caused by access issues.

One study [130] compared abstracts with articles in three aspects: information distribution, syntax and the performance of mining tools. The results show significant difference in all three aspects, which indicates potential challenges in the transition of text mining from abstracts to full-text. Except for its complexity, full-text based text mining, with integration of additional essential information from tables, figures, and references, is expected to be more valuable in the near future as the number of electronically available full-text documents in open access repositories increases [131]. Though there is still no comprehensive free-access full text corpora, some efforts have appeared, e.g. PubMed Central and Highwire, as full-text article repositories [132].

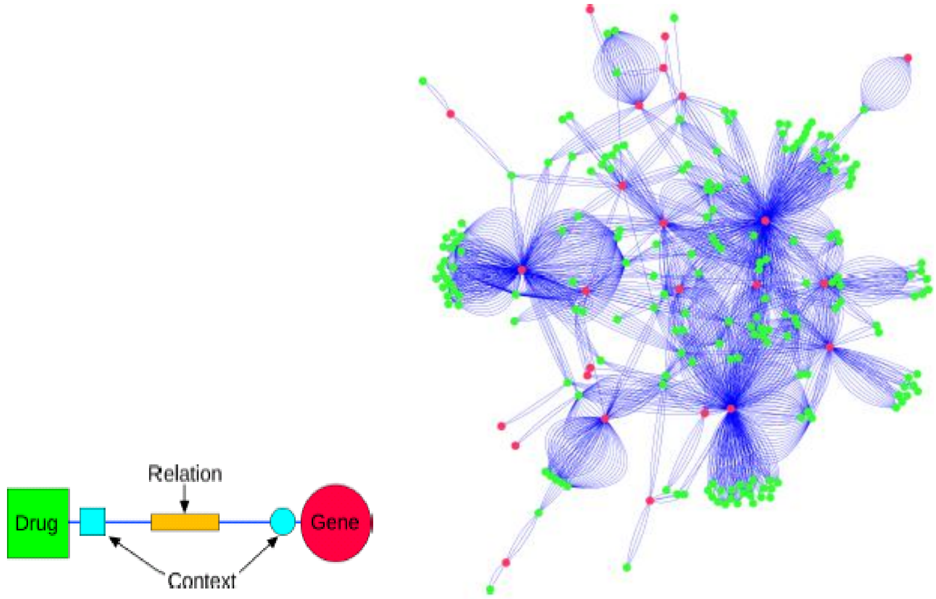
Some preliminary TM studies checked the Information distribution [133, 134] of full text articles. The Open Text Mining Interface [135] made efforts until 2009 to provide a consistent format for text processing. In addition to these efforts, there has been some research on information extraction on full-text, such as BioRAT (information extraction) [136], @note (TM workbench) [137], Pharmspresso (pharmacogenomic IE) [138], KiPar (pathway kinetic parameter extraction) [139], BioText (figure extraction) [131]. These studies work in a specific biological domain and some of them [138, 139] achieve very promising precision and recall performance. BioCreative II.5 [98] was a challenge to identify interaction proteins, protein pairs and protein-protein interaction (PPI) relevant articles based on full-text articles. BioCreative III continued all three tasks from II.5 and tried to make a breakthrough. The last workshop (BioCreative’12 [140]) started to focus on the application of text mining in the curation of real word biomedical databases, which reflects the fast growth and increasing maturity of research in this field.

1.3 Review of Pharmacokinetics-Related Literature Mining

For the pharmaceutical industry, TM systems are a valuable resource as part of drug discovery and target selection systems, and also for identifying adverse drug effect descriptions [141]. A statistical study [142] has shown the increasing need of TM in drug development. For example, the importance of TM in pharmacogenomics is getting so much attention [143] that the Pacific Symposium on Biocomputing (PSB) has dedicated one edition to pharmacology studies [144]. Though the focus of this thesis is PK parameter extraction, we will first review the study of TM in some drug related fields to illustrate how this type of information is handled by TM.

1.3.1 Drug-Drug Interaction Mining

Drug-drug interaction (DDI) is a situation in which a drug's activity, in the aspect of pharmacodynamics, pharmacokinetics and drug efficacy, is changed by another drug if administered together. DDI discovery is a fairly new area but the application of TM has already shown some progress. One study discover two interacting drugs via the network of genes they are both connected with [145]. This DDI prediction is based on pharmacogenomics facts extracted with biomedical TM of Medline abstracts, from which predefined and normalized gene names, drug names and relationship terms are extracted at the sentence level (**Figure** 1**.**1 a). A semantic network was constructed for these extracted entities (**Figure** 1**.**1 b), which are used as features for a drug pair which is connected by only one gene. This paper utilizes random forest [84] as the classifier for all drug pairs and claimed it outperformed both logistic regression [85] and support-vector machine. The classifier recognizes the combinations of relationships, drugs and genes that are most associated with the gold standard DDIs, correctly identifying 79.8% of assertions relating interacting drug pairs and 78.9% of assertions relating non-interacting drug pairs. Another paper focuses on the metabolism-based interactions, which means the DDI is linked via related enzymes and transporters [146]. So instead of a common gene as shown in (**Figure** 1**.**1 a), here enzymes or transporters define the interaction network. The method in this second paper uses natural language processing and logic reasoning [86] (i.e. rule application). This study is based on Medline abstracts, and claims to detect 81.3% DDIs correctly.



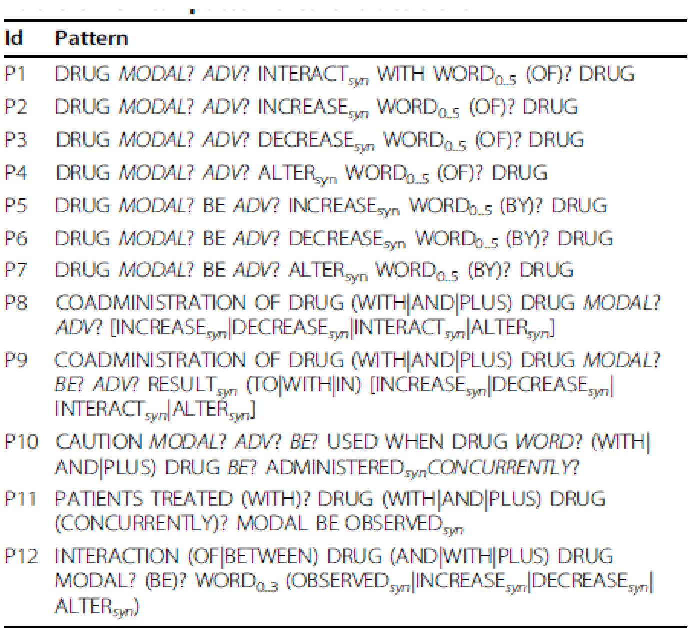
(a) (b)

Figure 1.1 : Example of Entity Process for DDI Prediction [145].

Another type of DDI prediction strategy is to mine direct interaction statements from literature or pharmacological documents (e.g. drug labels). One such study worked on drug descriptions downloaded from the "Interactions" field in DrugBank [147]. A linguistic rule-based approach, combining shallow parsing and pattern matching, was applied in this study to extract DDIs from text. Example rules are listed in Figure 1.2. Unfortunately, this approach yielded poor results (precision = 48.89%, recall = 24.81%, F-measure = 32.92%). The authors argued that the UMLS MetaMap Transfer (MMTx) tool [148] used in this paper is not powerful enough in NLP processing, not able to determine the syntactic type of a phrase, classifying it as an unknown phrase. Also its sentence clause splitting algorithm should be improved. This paper also discussed the impact of negation cases in DDI prediction. If not considered properly, negation DDIs can compromise the mining performance. Thus a DDI study should carefully classify such cases. This paper contributed an annotated drugDDI corpus (http://labda.inf.uc3m.es/DrugDDI/) which collects over 400 DDI sentences most about drug efficacy and pharmacokinetics.

In follow-up work, the same group developed a supervised machine learning technique[149]. The new approach achieved a precision of 51.03%, a recall of 72.82% and an F-measure of 60.01% on the same drugDDI corpus. It is based on Shallow Linguistic Kernel (SLK) method, which has successfully been applied to the extraction of protein–protein interactions (PPIs) [150]. So this paper demonstrates that methods applied in PPI studies have the potential to be ported to DDI prediction. The SLK algorithm is basically implemented as a context-feature scoring method here. It contains two kernel functions for global context and local context. The global context kernel is designed to discover the presence of a relation between two entities by using information from the whole sentence. The local context kernel is based on the hypothesis that the contextual information of candidate entities is particularly useful for the verification of a relationship existing between them. In particular, windows of limited size around entities provide useful clues for the identification of the entities' roles within a relation. The scores from the

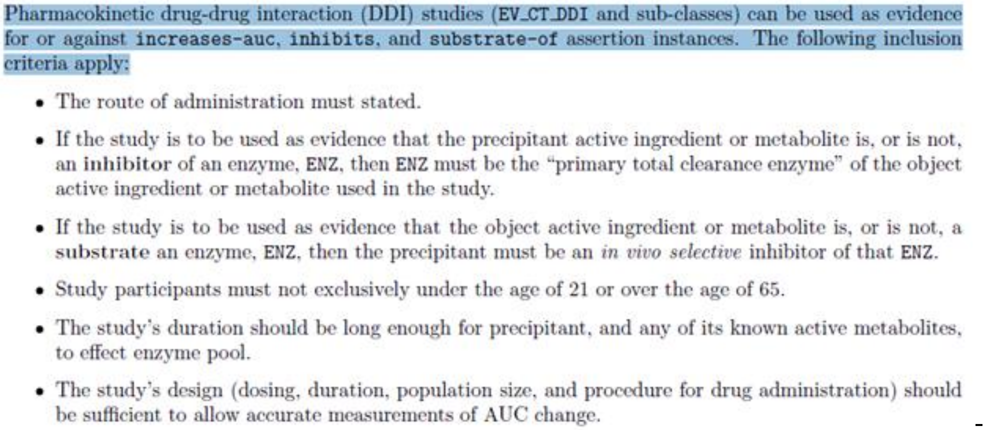
SLK module were then fed to an SVM for DDI classification. This paper thus concluded that the ML approach is far more efficient than the pattern-based approach for tackling DDI extraction from texts.



Id	Pattern
P1	DRUG *MODAL? ADV?* INTERACT$_{syn}$ WITH WORD$_{0..5}$ (OF)? DRUG
P2	DRUG *MODAL? ADV?* INCREASE$_{syn}$ WORD$_{0..5}$ (OF)? DRUG
P3	DRUG *MODAL? ADV?* DECREASE$_{syn}$ WORD$_{0..5}$ (OF)? DRUG
P4	DRUG *MODAL? ADV?* ALTER$_{syn}$ WORD$_{0..5}$ (OF)? DRUG
P5	DRUG *MODAL?* BE *ADV?* INCREASE$_{syn}$ WORD$_{0..5}$ (BY)? DRUG
P6	DRUG *MODAL?* BE *ADV?* DECREASE$_{syn}$ WORD$_{0..5}$ (BY)? DRUG
P7	DRUG *MODAL?* BE *ADV?* ALTER$_{syn}$ WORD$_{0..5}$ (BY)? DRUG
P8	COADMINISTRATION OF DRUG (WITH|AND|PLUS) DRUG *MODAL? ADV?* [INCREASE$_{syn}$|DECREASE$_{syn}$|INTERACT$_{syn}$|ALTER$_{syn}$]
P9	COADMINISTRATION OF DRUG (WITH|AND|PLUS) DRUG *MODAL? BE? ADV?* RESULT$_{syn}$ (TO|WITH|IN) [INCREASE$_{syn}$|DECREASE$_{syn}$|INTERACT$_{syn}$|ALTER$_{syn}$]
P10	CAUTION *MODAL? ADV? BE?* USED WHEN DRUG *WORD?* (WITH|AND|PLUS) DRUG *BE?* ADMINISTERED$_{syn}$*CONCURRENTLY?*
P11	PATIENTS TREATED (WITH)? DRUG (WITH|AND|PLUS) DRUG (CONCURRENTLY)? MODAL BE OBSERVED$_{syn}$
P12	INTERACTION (OF|BETWEEN) DRUG (AND|WITH|PLUS) DRUG MODAL? (BE)? WORD$_{0..3}$ (OBSERVED$_{syn}$|INCREASE$_{syn}$|DECREASE$_{syn}$|ALTER$_{syn}$)

Figure 1.2 : Example of Lexical Patterns to Extract DDIs [147].

Though the above machine learning method has shown superior performance on the drugDDI corps, rule based methods are still being applied and under improvement for DDI discovery as such methods have the potentiality to perform effectively on certain topics if rules are designed properly. For this purpose, a comprehensive repository of knowledge about drug mechanism was developed, the Drug Interaction Knowledge Base (DIKB) [151], with standard DDI evidence inclusion criteria. In DIKB, many types of evidence are defined, e.g. the criteria for PK DDI study in Figure 1.3. Evidence is further classified into different levels. Based on DiKB, rules can be designed to make DDI assertions using levels of evidence (LOE) and different combination of LOEs as assertion criteria. Its effectiveness was tested using clinical records from PubMed search and drug labels which showed variant recall (0.88-1) and precision (0.8-1.0) for different prediction criteria tests [152].



Pharmacokinetic drug-drug interaction (DDI) studies (EV_CT_DDI and sub-classes) can be used as evidence for or against increases-auc, inhibits, and substrate-of assertion instances. The following inclusion criteria apply:

- The route of administration must stated.
- If the study is to be used as evidence that the precipitant active ingredient or metabolite is, or is not, an **inhibitor** of an enzyme, ENZ, then ENZ must be the "primary total clearance enzyme" of the object active ingredient or metabolite used in the study.
- If the study is to be used as evidence that the object active ingredient or metabolite is, or is not, a **substrate** an enzyme, ENZ, then the precipitant must be an *in vivo selective* inhibitor of that ENZ.
- Study participants must not exclusively under the age of 21 or over the age of 65.
- The study's duration should be long enough for precipitant, and any of its known active metabolites, to effect enzyme pool.
- The study's design (dosing, duration, population size, and procedure for drug administration) should be sufficient to allow accurate measurements of AUC change.

Figure 1.3 : Example of Inclusion Criteria in DiKB [151]

1.3.2 Kinetics Parameter Mining

PK parameter mining has not been reported by other research groups so far. However, there have been studies addressing general kinetics parameter mining problems. Ordinary differential equations (ODEs) used for general biological kinetic system (e.g. enzyme kinetics) modeling are quite similar to those used in drug modeling. Thus, strategies for kinetics parameter mining have high potentiality to drug PK parameter mining.

One such TM system classifies documents regarding the question of whether or not they contain experimentally obtained parameters for kinetic models using a support vector machine [153]. This system is illustrated in Figure 1.4, which includes manual assisted SVM model training and automatic document classification. It works on PDF full text documents. A tool named *PDFTOTEXT* [154] was used to convert documents from PDF to ASCII formats. Classifying 791 pre-selected publications with SVM model yielded a precision of 60% at 50% recall. This system only focuses on the classification of relevant publications with the lack of automatic kinetic parameter extraction.

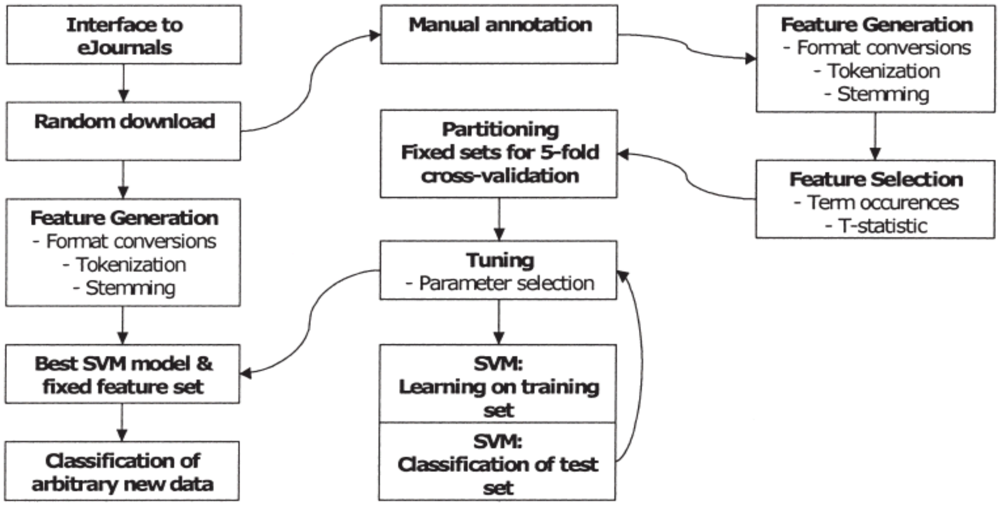
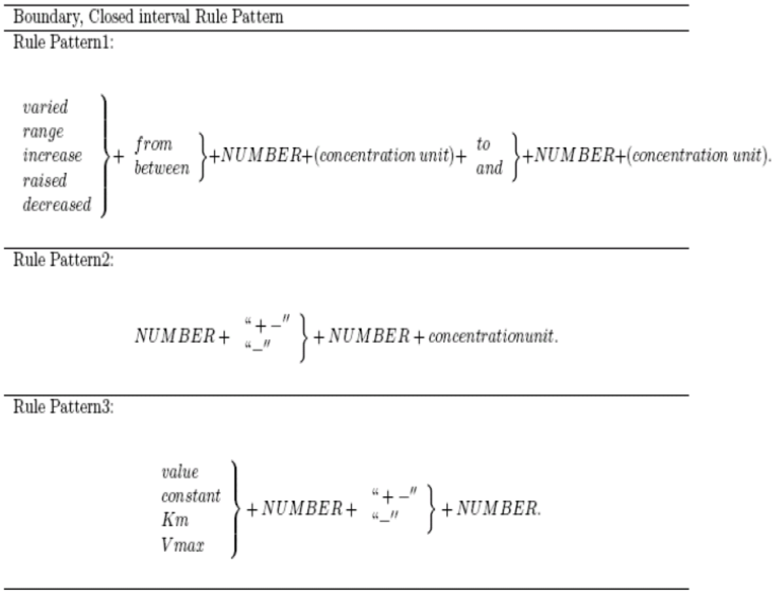


Figure 1.4 : Work Flow of A Kinetics Mining System [153]

Compared with mining kinetics data for general biological systems [153], KiPar is an IR application developed specifically for retrieving relevant documents with enzyme kinetic parameters for quantitative modeling of yeast metabolism [139]. This is a rule based system relying on entity recognition. Instead of working on the relevance analysis of abstracts, KiPar searches PubMed and PubMed Centeral (PMC) using customized query patterns. Such query patterns are made of entities from relevant open resources, e.g. enzyme names from the KEGG enzyme database [155], gene names from the Gene Ontology [156], and kinetic parameters from the Systems Biology Ontology (SBO) [157]. Retrieved documents were then scored based on the relevant entities contained. Its retrieval performance was compared with basic Boolean search by using Entrez, which shows 36% improvement for abstracts (PubMed) and 100% for full text (PubMed Central).

Research on kinetics mining has gone further from IR to IE today. One paper [158] presented a method of kinetics parameter extraction. In its name entity recognition (NER) step, POS tagging and orthographic feature detection were applied to labels describing kinetic parameter type, value and annotation for sentences (examples in Figure 1.5**)**. Then the labeled entities were associated using a set of matching rules to recognize and extract kinetics parameters and their related annotations. This tool shows an overall 76% precision and 87% recall in relevant sentence classification, and 75% precision and 90% recall in parameter recognition. Another similar rule and dictionary-based TM algorithm [159] for chemical and biological kinetics data relies on the identification of entities in the text and a rule-based linkage of these units. Its NER is based on dictionaries manually constructed according to expert knowledge. This method was tested using PubMed abstracts. A manual verification of the results yielded a recall between 51% and 84% and a precision ranging from 55% to 96%, depending on the category searched (e.g. enzyme, organism, or ligand). The results were stored in a database (KID kinetic Database) which is available via http://kid.tu-bs.de/ and the source code is also provided.



Boundary, Closed interval Rule Pattern

Rule Pattern1:

$$\left.\begin{matrix}\textit{varied}\\\textit{range}\\\textit{increase}\\\textit{raised}\\\textit{decreased}\end{matrix}\right\}+\left.\begin{matrix}\textit{from}\\\textit{between}\end{matrix}\right\}+NUMBER+(\textit{concentration unit})+\left.\begin{matrix}\textit{to}\\\textit{and}\end{matrix}\right\}+NUMBER+(\textit{concentration unit}).$$

Rule Pattern2:

$$NUMBER+\left.\begin{matrix}\text{“}+-\text{”}\\\text{“}-\text{”}\end{matrix}\right\}+NUMBER+\textit{concentrationunit}.$$

Rule Pattern3:

$$\left.\begin{matrix}\textit{value}\\\textit{constant}\\Km\\Vmax\end{matrix}\right\}+NUMBER+\left.\begin{matrix}\text{“}+-\text{”}\\\text{“}-\text{”}\end{matrix}\right\}+NUMBER.$$

Figure 1.5 : Rule Pattern for A Kinetics Mining System [158]

1.3.3 Pharmacokinetics Parameter Mining

Until today, most of the drug-related TM still focused on relationships, especially in the prediction of drug-drug interactions (DDIs). TM for DDI prediction has been processed either for pathway analysis (i.e. association study between drugs and genes, enzymes or transporters) [145, 146] or for extraction of direct DDI statements in text [147, 149]. Another potential DDI prediction strategy is based on drug PK modeling, which relies heavily on the availability of related PK parameters, most of which are still uncollected from scientific publications. As TM has been proved effective in biomedical data collection from previous studies, it should be applied for PK parameter extraction from literature. No TM research in this field exists yet. Thus

PK parameter mining serves not only as facilitation for DDI prediction study but also a meaningful span of TM to drug research area.

Most closely related existing TM studies are kinetic parameter mining for biological systems. Some effective kinetics IR systems [139, 153] have been developed to classify relevant articles. Furthermore, some dictionary and rule based IE methods [158, 159] were also presented for automatic kinetic parameter extraction. As the kinetic system modeling has some similarity with PK modeling, i.e. the expression of Ordinary Differential Equations (ODEs), and expressions in the aspect of kinetic parameter symbols, numeric and units, the strategies used by kinetic mining can thus provide good reference to PK mining. Based on our observation, only one system [153] applied a machine learning (i.e. SVM) strategy in document classification, while the other three systems [139, 158, 159] each used a dictionary and rule based method for either classification or extraction. This situation may be caused by the lack of good training data for machine learning algorithms, while expert knowledge based rules can be developed without such limit. However, the rules can be really complex and need to be comprehensive enough to cover all relevant data and specific enough to filter out irrelevant data. Thus rules should be carefully developed and they can hardly be ported to another mining task.

Kinetic IE systems [158, 159] only enumerate all extracted kinetics data as results. Though a performance test was provided based on a subset of extracted data or a group of artificial test data, they usually lack a mechanism to evaluate all of the extracted data on the population level. Thus it is hard to predict the influence of such mined kinetics parameters to the system modeling. Considering that the kinetics parameters were extracted from various systems, it is debatable whether such parameters fit one specific user defined biologic system model. Furthermore, despite the presentation of methods and results for kinetics parameter mining, these systems lack a mechanism for user customized and guided data collection. Thus they have not reached the stage of being applied in real world automatic kinetic parameter data collection, though their assistance to certain kinetics database curation is very promising.

Though existing kinetic mining systems provided valuable reference for PK parameter mining, none of them tried information extraction from tables that usually contain important kinetic information but too complex to be handled. This situation has inspired a discussion to standardize table structure [160], which proposes approaches turning human-readable tables from literature into structured digital tables on the Semantic Web (in the form of machine-readable triples). Such machine-readable tables (produced by individual authors/curators/editors) can be automatically/semantically linked to each other and then can be easily mined by programs developed by other researchers (possibly in some other discipline).

1.4 Thesis Overview

This thesis focuses on exploring TM in the pharmaceutical area; specifically in the extraction of PK data from both abstracts and full-text articles. It covers most of the TM areas discussed above: ontology construction, information retrieval, information extraction, text mining of abstracts and full-text.

1.4.1 Pharmacokinetics Ontology

Biomedical literature mining usually involves heavy use of domain-specific terminology. However, the ambiguity of terms and heterogeneous description of studies by different researchers make the information extraction for drugs a very difficult job. One solution is to recognize the key entities with the assistance of standardized terminologies and ontologies [47], which enable the consistent large scale extraction of relations ideally leading to a performance close to the reasoning process of human curators.

For information retrieval purposes, we construct an entity template to classify relevant articles. The performance of TM for drug PK parameters is based on correct recognition of drug names and experiment design terms. The coexistence of other drugs also needs to be tagged for syntactic analysis, and their interactions with the object drug should be considered to provide extra guidance to locate the target information. Thus, the construction of a PK ontology, as a comprehensive summary of terms and concepts in PK studies, should provide valuable reference to both the semantic and syntax analysis in the mining process.

In this thesis, we collect data from multiple resources and link several existing biomedical ontologies to cover the essential knowledge structure for drug studies. We build a PK ontology, which is formalized in Protégé [161] and uploaded to BioPortal [162] website for public reference. It can be applied in text processing for feature selection, domain knowledge summary, template construction, and semantic tagging etc. Overall, this ontology serves as a quick reference of the domain knowledge and also a product of the research described in this dissertation.

1.4.2 Text Mining on Abstracts

Literature mining for PK parameters is highly unique. Firstly, important PK parameters (entities) are specifically defined. These PK parameters are usually available from different drug studies, which may vary by factors such as units, sub-populations, study designs, and dose regimens. Secondly, each PK study focuses on certain aspects of a drug’s PK features. Some key PK parameters, e.g. concentration and clearance, are generally measured, while some others are not. So the retrieved information from publications can be incomplete for some PK parameters and abundant for others. Thirdly, as the mined PK numerical data is applied on drug PK models, the quality of mined data determines the performance of the modeling. Thus false positive findings need to be filtered from mined results as thoroughly as possible. Also, one barrier that literature mining faces is the relative lack of standards to evaluate the performance of mining strategies.

Therefore, we tested the feasibility of PK numerical information extraction in published scientific literature abstracts before initiating further work [163]. For this purpose, we did some preparation work which include curation of PK data as a gold standard for model training and evaluation, as well as the construction of a PK ontology for drug PK studies.

1.4.2.1 Manual Drug Pharmacokinetics Data Curation

In PK TM using abstracts, it is difficult to evaluate performance because of the lack of existing drug PK data as a gold standard. Therefore, we manually curate PubMed abstracts for a test drug (Midazolam) to find relevant articles. The whole process is tedious and time consuming, a painful way to highlight how much we are in an urgent need to improve the performance of literature mining for drug studies. However, ultimately we curate high quality data which can be used as a gold standard for subsequent mining projects. The PK parameter data generated from this manual collection step, together with validated data from mining, is saved into a PK database as a result of this dissertation. However, such tedious manual curation can only work on a limited number of drugs, collecting PK data for all thousands of FDA drugs needs assistance of TM.

1.4.2.2 Abstract Pharmacokinetics Parameter Mining

We develop a sequential mining strategy to discover PK relevant studies in abstracts, extract the PK parameter data, evaluate the results and remove false positives. This approach is tested by extracting clearance data for drug *Midazolam*. Firstly, an entity template library is built based on the PK ontology to retrieve pharmacokinetics relevant abstracts. Then a set of rules are designed to tag and extract PK data from the sentences. Because the extracted PK data are from various publications (studies) which adopt various experiment conditions (e.g. subject races, gender), their population means and between study variance are also estimated to be useful for PK modeling.

1.4.3 Text Mining of Full Text Documents

The main purpose of this dissertation is to collect drug PK data on a large scale, especially published PK numerical data which has not been documented in any databases. After mining on PubMed abstracts, we may still miss a large amount of available data in the full text literature, which usually contains much more PK information than abstracts. Thus full-text articles are also used as document resources for building the PK database. The full-text based mining pipeline, in addition to TM on abstracts, is the main product of this dissertation.

1.4.3.1 Article clarification

The proposed IR strategy is based on a manually constructed entity template. Such template can summarize the key features of a relevant article, such as the existence of certain PK parameters in the abstract. However, PK existence can also be determined from non-PK related studies, e.g. drug-drug interaction studies, or from a non-target drug PK studies. Thus the template-based IR can achieve high recall, but comparatively low precision. A low precision means the inclusion of many false positive articles, which can bring too much noise to the subsequent full-text mining pipeline. So the IR step needs to improve on precision while not sacrificing recall. For this purpose, we proposed a two-step IR system, combining template-based filtering, on both titles and abstract sentences, with a machine learning method, i.e. Conditional Random Field (CRF) [123], on filtered sentences from the first step.

1.4.3.2 Article Collection

The automatic retrieval of full-text articles still remains a challenge for text mining studies, due to lack of comprehensive article repositories. One solution is to collect full-text literature from open resources such as PubMed Central, Google Scholar, Highwire, and OVID, or using existing article retrieval systems such as AJAXSearch [164], Open Search Server [165]. However, most vendors set a stringent limit on bulk article downloading. For now, the best solution would be to download a small number of articles for a drug at a time from an open resource. In this thesis, we use PubMed as the resource for article collection. To check the availability of full-text articles, we did a complete search and found that 106 (58%) out of 183 Midazolam PK relevant papers can be downloaded through PubMed external links or PubMed Central.

1.4.3.3 Tabular Data Extraction

To have a rough idea of PK data distribution in the full text, we browsed PK relevant articles for the drug Midazolam. It turned out around 60% of them have one or more numerical data tables, which summarize the PK study results. As PK information highly concentrated area, the table becomes an essential part of full text for the PK data extraction. However, there is little research on information extraction from tabular data, with the rare exception of the extraction of relevant tables [166] or table captions instead of extracting specific data from tables. In this dissertation, we developed an effective solution of extracting tabular data.

1.4.3.4 Pharmacokinetics Information Extraction from Full Text Documents

The performance of information extraction from full text has been reported to be improved by finding relevant sentences first [167, 168]. So our strategy of TM of articles is to focus on sentences with PK information. As the full text shares high similarity with abstracts in aspects of entities and expressions for PK data in sentences, we can expect similar performance by adopting same mining strategies here as in abstract based feasibility test. After IE from full-text articles, we compared information content increase with IE from abstracts and tables.

1.4.4 Pharmacokinetics Parameter Transformation

One of the main purposes of this dissertation is to provide parameter data support for PK modeling, which is the central piece of model-based drug development. Such modeling divides the human body into multiple compartments according to a target drug's varying kinetics characteristics in its functioning path in human body. However, the mined numerical PK parameter data is mostly from clinical trials, which take the human body as a whole. Thus we propose a meta-analysis approach [169] to transform PK parameters from TM to be modeling compatible, which makes TM really meaningful in real world.

1.4.5 Thesis structure

The whole thesis is illustrated in figure 1.1. The data is from PubMed search, for both abstracts and full text articles. Chapter 2 covers PK ontology construction. Abstract based text mining is presented in chapter 3, which also discusses approaches to clean up the PK parameter collection results. Text mining of full-text documents is presented in chapter 4. The PK parameter transformation and database construction for PK TM are discussed in chapter 5 and 6 respectively.

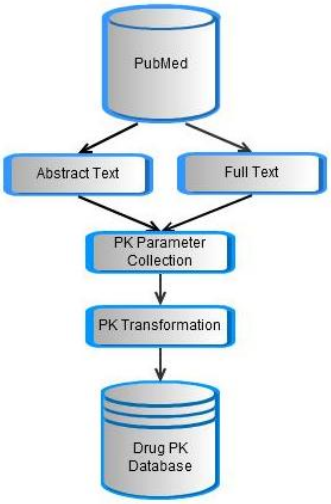


Figure 1.6 : The Scheme of This Thesis.

Chapter 2.

PHARMACOKINETICS ONTOLOGY

A well-annotated pharmacokinetics corpus and ontology can facilitate the development of text mining tools, data collection, and integration from multiple databases in pharmacokinetics. The comprehensive pharmacokinetics ontology we develop in this chapter serves this purpose. It can annotate all aspects of *in vitro* pharmacokinetics experiments and clinical pharmacokinetics studies. This work was a collective effort developed in a Dr. Lang Li's lab.

2.1 Pharmacokinetics Ontology Construction

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

Experiment specifies *in vitro* and *in vivo* PK studies and their associated PK parameters. The PK parameters of the *single-drug metabolism experiments* 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}) [170]. 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* [171]. There are also multiple drug-interaction mechanisms: *competitive inhibition*, *non-competitive inhibition*, *uncompetitive inhibition*, *mechanism based inhibition*, and *induction* [172]. 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.

Table 2.1 : PK Ontology Categories

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

In vitro experiment conditions are also included in PK ontology. Metabolism enzyme experiment conditions include buffer, NADPH sources, and protein sources. In particular, protein sources include recombinant enzymes, microsomes, hepatocytes, *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 this information can be found in the FDA website http://www.abclabs.com/Portals/0/FDAGuidance_DraftDrugInteractionStudies2006.pdf .

In vivo PK parameters were summarized from two books [173, 174]. There are several main classes of PK parameters: Area under the concentration curve parameters (AUC_{inf}, AUC_{SS}, AUC_t, AUMC), drug clearance parameters (CL, CL_b, CL_u, CL_H, CL_R, CL_{po}, CL_{IV}, CL_{int}, CL_{12}), drug concentration parameters (C_{max}, C_{SS}), extraction ratio and bioavailability parameters (E, E_H, F, F_G, F_H, F_R, f_e, f_m), rate constants (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, $\lambda_{2;}$ 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).

We also account for two types of pharmacokinetics models that usually presented in the literature: *non-compartment model* and *one or two-compartment models*. There are multiple items that 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 radiography.

CYP450 family enzymes predominantly exist in the gut wall and liver. Transporters are tissue specific. 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. The information about the probe inhibitors, inducers, and substrates of CYP enzymes and all transporters were collected from http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm064982.htm, reviewed in the top pharmacology journal [175].

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 [176]. 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 serve 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 field was created using drug names from DrugBank 3.0 [177]. DrugBank consists of 6,829 drugs which can be grouped into different categories of FDA-approved, FDA approved biotech, nutraceuticals, and experimental drugs. Drug names are mapped to generic names, brand names, and synonyms.

Subject includes existing ontologies for disease, physiology, and population from http://bioportal.bioontology.org. The PK ontology was implemented with Protégé [178] and uploaded to the BioPortal ontology platform.

2.2 Utility of Pharmacokinetics Ontology

2.2.1 Study Annotation

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 [179]. 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.

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 Table 2.2, the trial summary is well organized by the PK ontology.

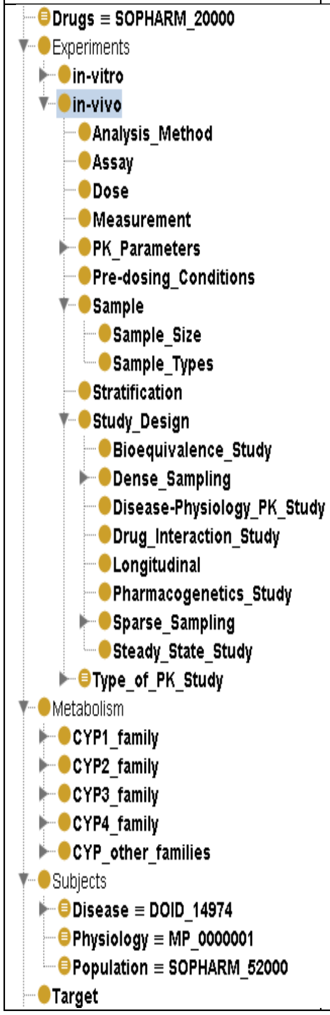
Example 2 Midazolam/Ketoconazole Drug Interaction Study

This was a cross-over, three-phase, drug interaction study [180] (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 annotation is shown in Table 2.2 column three.

Example 3 in vitro Pharmacokinetics Study

This was an *in vitro* study [181], 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 annotation is shown in Table 2.3.

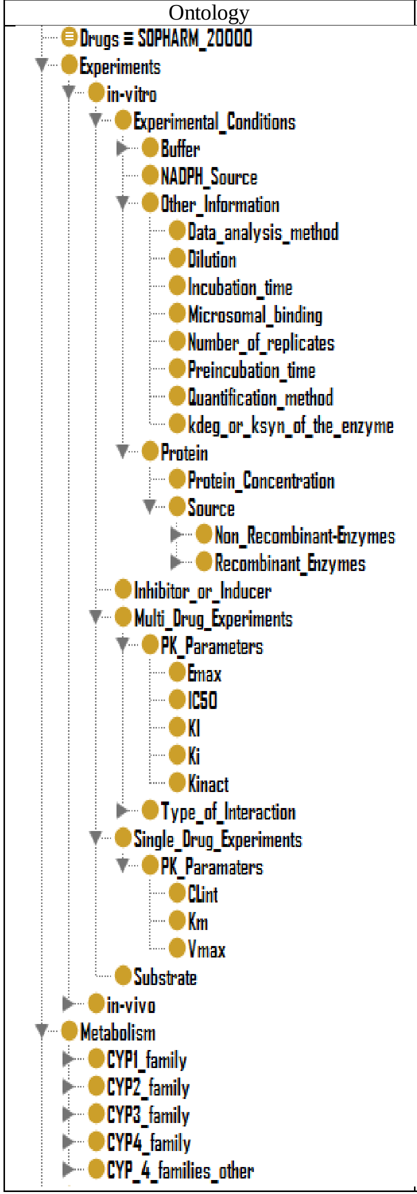
Table 2.2: Clinical PK Studies



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	298	24
Sample_Size	Blood	blood
Sample_Types	prior chemo, menopausal	
Stratification		
Study_Design		
Bioequivalence_Study		
Dense_Sampling		
Disease-Physiology_PK_Study		inhibition
Drug_Interaction_Study	Longitudinal	three-phase crossover
Longitudinal	prospective, single arm	prospective, single arm
Pharmacogenetics_Study		
Sparse_Sampling	steady state	
Steady_State_Study		
Type_of_PK_Study		
Metabolism		
CYP1_family	CYP2D6, 2C9, 2B6	
CYP2_family	CYP3A4/5	CYP3A4/5
CYP3_family		
CYP4_family		
CYP_other_families		
Subjects	breast cancer	healthy volunteers
Disease ≡ DOID_14974		
Physiology ≡ MP_0000001	Caucasian/African American	
Population ≡ SOPHARM_52000	ESR1/ESR2	
Target		

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.

Table 2.3 : *in vitro* PK Study



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	

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.

2.2.2 Pharmacokinetics Corpus

To illustrate the application of our PK ontology, a PK abstract corpus was constructed to cover four primary classes of PK studies: clinical PK studies (n = 60); clinical pharmacogenetic studies (n = 60); *in vivo* DDI studies (n = 218); and *in vitro* drug interaction studies (n = 208). The PK corpus construction process was a manual process that calls us to test various ML later on. The abstracts of clinical PK studies were selected from PubMed search results using the most popular CYP3A substrate, midazolam and *pharmacokinetics* as query terms. The clinical pharmacogenetic abstracts were selected based on the most polymorphic CYP enzyme, CYP2D6. These two selection strategies represent very well all the *in vivo* PK and PG studies, constituting about 50% of total CYPs in the human body. For drug interaction studies, abstracts were selected via a PubMed search using probe substrates/inhibitors/inducers (see section 2.1) for metabolism enzymes as query terms followed by manual screening.

Once the abstracts were identified in four classes above, they were annotated manually by curators (3 masters and one Ph.D.) with different training backgrounds: computational science, biological science, and pharmacology. In addition a random subset of 20% of the abstracts that had consistent annotations among four annotators, were double checked and reviewed by two Ph.D. level scientists.

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

Keyterms include drug names, enzyme names, PK parameters, numbers, mechanisms, and change. These terms among different annotators were recognized by the following standard.

- *Drug names* were defined mainly on DrugBank 3.0. 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) [182].
- *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|450|1-26)(A-Z)(1-99|*1-99).
- *PK parameters* were annotated based on the defined *in vitro* and *in vivo* PK parameter ontology. 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(ing|s|ed|tion|or), catalyz(ing|es|ed), correlat(ing|es|ed|tion), metaboli(zing|zs|zed|sm|or), induc(e|es|ed|or|tion|ing), form(s|ing|ation|ed), stimulat(e|es|ed|ing|ion), activ(e|ate)(ated|ates|ating|ation), and suppres(s|ses|sed|sing|sion).
- *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, probably, may, might, minor, little, negligible, doesn't interact, affect(s|ed|ing), reduc(e|es|ed|tion|ing), and increase(s)(ing)(ed).

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:

- <u>Clear DDI Sentence</u> (CDDIS): two drug names (or drug-enzyme pair in the in vitro study) are in the sentence with a clear interaction statement.
- <u>Vague DDI Sentence</u> (VDDIS): One drug or enzyme name is missing 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. <u>In *in vivo* studies</u>, three types of DDI relationships were defined: *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. DDI, ADDI, and NDDI can be thus 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}. <u>In *in vitro* studies</u>, six types of DDI relationships were defined. *DDI*, *ADDI*, *NDDI* were similar to in vivo DDIs, but three more drug-enzyme relationships were further defined: *DEI*, *ambiguous DEI* (ADEI), and *non-DEI* (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 define DEI relationship quantitatively. The priority rank of *in vitro* PK parameters is Ki > IC50.

Krippendorff's alpha [183] 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 2.4. 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 [184]. 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 colors for the entities in the corpus. This feature allows the users to visualize annotated entities. A DDI Corpus was recently published by another team, 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.

Table 2.4 : Annotation Performance Evaluation

	Annotation Categories	Frequencies	Krippendorff's alpha
Keyterms	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	

2.2.3 A Drug Interaction Text Mining

After PK corpus was constructed, we can further test the application of PK ontology by a text mining study on drug interaction. Prior to performing DDI extraction, the testing and validation DDI abstracts in our corpus was pre-processed and converted into the unified XML format [185]. 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, PoS tags and dependency tree generation using the Stanford parser [186].
- 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 [185] using interacting components (drugs) (Figure 2.1). 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 [186] 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 [185].

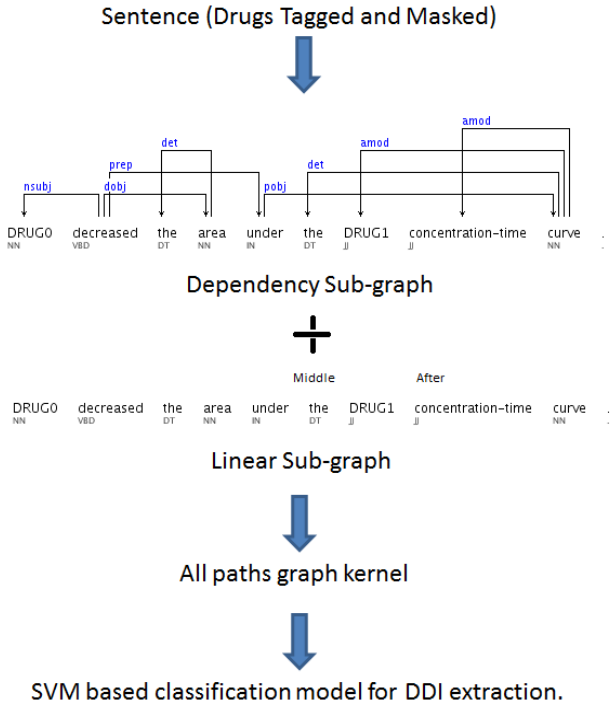


Figure 2.1 : Graph Kernel Approach

DDI extraction was implemented in the *in vitro* and *in vivo* DDI corpus (see section 2.2.2) as test data separately. In extracting *in vivo* DDI pairs, the precision, recall, and F-measure 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, 0.52 respectively. In our early DDI research published in the DDIExtract 2011 Challenge [187], 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.

2.3 Conclusions and Discussions

The comprehensive PK ontology is available at http://rweb.compbio.iupui.edu/corpus/ontology/, It annotates both *in vitro* PK experiments and *in vivo* PK studies. Using our PK ontology, a PK corpus was also developed as described in section 2.2.2. This PK corpus is valuable at http://rweb.compbio.iupui.edu/corpus/, using it we extracted drug interactions relationship via text mining. This DDI text mining demonstrates how our PK ontology can facilitate the development of text mining tools.

Just as general biomedical literature mining usually dealing with heavy use of domain-specific terminology, PK numerical parameter collection also needs to recognize entities used in PK studies. As the essential knowledge structure for drug studies was summarized in our PK ontology, it can be applied in text processing for feature selection, domain knowledge summary, template construction, and semantic tagging etc. Overall, this ontology serves as a quick reference of the domain knowledge.

In the feasibility test of PK parameter data collection (next chapter), we constructed an entity template to classify relevant articles for information retrieval purposes with the assistance of standardized terminologies in our PK ontology. Also, the performance of drug PK parameters extraction is based on correct recognition of drug names and experiment design terms. The coexistence of other drugs also needs to be tagged for syntactic analysis. Thus, our PK ontology, as a comprehensive summary of terms and concepts in PK studies, should provide valuable reference to both the semantic and syntax analysis in the drug PK related mining process.

Chapter 3.

TEXT MINING OF PUBMED ABSTRACTS

In this chapter[1], we investigate the feasibility of extracting drug PK parameter data in numerical form using a sequential mining strategy. Firstly, an entity template library is built to retrieve pharmacokinetics relevant articles. Then a set of tagging and extraction rules are applied to retrieve PK data from the article abstracts. To estimate the PK parameter population-average mean and between-study variance, a linear mixed meta-analysis model and an E-M algorithm are developed to describe the probability distributions of PK parameters. Finally, a cross-validation procedure is developed to ascertain false-positive mining results. Using this approach to mine midazolam (MDZ) PK data, an 88% precision rate and 92% recall rate are achieved, with an F-score = 90%. It outperforms a support vector machine (SVM) based mining approach, which leads to an F-score of 68.1%. Repeating the methodology on 7 additional drugs leads to similar performance.

[1] This chapter was published as ref. 163. Wang, Z., et al., *Literature mining on pharmacokinetics numerical data: a feasibility study*. J Biomed Inform, 2009. **42**(4): p. 726-35.

3.1 Midazolam Case Study Overview

The goal here is to extract all pharmacokinetics (PK) related information for a given drug. We used midazolam (MDZ) as the test drug. The mining is performed on abstracts from PubMed. One example of a MDZ PK relevant abstract [123] is:

> *To study the effects of cirrhosis of the liver on the pharmacokinetics of midazolam single IV (7.5 mg as base) and p.o. (15.0 mg as base) doses of midazolam were administered to seven patients with cirrhosis of the liver and to seven healthy control subjects… The elimination of midazolam was significantly retarded in the patients as indicated by its lower total clearance (3.34 vs. 5.63 ml/min/kg), lower total elimination rate constant (0.400 vs. 0.721 h-1), and longer elimination half-life (7.36 vs. 3.80 h). The bioavailability of oral midazolam was significantly (P less than 0.05) higher in patients than controls (76% vs. 38%)…*

The search engine of PubMed is not powerful enough to limit the search results to a specific topic, i.e. human PK study. So a further filtering step is necessary to remove irrelevant articles from PubMed search results, and keep the PK relevant abstracts which usually contain information of the following relevant key-term categories:

- Subject type (race, age, sex etc.) and size
- MDZ dose and administration route (oral, intravenous and etc.)
- PK parameters, such as AUC (area under the concentration-time curve), half-life, bioavailability, clearance, and etc.

Besides the key-term categories above, we limit mining to PK data from healthy human subjects and for the target drug (i.e. no other factors involved such as drug inhibitor/activator) to comply with the requirements of drug PK study.

Hence, in the example abstract, the literature mining tool should be able to extract "seven healthy control subjects" as subject, "IV (7.5 mg as base) and p.o. (15.0 mg as base)" as dose, "3.34 vs. 5.63 ml/min/kg" as total clearance, "0.400 vs. 0.721 h-1" as elimination rate, "7.36 vs. 3.80 h" as half-life, and "76% vs. 38%" as bioavailability. To be more precise, the mining tool should be able to recognize which clearance value of MDZ refers to healthy subjects, e.g. for the total clearance data, "3.34 ml/min/kg" is from patients and "5.63 ml/min/kg" is from healthy subjects.

For PK mining on abstracts, we developed a rule-based information extraction system. The architecture is shown in Figure 3.1. Abstracts are downloaded from PubMed after an initial query for a target drug, i.e. midazolam. Text is preprocessed such that it is divided into sentences, and different forms of the same terms are stemmed. The next step is entity recognition. It determines sentence relevance, and tags the stemmed sentence terms as various entity classes. At the end of this step, only the more relevant abstracts are left and well tagged. In the information extraction step, a set of extraction rules are manually created and implemented. Then the mined data are analyzed by a statistical model to detect and remove outliers, which are potentially false positive items.

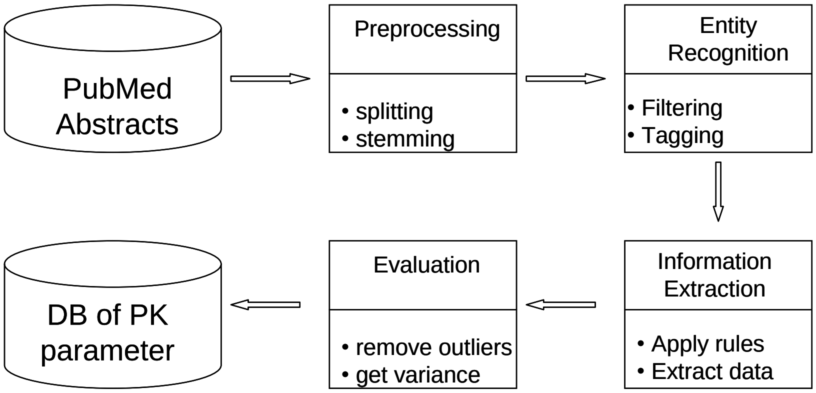


Figure 3.1 : The Architecture of Abstract Based PK Mining

3.2 Abstract Mining Methods

3.2.1 Text Preprocessing

Our PubMed search uses the drug name, e.g. midazolam, as the unique key-term in a query. The search results are downloaded with the XML format to get the structured abstract information. In the following mining process, only article title (<ArticleTitle>), abstract (<AbstractText>) and paper type (<PublicationType>) information is utilized from the XML format abstract.

The goal of the preprocessing step is to split the abstract text into units of sentences. There are some existing tools to do this job (e.g. SentenceDetector [188], MxTerminator [189]). Considering the simple grammar of the abstracts, we applied a Perl module (Lingua::EN::Sentence) for sentence splitting. The Porter stemming algorithm [190] is used to deal with the common morphological and inflectional endings from words in English. After stemming, each word in the abstracts is normalized into a standard form.

3.2.2 Entity Recognition

3.2.2.1 Entity Template Library

PubMed search with just a drug name as key term usually returns a large number of abstracts, e.g. 7,129 abstracts for midazolam. To increase the precision of the mined results, an abstract filtering step is necessary following text preprocessing. Firstly, as we limit the mining of PK data from healthy human subjects, the studies on diseased subjects should be removed. The human subject information (health status, race, weight…) is highly important and usually reported in pharmacokinetics studies. Most article abstracts state clearly whether the human subjects are healthy or diseased (patients). So if one abstract only mentions patients or diseased subjects, it is usually irrelevant; but if there is co-existence of healthy subject information (this is usually the control in clinical studies), it is still considered as subject relevant. For abstracts without any subject information, we kept them as relevant in case of data loss. Secondly, an entity template library is built upon the PK ontology for the further abstract filtering. It summarizes key factors in determining an abstract's relevance. Table 3.1 is a library example, which contains a list of relevant key-terms and a list of forbidden terms. The terms are in the stemmed format. Because some relevant abstracts do not have human subject information, subject terms are not included in the key-term list. Thus, these articles can be kept as relevant for future full text mining purposes. In addition, the drug terms should correspond to the studied drug. For midazolam, such terms include "midazolam" and "mdz".

Table 3.1 : Key Terms and Forbidden Terms

Key Terms		Forbidden Terms	
DRUG	<drug terms>	NTITLE	mice mouse rat animal penguins pig horse human liver microsom review
ROUTE	oral orally introven administr i.v. intramuscular		
PK	clearance pharmacokinet concentr bioavail auc elimin c(max) half-lif	NTYPE	review

The entity template library is a representation model for the relevant abstracts. The PubMed search abstracts are further filtered by this library. An abstract is considered relevant if it contains at least one term from each of the key-term categories, which include drug administration routes (ROUTE), PK parameters and DRUG (Table 3.1); and the abstract is considered as irrelevant if it contains one or more forbidden terms in either <NTITLE> or <NTYPE> (Table 3.1). As MDZ is primarily a CYP3A substrate, all of its DRUG key-terms are related to this metabolic enzyme. The <NTITLE> is the forbidden term list for article titles. These terms mostly represent animal and in-vitro studies (Table 3.1). The other forbidden term, <NTYPE>, is used to recognize the review articles. Since review articles contain PK data only from other publications, they don’t provide additional information.

3.2.2.2 Tagging Entities

All information in the key-term categories is necessary for a drug PK study. Obviously, in order to extract all the PK data from the abstract, we need to properly recognize all these relevant terms in each sentence. Thus the quality of the PK ontology becomes very critical for mining.

Subject Tagging

The subject information usually contains all or part of the following four key components: size, description, race and subject types. <SUB_part : term> is used to represent a term in each component, e.g. "seven healthy control subjects" can be tagged as "<SUB_N : seven> <SUB_D : healthy> control <SUB_T : subjects>".

Drug Tagging

A drug name dictionary is built based on drug list from PK ontology. Thus, the drug entities in the abstracts can be correctly tagged, e.g. midazolam to <DRUG : midazolam>.

Dosing Tagging

The dosing tagging covers drug administration ways, and dosing units. The dose is located by searching the numerical data lying ahead of its unit. In sentences, the administration routes and units after numerical data are important dosing tags. As these tags are highly compact, they usually occur together. For example, the following two dosing related text segments

- Midazolam oral (15 mg) and intravenous (0.05 mg.kg-1) was given
- 7.5 mg dose of midazolam was given orally

are tagged as

- <DRUG : Midazolam> <Dose_A : oral> (15 <Dose_U : mg>) and <Dose_A : intravenous> (0.05 <Dose_U : mg.kg-1>) was given

- 7.5 <Dose_U : mg> dose of <DRUG : Midazolam> was given <Dose_A : orally>.

PK Parameter Tagging

Drug clearance is chosen as the test PK parameter data mining performance, since it has comparably more numerical data available in the abstracts. The important tags for the clearance relevant value and unit are,

- clearance terms (T) : [systemic / oral] clearance
- Value (V)
- Unit (U) examples : ml/min/kg; l/kg/h; ml/min; l/hr …

As there are two types of clearance, systemic clearance and oral clearance, a type classification is needed in the following data analysis step. The clearance value is usually reported in both sample mean and standard deviation. The co-existence of the clearance key-terms and units is a unique identification, and the tagging is done by identifying them together in one sentence. For example, the phrase "the systemic clearance of midazolam was unchanged (37.7 +/- 11.3 l/h)" is tagged as "the <CLR_T : systemic clearance> of <DRUG : midazolam> was unchanged (<CLR_V : 37.7+/-11.3> <CLR_U : l/h>)".

After the tagging process, the relevant elements in each sentence are recognized. The tagged sentences in each abstract are kept for the following information extraction. All the untagged sentences are removed.

3.2.3 Information Extraction

In this stage, we need to extract the information from three prescribed tagging items: dosing, subject, and PK parameters. The subject and dosing information can be extracted easily given a well tagged sentence. For example, given the tagged phrase "<SUB_N : seven> <SUB_D : healthy> control <SUB_T : subjects>", the machine easily locates the subject information. Similarly, the tagged phrase "7.5 <Dose_U : mg> dose of <DRUG : Midazolam> was given <Dose_A : orally>" clearly shows "7.5mg orally" as the dosing information for midazolam. The tagged sentence "<DRUG : Midazolam> <Dose_A : oral> 15 <Dose_U : mg> and <Dose_A : intravenous> 0.05 <Dose_U : mg.kg-1> " indicates a simple sequential parsing of information for oral dosing and intravenous dosing as "oral 15mg; intravenous 0.05 mg.kg-1".

PK parameter data extraction is more complicated. As multiple drugs are usually involved in the PK studies, one abstract sentence may contain PK data for both target drug and other drugs. Even if one sentence discusses the target drug only, the data can reflect its PK value change caused by other study drugs. The following sentence reflects this complexity,

> *Rifampin significantly (P<.0001) increased the systemic and oral clearance of midazolam from 0.44+/- 0.2 L. h/kg and 1.56 +/- 0.8 L x h/kg to 0.96 +/- 0.3 L x h/kg and 34.4 +/- 21.2 L x h/kg, respectively.*

Two drugs, midazolam and rifampin, are mentioned in this sentence, and the clearance values contain both control and affected cases. The information extraction needs to make the correct decision that this sentence discusses midazolam, but not rifampin; and the control clearance values come first (0.44+/-0.2 for systemic clearance; 1.56+/-0.8 for oral clearance). There are two steps to discriminate the target drug. First, if the title or the occurrence frequency of term "midazolam" shows strong signal that the abstract is about midazolam but not rifampin, this sentence is most likely to be midazolam. Secondly, it is still possible that one clearance value is for rifampin for the sake of comparison. In order to deal with this case, a set of extraction rules are created. The rules are explained in detail in the follow up example. After the tagging step, this sentence example is converted to

> *<DRUG : Rifampin> significantly (P<.0001) <CHG : increased> the <CLR_T : systemic> and <CLR_T : oral clearance> of <DRUG : midazolam> from <CLR_V : 0.44+/- 0.2> <CLR_U : L. h/kg> and <CLR_V : 1.56 +/- 0.8> <CLR_U : L x h/kg> to <CLR_V : 0.96 +/- 0.3> <CLR_U : L x h/kg> and <CLR_V : 34.4 +/- 21.2> <CLR_U : L x h/kg>, respectively.*

The tag "<CHG>" is an important one to show the change of clearance value caused by the co-existence of other drugs. Hence, the "increased" case of <CHG> tag, the smaller value of clearance data is usually the control, i.e. study with no drug interaction effect, which should be extracted. Now the simple representation pattern for this sample sentence is "<DRUG1> <CHG> <CLR_T1> <CLR_T2> <DRUG> <CLR_V1> <CLR_V2> <CLR_V3> <CLR_V4>". The rules to extract clearance information for this type of pattern are listed below,

- Find clearance type <CLR_T1> <CLR_T2>.
- Find value change type <CHG>.
- Each value change involves two clearance values for one clearance type, hence there should be four clearance values (<CLR_T1_V1> to <CLR_V4>).
- The clearance values for <CLR_T1> can be (<CLR_V1> <CLR_V2>) or (<CLR_V1> <CLR_V3>). Choose the pair with the smaller difference, and the smaller value in that pair is <CLR_T1>. For example, the systemic clearance is 0.44 +/- 0.2 L. h/kg.
- The other two values are for <CLR_T2>. Similarly, the smaller value of the two is chosen for <CLR_T2>. For example, the oral clearance is 1.56 +/- 0.8 L x h/kg.

These extraction rules cover regular expressions of clearance data, considering single and multiple drug occurrences, different clearance types, and clearance value changes.

3.2.4 Linear Mixed Model Meta-Analysis for Outlier Detections

Because the mined PK parameter numerical data may contain some false positive values, an evaluation mechanism is needed to remove them as outliers. The population mean and variance of PK parameters are also requested to be estimated. We developed a linear mixed model meta-analysis approach for this purpose. The PK parameter values are assumed to follow the normal distribution as illustrated in Eq. (1).

$$\begin{aligned} &\bar{\theta}_{\cdot k} \sim N(\theta_k, se_k^2) \\ &\theta_k \sim N(\theta, \sigma^2) \end{aligned} \qquad (1)$$

The first normal distribution is at the study level, in which $\bar{\theta}_{\cdot k}$ (the sample mean of study *k*) has study-specific mean θ_k , sample standard error se_k^2, where *k = 1*,…,*K*, indicates the studies. The second normal distribution is at the population level, in which θ_k has the population mean θ , and σ^2 is its between-study variance. The population and study level PK parameters are two common statistics concepts in the pharmacokinetics meta-analysis literature [191]. The population PK parameter refers to its population-average mean, and a study-specific PK parameter refers to its sub-population mean, in which the study was sampled from. In this paper, we assume that PK data from one paper is a study, which is denoted by *k*.

In Eq. (1), $\bar{\theta}_{\cdot k}$ and se_k^2 are observed data from the literature mining results. The unknown parameters θ, σ^2 and θ_k are estimated by the following expectation and maximization algorithm. The expectation step estimates θ_k by Eq. (2).

$$\hat{\theta}_k = [\frac{1}{se_k^2} + \frac{1}{\sigma^2}]^{-1} \cdot [\frac{\bar{\theta}_{\cdot k}}{se_k^2} + \frac{\theta}{\sigma^2}] \qquad (2)$$

The values of population mean θ and population variance σ^2 are estimated in the maximization step by Eq. (3). The E-M iterative procedure stops when the estimated values are stable.

$$L(\theta, \sigma^2, \theta_k \mid \bar{\theta}_{\cdot k}, se_k^2) \propto \prod_k N(\theta_k, se_k^2) \cdot N(\theta, \sigma^2)$$

$$\frac{\partial}{\partial \theta} \log L = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{\sum_{k=1}^{n} \theta_k}{n} \qquad (3)$$

$$\frac{\partial}{\partial \sigma^2} \log L = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{\sum_{k=1}^{n} (\theta_k - \theta)^2}{n}$$

Based on the meta-analysis, the standard error of the estimated population mean is expressed in Eq. (4),

$$se(\theta) = \sqrt{\frac{1}{\sum_{k=1}^{K} \frac{1}{(\sigma^2 + se_k^2)}}} \qquad (4)$$

3.2.5 Validation and Classification

Some PK parameters have multiple types, but the abstracts do not always state clearly which type a numerical data refers to, e.g. some MDZ abstracts just use a single word “clearance” to represent either systemic clearance or oral clearance. In order to classify the unknown clearance type, the probability functions are established from known oral and intravenous clearance data with prescribed linear mixed model. Denote them as, $P[\bullet \mid \theta_{PO}, se_{PO}^2, \sigma_{PO}^2]$ and

$P[\bullet \mid \theta_{sys}, se_{sys}^2, {\sigma_{sys}}^2]$, respectively. For an unknown type sample mean clearance value, $\overline{\theta}_{\cdot k,unknown}$ is classified by Eq. (5).

$$\begin{cases} P[\overline{\theta}_{\bullet k,unknown} \mid \theta_{PO}, se_{PO}^2, {\sigma_{PO}}^2] > P[\overline{\theta}_{\bullet k,unknown} \mid \theta_{sys}, se_{sys}^2, {\sigma_{sys}}^2] \Rightarrow type = PO, \\ P[\overline{\theta}_{\bullet k,unknown} \mid \theta_{PO}, se_{PO}^2, {\sigma_{PO}}^2] < P[\overline{\theta}_{\bullet k,unknown} \mid \theta_{sys}, se_{sys}^2, {\sigma_{sys}}^2] \Rightarrow type = Systemic. \end{cases} \quad (5)$$

Then a leave-one-out strategy is implemented to validate the classification results. For each classified data set, one single data is taken out and the rest go through the prescribed linear mixed model meta-analysis. If this left-out data is 2.5 standard deviations from the population mean, it is considered as an outlier. To save the computation time, the data are ranked first, and this leave-one-out process is conducted iteratively from both the bottom and the top of the ranked data until the left-out data is not considered as outliers.

3.3 Abstract Mining Results

3.3.1 Evaluation of Each Mining Step

3.3.1.1 Entity Recognition

In this thesis, midazolam (MDZ) is used to test our literature mining strategy. The key-term “midazolam” in PubMed search returns over 7,129 article records. After applying the entity template, out of the 7,129 PubMed abstracts, 393 abstracts are considered as MDZ PK relevant. Among those 393 abstracts, 170 are determined truly relevant after being manually checked by author and validated by a PhD pharmacist. Thus, precision is 43%.

3.3.1.2 Information Extraction

From 393 abstracts, the information extraction returns 53 abstracts. 43 out of 53 abstracts contain true MDZ clearance data. Hence the precision improves to 81%. As actual MDZ clearance data in the 7,129 abstracts is unknown, we did not calculate its recall here but perform a thorough performance analysis in a separate section (3.3.2). The same information extraction rules are also applied directly to the starting 7,129 PubMed abstracts. It returns 120 abstracts, and a much lower precision, 36% (dashed lines in Figure 3.2). This analysis shows the importance and the power of the entity template step.

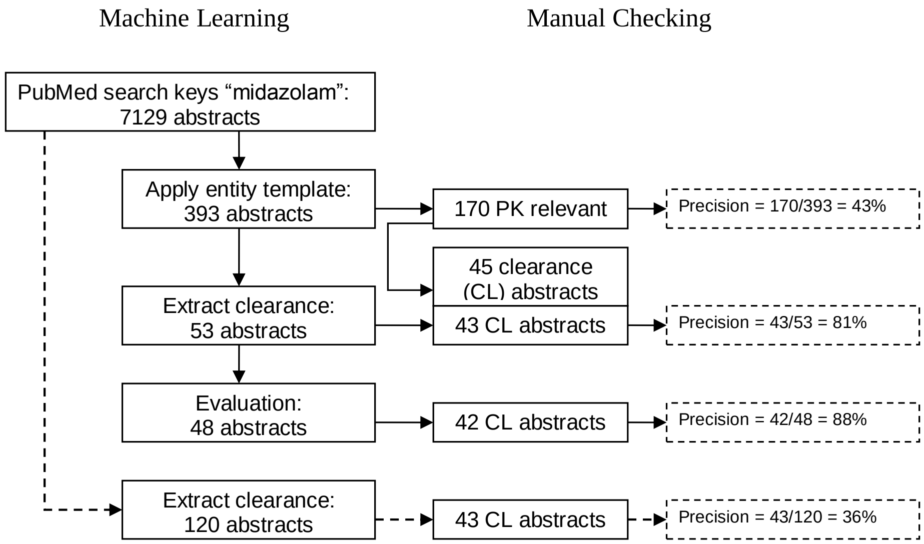

Machine Learning
Manual Checking
PubMed search keys “midazolam”: 7129 abstracts
Apply entity template: 393 abstracts
170 PK relevant
Precision = 170/393 = 43%
45 clearance (CL) abstracts
Extract clearance: 53 abstracts
43 CL abstracts
Precision = 43/53 = 81%
Evaluation: 48 abstracts
42 CL abstracts
Precision = 42/48 = 88%
Extract clearance: 120 abstracts
43 CL abstracts
Precision = 43/120 = 36%

Figure 3.2 : Precision Performance Analysis of the Machine Learning Algorithm in all MDZ Related Abstracts

3.3.1.3 Large Scale Evaluation

Linear mixed model meta-analysis is implemented to classify the oral and systemic clearances, and remove the outlier data and abstracts. After this evaluation step, only 48 final abstracts are left, and 42 of them are true (precision 88%). The precision of the mining goes from 43% in entity recognition to 81% in clearance data extraction, and reaches 88% after evaluation (see figure 3.2). A comprehensive performance analysis on a constructed test data set is provided in a separate section (3.3.2).

3.3.1.4 Midazolam Clearance Parameter Estimation and Outlier Detections

MDZ PK clearance data from information extraction are shown in the first row of Table 3.2. The mined clearance data have three types: oral clearance, systemic clearance and clearance with unknown mechanisms. The values are normalized based on an estimated average human body weight 80kg, and verified by manually going through the abstracts. False positive clearance data are labeled in red.

The mined clearance data are then fed to the linear mixed-model meta-analysis to estimate the distributions for the systemic/oral clearance and remove the outliers. The calculated distributions are displayed in Figure 3.3. The population mean ± se of systemic clearance is 27.8 ± 1.0 L/Hour, and its between-study standard deviation is 7.31; oral clearance is 78.1 ± 6.0 L/Hour, and its between-study standard deviation is 32.8.

Based on the distributions, the unknown type of clearance data were classified into oral clearance or systemic clearance, and outliers were removed. After the evaluation process, the final mined MDZ clearance data was shown in the second row of Table 3.2. The evaluation removes most of the false positive data. The left false positive data are comparable to the true clearance data, and they cannot be identified as outliers. Some true MDZ clearance data, labeled blue in the first row of Table 3.2, are considered as outliers by the evaluation. Figure 3.4 shows all mined MDZ clearance data before evaluation and outlier removal (a), compared with the MDZ clearance data after outlier removal (b). Obviously the meta-analysis can efficiently classify the data and remove the outliers.

Table 3.2 : Mined and Validated MDZ Clearance Data.

The mined clearance data have three types: oral, systemic and unknown type. The false positive data was labeled *red*; the false negative data which was removed in the validation step was labeled *blue*.

	Oral	Systemic	Unknown
Mined Clearance Data (L/Hour)	0.72, 4.9, 8.2, 31.98, 42.6, 43.2, 52.32, 68.64, 84.78, 109.2, 116.8, 124.8, 137, 152, 215.9, 1289	15.12, 18.6, 22.98, 28, 32, 33.06, 33.6, 35.2, 36.9, 37.7, 77.28, 84.78	0.81, 1.14, 2.016, 2.11, 2.26, 2.4, 3, 4.6, 5.58, 6.6, 14.94, 15.9, 16.75, 16.98, 19.02, 19.38, 19.5, 20.16, 21.12, 21.5, 22.2, 22.56, 23.28, 23.3, 23.4, 23.5, 23.52, 23.664, 23.94, 24, 24.8, 25.14, 25.2, 25.86, 25.92, 27.024, 27.78, 28, 28.2, 28.8, 28.96, 29.904, 30.12, 30.64, 32.16, 33.78, 34.08, 36.77, 36.96, 37.44, 37.92, 38.88, 39.17, 39.22, 40.8, 42.4, 45.12, 45.6, 46.08, 51.2, 52.8, 53.8, 54.6, 54.72, 58.56, 59.04, 59.2, 66.24, 78.6, 97.5, 99.36, 132, 144, 146, 166.56, 1281, 2272, 3328, 5472, 17616
Clearance After Evaluation (L/Hour)	Oral		Systemic
	42.4, 42.6, 43.2, 45.12, 45.6, 46.08, 51.2, 52.8, 53.8, 54.6, 54.72, 58.56, 59.04, 59.2, 66.24, 68.64, 78.6, 97.5, 99.36, 109.2, 116.8, 124.8, 132, 137, 144, 146, 152, 166.56		14.94, 15.12, 15.9, 16.75, 16.98, 18.6, 19.02, 19.38, 19.5, 20.16, 21.12, 21.5, 22.2, 22.56, 22.98, 23.28, 23.3, 23.4, 23.5, 23.52, 23.664, 23.94, 24, 24.8, 25.14, 25.2, 25.86, 25.92, 27.024, 27.78, 28, 28, 28.2, 28.8, 28.96, 29.904, 30.12, 30.64, 32, 32.16, 33.06, 33.6, 33.78, 34.08, 35.2, 36.77, 36.9, 36.96, 37.44, 37.7, 37.92, 38.88, 39.17, 39.22, 40.8, 42.4, 45.12, 45.6, 46.08, 51.2

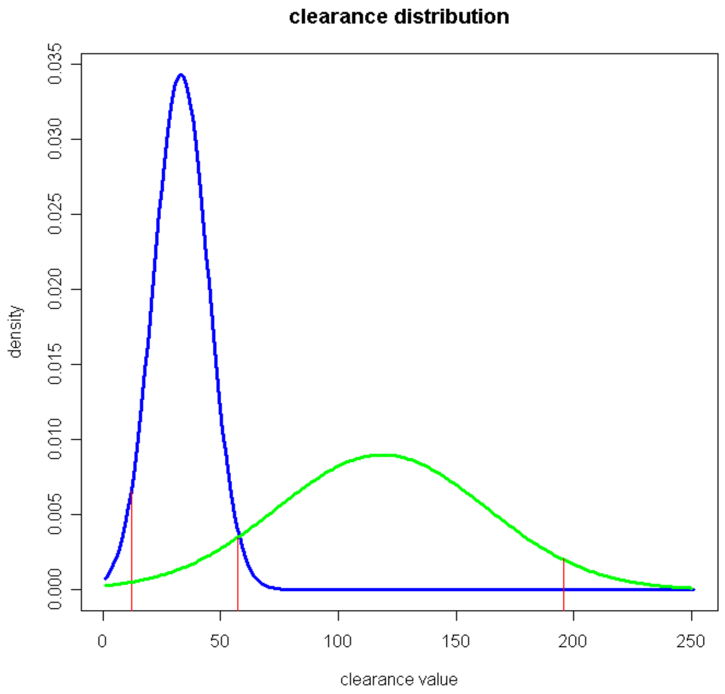


Figure 3.3 : Estimated Clearance Distribution

The *BLUE* curve shows systemic clearance; the *GREEN* curve shows oral clearance. The 95% confidence interval is marked on each curve using vertical lines.

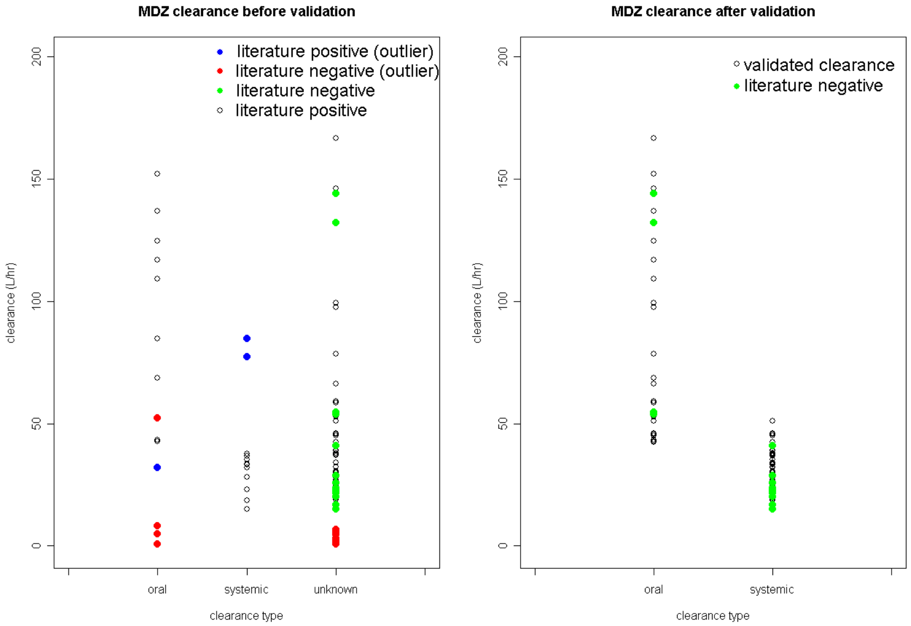


Figure 3.4 : MDZ Clearance Data

(a) contains all mined MDZ clearance data before evaluation and outlier removal, and (b) contains the MDZ clearance data after evaluation outlier removal. The *BLUE* dots are true clearance data from MDZ PK relevant abstracts; the *RED* and *GREEN* dots are false MDZ clearance data, in which the red ones were removed by EM validation as outliers and green ones were not.

3.3.2 Performance Evaluation on Constructed Test Data

3.3.2.1 Validation Data Generation

The classical way to evaluate the performance of information retrieval is to check its recall and precision. In this case study, the quality of the entity template determines how well the MDZ PK relevant abstracts can be retrieved. However since the sample data set (over 7,000 abstracts) from PubMed search is too big to be handled manually for the recall and precision analyses, a subset of the abstracts are generated to estimate the performance of each literature mining step. To build such a subset, one more key-term "pharmacokinetics" is included into the PubMed search. This decreases the size of the result abstracts to 819, a reasonable number for the manual performance check. The results are shown in Figure 3.5. The manual inspection of the 819 abstracts returns 164 PK relevant articles for drug MDZ. This figure shows the whole test design. The template based IR/IE was compared with SVM based IR/IE, and direct IE (more details are given in the following result sections). The template based IR/IE has shown competitive performance, precision and recall, in both IR and IE. This method can be applied in both relevant abstract collection and PK parameter extraction. However, the success of this method relies heavily on the quality of the manually constructed template. O~~I~~n the other hand, the SVM 100 IE results are also very competitive for an automated TM pipeline, from both a precision and recall viewpoint. It provides an effective substitution for PK parameter extraction even in the presence of no human expertise, under the assumption that a collection of positive and negative papers were given.

3.3.2.2 Entity Recognition

After applying the entity template, 220 out of the 819 abstracts are left in which 150 abstracts are truly relevant. The recall of this information retrieval step is 91% and the precision is 68% (Table 3.3). To evaluate the power of this entity template, we compare the performance of template based abstract classification with an automatic classifier implemented using a support vector machine (SVM). Training data were established by dividing the 164 relevant abstracts into three groups with about 55 abstracts in each, then adding to each group 55 randomly selected irrelevant abstracts. The group which generates higher F-score was recorded as SVM50. We applied a two-step process to determine proper features for SVM. First, a chi-square based feature selection filter was used to retain all features with the p-value below threshold 0.05. Then, the remaining features went through a principle component analysis [32] for dimensionality reduction, which was set to keep a cumulative proportion 95% of the original features. The final features were fed into SVM for model training and classification. We also tried a second training data set (SVM100), which was made up of 100 randomly selected abstracts from the 164 relevant articles and 100 randomly selected irrelevant abstracts. The SVM^{light} [192] was implemented with different kernels, and the best performance was shown in Table 3.3. SVM50 achieved higher precision. To further evaluate the potential of SVM, a 3-fold cross-validation method was applied on the 819 abstracts, which shows an average precision of 0.841 and recall of 0.562. So after training~~When~~ the SVM model ~~is trained~~ on unbalanced positive (2/3 of 164 abstracts) and negative (2/3 of 655 abstracts) data sets, its precision can be further improved (from 0.692 to a mean of ~~0.795~~0.841).

Comment [LMR1]: This is not clear on the table… where you show three fold results, one of which has 100%! The mean value of the 3 folds is 84.1… So what are you reporting exactly?

3.3.2.3 Information Extraction

For the clearance data, the manual inspection proves 39 out of the 164 relevant abstracts containing MDZ clearance numerical values (clearance relevant). Our information extraction step recognizes 45 abstracts as clearance relevant, in which 37 are true. Hence, the recall rate for clearance data extraction is 95% and the precision is 82%. The same information extraction rules are also applied directly to the starting 819 abstracts (**Error! Reference source not found.**). Without the application of entity template, the precision drops from 82% to 38%, and F-score reduces from 88% to 55%.

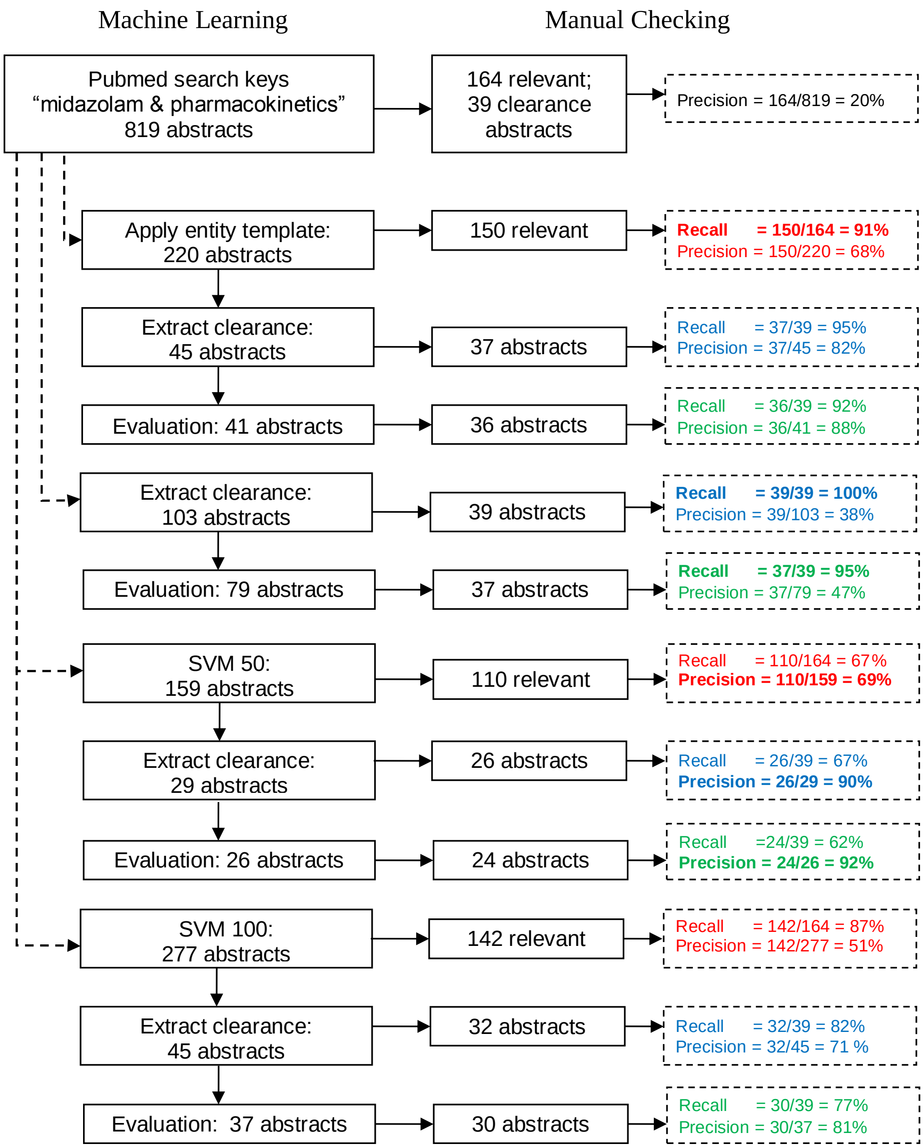


Figure 3.5 : Recall and Precision Performance Analysis of the Machine Learning Algorithm in a MDZ Abstracts Subset

Table 3.3 : Abstract Classification by Template and SVM on MDZ

The training data of SVM (50) contains 50 randomly selected relevant abstracts and 50 irrelevant; the training data of SVM (100) contains 100 randomly selected relevant and 100 irrelevant. (TP, FP, FN, TN) stand for true positive, false positive, true negative, and false negative, respectively.

MDZ-Relevance											
Method	Total	Query	TP	FP	FN	TN	Precision	Recall	F-Score	Accuracy	MCC
PubMed Query	NA	819	164	655	NA	NA	20.0%	NA	NA	NA	NA
PubMed Query & Entity Template	819	220	150	70	14	585	68.2%	**91.5%**	**78.1%**	**89.7%**	**0.73**
SVM (50)	819	159	110	49	54	496	**69.2%**	67.1%	68.1%	74.0%	0.59
SVM (100)	819	277	142	135	22	520	51.3%	86.6%	64.4%	80.8%	0.56
Entity Template	228	113	104	9	10	105	**92.0%**	91.2%	91.6%	91.7%	0.83
SVM (50)	228	84	68	16	46	98	81.7%	59.3%	68.7%	73.0%	0.47
Entity Template	128	63	58	5	6	59	**92.1%**	90.6%	91.3%	91.4%	0.83
SVM (100)	128	46	40	6	24	58	88.6%	61.9%	72.9%	77.0%	0.55
SVM (3 fold)	273	59	43	16	11	203	72.9%	79.6%	76.1%	90.1%	0.70
	273	44	35	9	19	210	79.5%	64.8%	71.4%	89.7%	0.66
	273	13	13	0	41	219	100%	24.1%	38.8%	85.0%	0.45

Clearance-Relevance

Table 3.4 : Clearance Extraction With and Without Entity Template

Method	Total	Query	TP	FP	FN	TN	Precision	Recall	F-Score	Accuracy	MCC
PubMed Query + CL Extraction	819	103	39	64	0	716	37.9%	**100.0%**	54.9%	92.2%	59.0%
PubMed Query + **Entity Template** + CL Extraction	819	45	37	8	2	772	82.2%	94.9%	**88.1%**	**98.8%**	**87.7%**
PubMed Query + **SVM** (50) + CL Extraction	819	29	26	3	13	777	**89.7%**	66.7%	76.5%	98%	76.4%
PubMed Query + **SVM** (100) + CL Extraction	819	45	32	13	7	767	71.1%	82.1%	76.2	97.7%	76.0%
PubMed Query + CL Extraction + Outlier Evaluation	819	79	37	42	2	744	46.8%	**94.9%**	62.7%	94.7%	64.6%
PubMed Query + **Entity Template** + CL Extraction + Outlier Evaluation	819	41	36	5	3	775	87.8%	92.3%	**90.0%**	**99.0%**	**89.5%**
PubMed Query + **SVM** (50) + CL Extraction + Outlier Evaluation	819	26	24	2	15	778	**92.3%**	61.5%	73.8%	97.9%	74.4%
PubMed Query + **SVM** (100) + CL Extraction + Outlier Evaluation	819	37	30	7	9	773	81.1%	76.9%	78.9%	98.0%	78.0%

3.3.2.4 Evaluation

The meta-analysis evaluation removed most outliers and false positive values. After this step, the clearance data from 41 abstracts are left and 36 of the abstracts are true MDZ clearance relevant. The recall rate becomes 92% and the precision is improved to 88%. Similarly, without the entity template step, both the F-score and precision drop significantly (**Error! Reference source not found.**), from (90%, 88%) to (70%, 53%).

3.3.2.5 Comparison of Midazolam Data Mining and Its Validation Analysis

Figure 3.2 and Figure 3.5 show the PK information comparison between single PubMed search key-term (“midazolam”) and two key-terms (“midazolam” and “pharmacokinetics”). Though the

PubMed search returns much more abstracts using a single key-term than using two key-terms (7129 vs. 819), only six more relevant abstracts are found in the single term search results (170 vs. 164). The difference of the number of clearance relevant abstracts is also six (45 vs. 39).

3.4 Abstract Mining Contributions

3.4.1 Compare Entity Template with Automatic Abstract Classification

To evaluate the power of this entity template, we compare its performance with SVM in both abstract classification and information extraction (Table 3.3, **Error! Reference source not found.**). The precision/recall is measured on information retrieval, finding relevant articles out of the test set of abstracts (Table 3.3). SVM (50) has slightly higher precision than our entity template in identifying MDZ relevant abstracts (69.2% vs. 68.2%), but worse recall (67.1% vs. 91.5%). Hence SVM (50)'s F-score is lower than entity template (68.1% vs. 78.1%). On the other hand, SVM (100) generates reduced precision, 51.3%, and improved recall, 86.6%. Its F-score becomes even worse, 64.4%. Overall, the entity template out-performs SVM in recall, F-score, accuracy and MCC score. Thus we choose entity template over SVM as abstract classification method. However, in a scenario the sacrifice of missing PK data is affordable, SVM outweighs entity template by obtaining higher precision. To further explore the advantage of SVM in aspect of precision, we apply SVM50 and SVM100 on *non-contaminated* test data sets which have no intersection with training data (i.e. using complementary relevant abstracts mixed with same number of irrelevant abstracts). It turned out the precision is quite impressive (81.7% and 88.6%), but their overall performance still falls behind of entity template.

In clearance extraction using abstracts from entity template and SVM classification, the performance comparison (**Error! Reference source not found.**) shows a similar pattern as in IR results (Table 3.3) except the less recall of entity template compared with direct extraction. Extraction from entity template out-performs SVM in recall, F-score, accuracy and MCC score but SVM50 leads in precision. Though entity template based extraction shows a superior overall performance compared with direct extraction and SVM based extraction, the choice of a certain mining strategy largely depends on the purpose of mining. As SVM shows higher precision, it is preferred in the fast retrieval of a small set of relevant articles and PK parameter data. However, if the target drug has very sparse data published, this method might not work effectively. Similarly, the direct extraction method can retrieve PK data fast and provide a quick reference for the range of PK data. However, its lower precision limits its efficiency to provide a proper set of articles as reference for the target drug because a certain amount of further work is still needed for precise classification, especially when the target drug is well studied hence coming with a large number of publications. Thus for our PK parameter collection study, we prefer the method of entity template based extraction, which shows second highest precision and second highest recall. However, under the circumstances that no human expertise can be consulted to build the template, SVM 100 also shows competitive performance from both precision and recall which indicates a promising option as an automated TM pipeline.

3.4.2 Information Content Comparison with DiDB

To better evaluate our literature mining method, we compare the extracted MDZ clearance data with those from DiDB database. DiDB [8] is the most complete PK database so far, which is built manually. DiDB MDZ clearance data are downloaded and are compared with the mining

MDZ clearance data. Table 3.5 lists detailed comparisons. DiDB provides 11 PK relevant articles for MDZ. We read through their abstracts and found only six clearance data records from relevant abstracts for healthy subjects. While the PubMed mining returned 170 PK relevant articles for MDZ, in which more than 70 clearance data records were extracted from the abstracts. Therefore, the literature mining method yields a 70/6=11.6 times fold increase in information content, in addition to the benefits of the automatic data extraction.

Table 3.5 : MDZ Clearance Comparisons among Known Data, DiDB, and Mining Results

This table shows the number of PK relevant articles (“relevant article #”) available, and number of clearance data records (“# of abstract PK”) extracted from abstracts.

	Manual			DiDB			Mining		
	# of Abstract PK	# of Relevant Article	$\theta \pm se$	# of Abstract PK	# of Relevant Article	$\theta \pm se$	# of Abstract PK	# of Relevant Article	$\theta \pm se$
Oral Clearance	25	170	83.6± 8.6	2	11	58.3± 16.8 (88.4± 7.3)†	28	170	78.1± 6.0
Systemic Clearance	50		32.3± 1.8	4		25.8± 3.1	59		27.8± 1.0

† After removing an outlier (publication error)

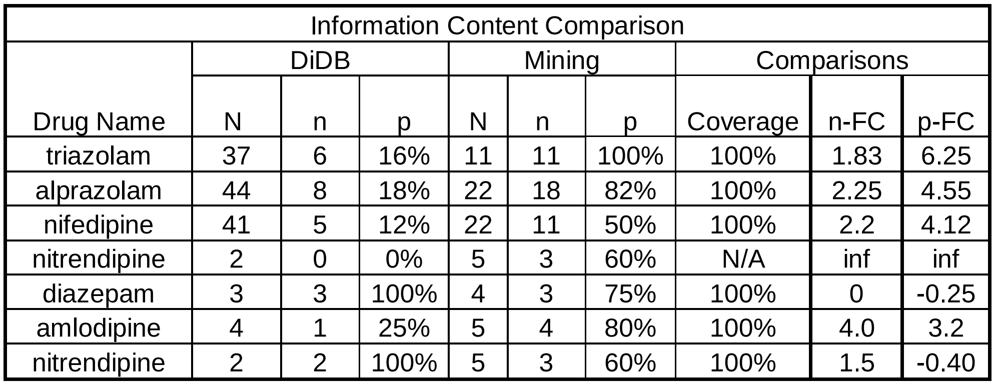
The true population mean and standard error ($\theta \pm se$) are benchmark, which come from manually accumulated clearance data from known relevant article abstracts. The population mean and its standard error are calculated for DiDB clearance data and the mined clearance data. For the oral clearance, the benchmark estimate is 83.6 ± 8.6 (L/Hour), while the DiDB and mining estimates are 58.3±16.8 and 78.1±6.0 respectively. Comparing to the benchmark, the DiDB estimate is much more biased than our mining approach, 30.3% vs. 6.6%; and DiDB estimate’s SE is 2.8 times higher than our mining approach. For the systemic clearance, comparing to the bench mark, DiDB estimate’s bias = (32.3-25.8)/32.3×100% = 20.1%, and mining estimate has a bias of 13.9%. DiDB estimate’s SE is 3.1 times higher than the SE of our mining estimate.

One observation on the DiDB oral clearance data is the influence of the publication errors on the data analysis. PubMed PID 15470333 reported oral clearance for midazolam as 533 +/- 759 mL/min by typo in the abstract. The correct value should be 1533 +/- 759 mL/min in the full text. In the meta-analysis of our text mining, the influence of such error is eliminated by the outlier detection. However, DiDB database suffers from this type of publication error, and we suspect that DiDB only reads the abstract sometimes.

Table 3.5 shows that our literature mining approach collects 11 times more MDZ clearance data than the manually curated DiDB database contains. To test the generalization potential of our literature mining method, we tried it on 7 other Cytochrome P450 3A Subfamily drugs and extracted their clearance data from PubMed abstracts as for Midazolam. The same drugs were also searched in DiDB database (Sept, 2008), and clearance data was also analyzed. The comparison is shown in Table 3.6. Among 5 out of 7 drugs, comparing to DiDB, literature mining generated 1.83 to 4.0 fold more information contents in CL, and precision increased three folds and higher. Among those two drugs that DiDB out-performed literature mining, our approach only missed two abstracts in total.

The impressive performance shows the great potential of text mining as a drug PK data curation tool. The high precision and recall score of our mining method even indicates the feasibility of TM based PK database construction. Furthermore, to make such a PK database more reliable, certain manual validation will be a very helpful supplement which makes existing relevant data repositories, e.g. DiDB, valuable reference.

Table 3.6 : CL Data Extraction on More Drugs: DiDB vs. Literature Mining



Information Content Comparison									
	DiDB			Mining			Comparisons		
Drug Name	N	n	p	N	n	p	Coverage	n-FC	p-FC
triazolam	37	6	16%	11	11	100%	100%	1.83	6.25
alprazolam	44	8	18%	22	18	82%	100%	2.25	4.55
nifedipine	41	5	12%	22	11	50%	100%	2.2	4.12
nitrendipine	2	0	0%	5	3	60%	N/A	inf	inf
diazepam	3	3	100%	4	3	75%	100%	0	-0.25
amlodipine	4	1	25%	5	4	80%	100%	4.0	3.2
nitrendipine	2	2	100%	5	3	60%	100%	1.5	-0.40

N: total number of reported abstracts in DiDB; and number of extracted abstracts from text mining.
n: clearance relevant abstracts.
p: precision = n/N.
coverage: the percentage of DiDB clearance relevant abstract covered by text mining approach.
n-FC: fold-change from DiDB to mining in clearance relevant abstracts, n.
p-FC: fold-change from DiDB to mining in precision, p.

3.5 Abstract Mining Conclusions

In this chapter, an approach to mine MDZ PK data was presented with an ***88% precision rate*** and ***92% recall rate***. A conventional data mining approach, SVM, is compared to this entity template approach. Though SVM shows higher precision, we prefer the higher recall and overall performance of our manually designed entity template for this type of data collection. This mining approach recollects ***11 times*** more MDZ clearance data than a manual accumulated DiDB database has. Interestingly, it also identifies a publication error of midazolam clearance data in the DiDB database. In addition, we also established the first validation set for more general data mining methodology development for PK data.

We further investigated this abstract mining approach in 7 more CYP3A substrate drugs. Among five out of seven drugs, comparing to DiDB, abstract mining generates 1.83 to 4.0 folds more information contents in CL, and precision increases from 3.2 to infinite folds higher. Among those two drugs that DiDB out-performs mining, our approach only misses totally two abstracts. Therefore, from the information content point of view, our data mining approach outperforms manual PK data curation. In the meantime, since we implemented statistical model based evaluation strategies for the mining data, our integrated approach can identify outliers for quality control (QC). As a side production of QC, we provide not only population PK parameter estimates of PK parameters, but also their variations estimates.

Chapter 4.

TEXT MINING OF FULL TEXT DOCUMENTS

The abstract mining provides a good starting point for the next step: mining drug PK data from full-text documents which usually contain much more PK numerical data (**Error! Reference source not found.**) as well as annotation information. We use articles, in addition to abstracts with no full-text available in PubMed, as the data set for drug PK data extraction. Drug Midazolam is still used as the test drug in this stage.

4.1 Revised Information Retrieval

One major modification in the IR step for full text mining is the criteria for determining relevance. In Chapter 3, for mining abstracts, we included abstracts about drug-drug interaction (DDI) studies as long as PK data was presented in the abstracts. However, PK data extraction from DDI studies usually needs to deal with complicated entity relationship recognition, and to avoid bias in the extracted PK data which comes from mismatch of PK parameters to subject drugs or the actual PK parameter values affected by interacting drugs. Therefore, we split DDI cases from this mining study. In this section, we will just focus on non-DDI studies extracted from full-text articles.

The abstract IR strategy starts with a manually constructed entity template, which aims at high recall so no sparse PK data can be missed. Furthermore, we also need to remove as much noise as possible from false positive articles for the following full-text mining steps, as false positive articles will not only increase the complexity of article collection but also include unnecessary irrelevant data into the mining step. In our abstract mining study, we have compared the performance of entity template with SVM in abstract relevance classification for drug MDZ (Table 3.3). In the full-text mining, we combined the specificity of template-based filtering with the generality of a machine learning method as a two-step IR system, which can be more portable for new mining criteria.

4.1.1 Text Pre-processing

After evaluating text processing tools which can be applied in the biomedical field, such as Porter stemming, Perl module Lingua::EN::Sentence, Stanford NLP, UMLS MetaMap [193], NLM Lexical Variant Generation tool Norm [194], and RxNormNorm [195], we realized there are no existing tools that can process text efficiently as well as recognize PK related entities correctly. Thus we decided to apply the widely used Stanford NLP tool (http://nlp.stanford.edu/) together with our PK ontology for this purpose. The advantage of Stanford NLP is its comprehensive functions in text processing, which include tokenization, sentence split, part-of-speech tagging, lemma recognition, name entity recognition, parsing etc. We also compared the template based IR for Porter stemming and the lemma annotation of Stanford NLP in SVM modeling, and it turns out the performance is comparable but lemma based template and filtering rules are easier for reading and understanding.

One concern in the text pre-processing step is the existence of special characters in text such as “&” and “*”. Usually such characters can be removed without affecting the mining results, however, the extraction of numerical data requires us to keep symbols which indicate the relationship between numbers such as “+/-“, “+-“, “/”. Special characters in text are processed as following:

- rewrite <, >, <=, >=, +/- to Slst Slgt Slse Slge Spsm (for easier test manipulation)
- replace “-“ ,“+”, and “/” with a blank space (e.g. “midazolam-ketoconazole”) unless it is followed by number (e.g. “mg min-1”, “ml.min-1 . kg-1 “)
- remove all other punctuations as they can be special symbols in regular expressions

Another concern is the entity abbreviation recognition. As most PK parameters use standard symbols (e.g. “clearance” to “CL”), this abbreviation recognition process is only applied on drug names, for example “Midazolam” is mentioned as “MDZ”, “Mid” and even “M” in the literature. This abbreviation recognition is quite specific, and the drug abbreviation format usually follows such a pattern: 1) its first appearance falls in a parenthesized expression right after the target drug name; 2) the abbreviation is made up of letters from the full drug name. So instead of using abbreviation analysis algorithm/tools mentioned previously (section 1.2.2), we developed a simple and fast drug name abbreviation detection tool by matching the above two abbreviation forms, which has been able to recognize all drug midazolam’s abbreviations correctly.

4.1.2 Hybrid Information Retrieval Method

We search PubMed using a target drug name, i.e. Midazolam, as the only query term. Then we apply a template library as a first filtering of the PubMed search results. Compared with the template library used in abstract mining, this template library is much less stringent and includes simply the key PK terms (i.e. the target drug name and PK parameters) included in abstract template library to guarantee article coverage (high recall). Precision improvements are handled by further filtering steps. As titles usually contain critical information about the theme and scope of publications, the filtered abstracts were subsequently screened using the following rules applied to titles:

1) remove abstracts with "author's transl" in title (no full text in English)
2) remove abstracts with animal mentioned in title (even with human mentioned, it is assumed to be about animal studies)
3) remove DDI studies if the title indicate clearly DDI relevance (see DDI rules in next paragraph; “control/comparison/correlation” are not used as DDI indicating terms), and target drug name is also mentioned in the title.
4) if the title indicates in-vitro ("human hepatocytes"...), patient (cancer, tumor...), preterm/infants/children/neonates, or pregnancy studies, check if relevant terms, e.g. ill patients vs. healthy subjects, are also mentioned; if not, remove it.
5) Keep the abstract only if target drug (e.g. midazolam/mdz) is mentioned in the title, with PK mentioned but without DDI.

After applying rules above on abstract titles, all the rest abstracts are kept for sentence based analysis, including modeling and prediction studies (with key terms in title: model, hypothesis, prediction, predict, Bayesian, NONMEM, SimCyp...), and Pharmacodynamics studies (e.g. Electroencephalographic…). In this section, we will just work on non-DDI articles. Therefore, we need to summarize DDI patterns and design rules to detect DDI types of abstracts. Such patterns and rules were also used when working on abstract sentences.

1) DDI patterns

  a) effect of [drug/food] on [pk] of [mdz] (effect can also be “influence”)
  b) effect of [drug/food] on [mdz] ([pk])
  c) [drug/food] effect on [mdz]
  d) Effect of [mdz] on (no other [drug/food] before [mdz])
  e) Interaction between/of [drug/food] and [mdz] (vice-versa)
  f) ([pk]) interaction of [mdz] with [drug/food] (vice-versa)
  g) Drug interaction(s)
  h) [drug/food] | [mdz] | [induction/inhibition] (recognize induce, inhibit, inhibitory etc)
  i) [drug/food] impair/affect/influence/impact/reduce/increase [pk] of [mdz] (or [mdz] [pk])
  j) Note: [pk] can be any PK parameters, such as clearance, elimination, bioavailability, concentration etc;

2) Remove “[drug/food] [interaction|prediction|effect]” type of titles, ONLY if [mdz] does not appear in title
3) Relevant standards ([mdz] only with [pk] mentioned, after going through rules above)
  a) [pk] of [mdz] (kinetics, pk, elimination, absorption, half-life …)
  b) [mdz] [pk]

Though titles contain key information about an article, the filtering process can be biased by focusing on key terms and rules only. This is why we made the filtering standards very specific, removing highly irrelevant abstracts while keeping all possibly relevant abstracts for the following sentence based retrieval classification. Similarly as title filtering, the abstract sentences were also filtered by a set of rules:

1) Relevant abstracts must contain at least one PK relevant sentence, containing both target drug name and PK parameters; trace back to previous sentences for a drug name if only PK parameters are mentioned in a sentence.
2) Check the PK relevant sentences to make sure it is not about DDI. In addition to the DDI rules applied for titles, also consider:
    a) DDI specific key terms: inhibit, induct, ki, K(i), IC(50)…
    b) Words about PK change such as changes/measurements changes/increase/decrease/reduce/enhance/alter/affect/studied/assess/different/difference/observe/determine/estimate/measure are allowed for relevant sentences if DDI terms, e.g. induction/inhibition, are not mentioned.

The rule-based sentence filtering is very similar to the one used for title flittering, the same strategy is applied at different levels. The rule design principle is also the same, keeping high recall at first while improving precision gradually. We need to make sure the articles removed are highly irrelevant, so the filtering rules cannot be very stringent. For this reason, a big portion of remaining abstracts are false positive (irrelevant). Thus we will do a final filtering using a machine learning method to further classify these abstracts. One contribution of this final filtering is to uncover features missed which can be used for classification. The way to classify a relevant abstract is similar as the rule based sentence filtering, an abstract is considered relevant only if it contains relevant sentences and no DDI sentences. The difference in the last filtering which is that the relevant sentences are detected by Conditional Random Field (CRF) [123] modeling instead of applying rules.

CRF modeling was used as the machine learning algorithm for the last sentence filtering step. Compared with SVM, CRF can adopt variant features which can contribute to classification, and easily define multiple role of features and utilize feature relationship for classification. The features chosen for CRF modeling include:

- Lemma of words
- POS tag
- NER tag
- Above features of three adjacent words upstream and downstream

Eventually, a classification model is trained on top of these features, which is label input sentences as PK relevant (*PPKP*) or irrelevant (*NPKP*). If any input sentence from an abstract is labeled as *PPKP*, the abstract is considered relevant.

4.1.3 Results and Discussions

We used *Midazolam* as test drug and also query term for PubMed search, which returned 9293 abstracts. The first step, template based filtering, was able to trim down the data set to 645 potentially relevant abstracts. By a manual check, these 645 abstracts contain 104 relevant ones and 541 irrelevant. Then we tested the hybrid IR method using these 645 abstracts. The performance is summarized in **Error! Reference source not found.**. By applying rule filtering alone, we were able to further screen out irrelevant articles and keep 218 potentially relevant one (103 true positive), which shows a precision 0.47 and recall 0.99. The following machine learning IR was tested on both CRF and SVM. The CRF model was trained on manually selected relevant and irrelevant sentences, 50 each. SVM was trained on 1/3 of known relevant abstracts with same number of randomly selected irrelevant abstracts from the 645 test abstracts.

Eventually, CRF was able to achieve a precision 0.59 and recall 0.99, while SVM shows a precision 0.57 and recall 0.77. So the hybrid IR of rule and CRF can gradually increase the precision with keeping a high recall, meeting our goal of keeping all relevant abstracts while removing as many irrelevant as possible.

To show the power of the hybrid IR method, we compared it with IR using CRF or SVM alone. It turned out CRF alone can still keep high recall (0.98) but its precision dropped to 0.24, and SVM also keeps its recall (0.77) but its precision lowered to 0.31. We also combined all these three methods and checked its performance; however, it did not show any improvement partly because of the recall drop. As SVM has shown an advantage of achieving high precision in previous abstract mining study, we further tested SVM using 3-fold validation method. After being trained on unbalanced positive/negative data sets, it shows a mean precision 0.57 (s.d. 0.20) and recall 0.19 (s.d. 0.06). One fold can even reach a very high precision (0.75) but overall recall values are still quite low. The precision performance SVM shows here is quite similar as in previous abstract mining, however, there is wide variation among each fold. This situation was caused by the high similarity of TP and FP articles and the small size of test data, i.e. any small misclassification can change the performance a lot. For example, in one fold with precision =0.75, there is a TP=6 and FP=2, thus even the number just changes 1, the precision changes 10%. So if several FP turns out to be very similar as TP and happen to fall into test data, they can cause big performance drop.

Table 4.1 : Information Retrieval Performance

Method		Query	TP	FP	TN	FN	Precision	Recall	F-score	Accuracy	MCC
Manual		645	104		541						
Hybrid	Rule	645	103	115	426	1	0.47	**0.99**	0.64	0.82	0.60
	Rule+CRF	645	103	71	470	1	**0.59**	**0.99**	**0.74**	**0.89**	**0.71**
	Rule+SVM	645	80	61	480	24	0.57	0.77	0.65	0.87	0.58
	Rule+CRF	645	88	61	480	16	0.59	0.85	0.70	0.88	0.64

	+SVM										
CRF		645	102	332	209	2	0.24	0.98	0.38	0.48	0.29
SVM		645	80	182	359	24	0.31	0.77	0.44	0.68	0.32

4.2 Article Collection

After checking the existing article repositories (e.g. PubMed Central, Google Scholar, OVID etc) and consulting with publication management professionals, we reached a conclusion that the most effective article collection method for our study would be to download a small set of articles for one drug at a time, from an open resource such as PubMed. This situation is mostly determined by the limitation of publication copyright and the lack of a centralized and comprehensive article repository. Therefore, our strategy is to use PubMed searches, then crawl full-text links on the PubMed search result page, and to retrieve articles from a journal that has full-text article available. We have developed downloading modules for most popular pharmaceutical/biomedical journals linked to PubMed.

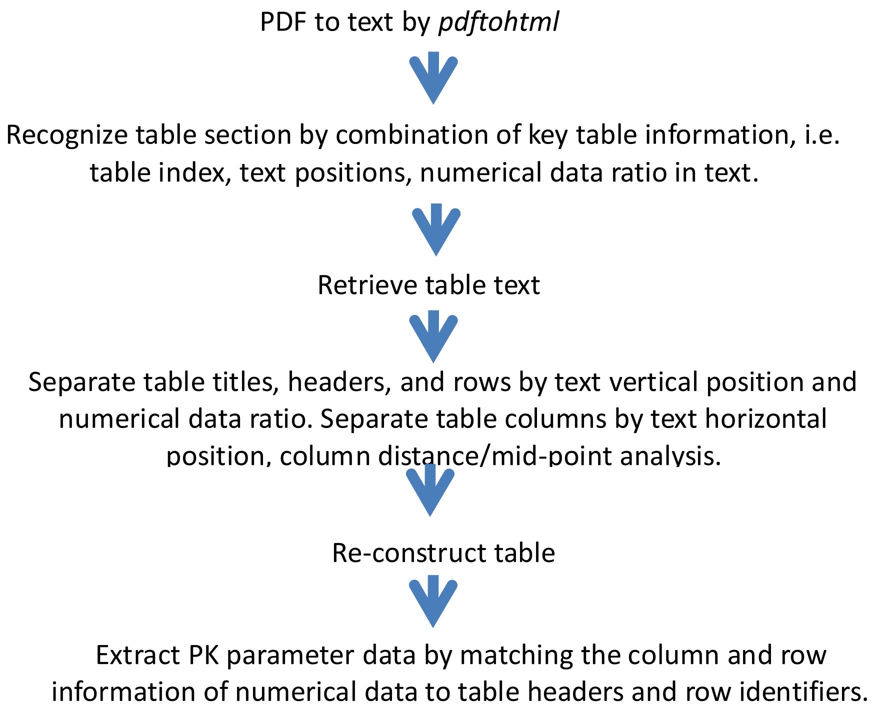
The demo website of our PubMed based PDF article downloading tool is available at http://rweb.biostat.iupui.edu/zhipwang/pubmed_pdf/. Using the drug Midazolam as a test, the manual full-text downloading of the 104 relevant articles, collected 59 PDF files, a full-text availability rate of about 60% (59/104) for drug MDZ through PubMed search results. Our crawler was able to download all of the 59 articles in PDF format. To achieve similarly high retrieval rate for other drugs, we have included crawling modules for most drug study related journals. Furthermore, this tool also has very good scalability, easily including missed articles from a new journal by adding corresponding downloading modules for this journal.

4.3 Tabular Data Extraction

4.3.1.1 Pharmacokinetics Data Extraction from PDF Tables

After the PDF articles were downloaded, a PDF processing tool, *pdftohtml* (http://pdftohtml.sourceforge.net/), was used to convert PDF files to XML format, which provides the physical position of each element/text of the PDF file. In turn, this position information was utilized to recognize a table segment. In addition to table index, most tables show significant position difference from the rest free text. We use this difference to recognize table segments from XML.

One unique feature of PK relevant tables is that the frequency of numerical data in a row (i.e. numerical data ratio) is over 50%. After table segments are separated, our table data extraction algorithm analyzes the table element position, column/row distance distribution, and numerical data frequency to recognize and reconstruct the original PDF table into free text format. Meanwhile, each numerical datum in a table cell is mapped to its corresponding row/column label if it is PK related. The whole procedure is summarized in Figure 4.1.



PDF to text by *pdftohtml*

Recognize table section by combination of key table information, i.e. table index, text positions, numerical data ratio in text.

Retrieve table text

Separate table titles, headers, and rows by text vertical position and numerical data ratio. Separate table columns by text horizontal position, column distance/mid-point analysis.

Re-construct table

Extract PK parameter data by matching the column and row information of numerical data to table headers and row identifiers.

Figure 4.1 : Table Data Extraction Procedure

To test the performance of this method, we used midazolam (MDZ) as the test drug and compared its PK parameter data from tabular data mining with manually extracted data from the same set of articles. Among the 59 downloaded PDF files, 34 articles contain tables of PK parameters. There are 111 tables total in these articles, out of which 42 tables are PK relevant for drug MDZ (true positive). Our method was able to recognize and reconstruct 40 relevant tables correctly with just one false positive table included. So the table retrieval recall is 0.95 and precision is 0.98.

To illustrate how the table was processed in our method, we used a table from PMID 10073325 [196] as an example. This table was recognized and reconstructed correctly by table data extraction component, as show in **Error! Reference source not found.** below. In the table re-construction process, table headers, columns and rows are identified separately. Thus each numerical data can be mapped to its corresponding PK parameter by column and row position, for example, “0.64^a (0.12)” is correctly mapped to “Premenopausal CL”.

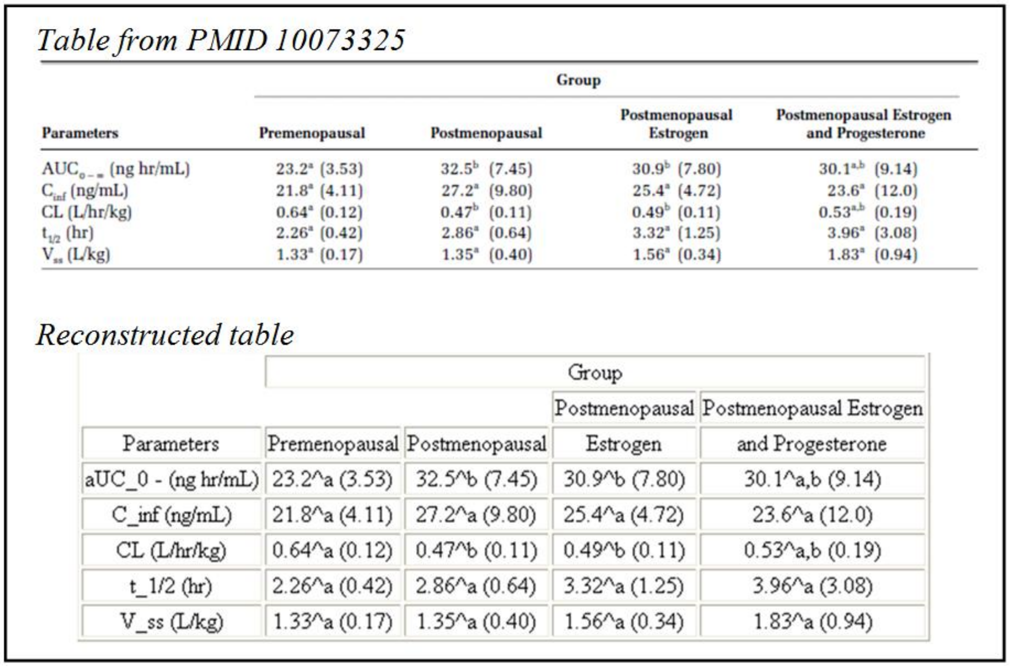


Table from PMID 10073325

Parameters	Group Premenopausal	 Postmenopausal	 Postmenopausal Estrogen	 Postmenopausal Estrogen and Progesterone
$AUC_{0-\infty}$ (ng hr/mL)	23.2[a] (3.53)	32.5[b] (7.45)	30.9[b] (7.80)	30.1[a,b] (9.14)
C_{inf} (ng/mL)	21.8[a] (4.11)	27.2[a] (9.80)	25.4[a] (4.72)	23.6[a] (12.0)
CL (L/hr/kg)	0.64[a] (0.12)	0.47[b] (0.11)	0.49[b] (0.11)	0.53[a,b] (0.19)
$t_{1/2}$ (hr)	2.26[a] (0.42)	2.86[a] (0.64)	3.32[a] (1.25)	3.96[a] (3.08)
V_{ss} (L/kg)	1.33[a] (0.17)	1.35[a] (0.40)	1.56[a] (0.34)	1.83[a] (0.94)

Reconstructed table

	Group			
			Postmenopausal	Postmenopausal Estrogen
Parameters	Premenopausal	Postmenopausal	Estrogen	and Progesterone
aUC_0 - (ng hr/mL)	23.2^a (3.53)	32.5^b (7.45)	30.9^b (7.80)	30.1^a,b (9.14)
C_inf (ng/mL)	21.8^a (4.11)	27.2^a (9.80)	25.4^a (4.72)	23.6^a (12.0)
CL (L/hr/kg)	0.64^a (0.12)	0.47^b (0.11)	0.49^b (0.11)	0.53^a,b (0.19)
t_1/2 (hr)	2.26^a (0.42)	2.86^a (0.64)	3.32^a (1.25)	3.96^a (3.08)
V_ss (L/kg)	1.33^a (0.17)	1.35^a (0.40)	1.56^a (0.34)	1.83^a (0.94)

Figure 4.2 : Table Reconstruction Example

PK data extraction performance was tested using six types of PK parameters (AUC, V, T1/2, Cmax, CL, Tmax) as shown in Table 4.2. The actual PK parameter data of each parameter was manually extracted from all 42 PK-relevant tables. Then we can calculate the performance of the tabular data extraction for each PK parameter. Most parameters show a high precision and recall, varying from 0.82 to 0.99 with an average precision 0.94 and recall 0.88. The F-scores are between 0.85 and 0.95, with an average value 0.91. Overall, this method of PK data extraction from tables has shown a strong performance.

Table 4.2 : PK Parameter Extraction from Tables for MDZ

PK	manual extraction	auto extraction				recall	precision	F-score
		TP	FP	TN	TN			
AUC	107	98	5	-	9	0.91	0.95	**0.93**
V	45	44	3	-	1	0.97	0.93	**0.95**
T1/2	77	64	8	-	13	0.83	0.88	0.85
Cmax	70	58	2	-	12	0.82	0.96	0.88
CL	127	114	1	-	13	0.89	0.99	**0.94**
Tmax	52	45	5	-	7	0.86	0.9	0.88

4.3.1.2 Information Content Comparison with Abstract

In the previous section, we tested the ability of our table processing component to recognize and re-construct table segment from transformed PDF files. Furthermore, drug PK database can be extracted precisely upon proper table reconstruction. This method of PK data extraction from tables has shown a very promising performance. To evaluate the contribution of extracted tabular information, we compared PK clearance data obtained from tables with that from the previous abstract mining study (Table 3.5). Even though CL data was only extracted from 59 PDF files, we are able to get a higher number (127) of CL data compared with the 87 (28+59) CL data from abstract mining. The increased number of CL data has shown the necessity of tabular data extraction from articles. The necessity of full text mining can be better illustrated by analyzing the PK data distribution in different parts of an article, i.e. abstract, tables and text, as shown in next section.

4.4 Pharmacokinetics Information Extraction from Full Text

4.4.1 Pharmacokinetics Data Distribution

To evaluate the PK information distribution of full text in the view of PK mining, we manually checked the information content within full text to see if information increase is possible from full text mining. We did this evaluation on the downloaded 59 PDF articles for drug MDZ using same PK parameter categories as in Table 4.2. The PK data distribution is summarized in **Error! Reference source not found.**.

Table 4.3 : PK Parameter Distribution Statistics

PK		**# of data in abstracts**		**# of data in tables**		**# of data in text**	**# of articles with table**	**# of articles w/o table**
AUC		4		**107**		5	0	3
V		5		**45**		3	1	1
T1/2		7		**77**		4	1	2
Cmax		6		**70**		2	0	2
CL		33		**127**		10	2	4
Tmax		3		**52**		0	0	0

The number of PK data in abstracts turned out to be around 4-25 times less than that from tables, which illustrates the importance of PK data extraction from tables. Beside abstracts and tables, the text of an article also contains PK data. There are only 24 pieces of additional PK parameter data from text of the 59 PDF articles. These PK parameters are from only 10 articles, four with tabular PK data and six without. **Error! Reference source not found.** listes the distribution details of these PK parameters among the 10 articles. By comparing the number of PK data from abstracts and tables with that from text, the information gain of extending the mining scope from abstract and tabular data to text seems very limited. Furthermore, after considering the amount of false positive data potentially recruited by full text based mining, the value of this mining scope escalation becomes truly arguable.

As summary, in articles with tabular PK data, when PK parameters are presented in full text, it is mostly in study result section which refers to PK data in tables. For articles without tabular PK data, we noticed the following phenomena: 1) PK parameters are well summarized and presented in abstracts, which means abstract based information extraction can collection most PK data for such articles; 2) some articles use one specific PK parameter to study the influence of different factors (e.g. genotype) to the subject drug. Such PK data can be illustrated in figures which cannot be extracted automatically yet; 3) some articles are studies for PK parameters out of the six PK categories chosen for information evaluation; 4) some articles are in image PDF format or secured, which keeps content of the files from being processed as free text; 5) Finally, articles that have novel PK parameters in text (other than abstracts and tables) are rare and the PK information content increase by including such articles is quite limited.

4.4.2 Pharmacokinetics Clearance Data extraction

Though abstracts and text of an article contain much less PK data than tables, the PK data extraction from all these three parts of an article is still evitable because PK data is very sparse for some drugs so any missing PK data is not affordable. For PK data extraction, we have worked on PubMed abstracts and PDF tables but not article text. However, article text is very similar to abstracts in aspects of entities and expressions of PK data in sentences. Also, we did not notice big difference when processing free text converted from PDF articles compared with PubMed abstracts handling during the information content evaluation step above. Thus the same IE method applied in PubMed abstract mining can be ported to article text.

To test the performance of PK extraction from articles, we still worked on the PK clearance data of drug MDZ. The clearance data was automatically extracted from the abstracts, PDF tables and PDF text of the known 104 PK relevant PubMed articles. The full text clearance extraction result was compared with the manually extracted data and mining data from abstract mining (Table 4.4). The number of clearance data from full text extraction is much higher than abstract extraction even on less number of articles, and the mean values are closer to the manually curated data.

Table 4.4 : Summary of PK Clearance Data Extraction

	Manual Abstract Extraction			Abstract Extraction			Full Text Extraction		
	# of PK	# of Abs	$\theta \pm se$	# of PK	# of Abs	$\theta \pm se$	# of PK	# of articles	$\theta \pm se$
Oral Clearance	25	170	83.6 ± 8.6	28	170	78.1 ± 6.0	**83**	104	**87.0 ± 4.3**
Systemic Clearance	50		32.3 ± 1.8	59		27.8 ± 1.0	**80**		**33.0 ± 3.4**

4.5 Full Text Mining Conclusion

In this part, we presented our solutions to four main challenges in full text mining for PK parameter data, i.e. retrieve abstracts with high relevancy, full article downloading, tabular data extraction and PK data extraction from full text. We were able to achieve promising performance in each test which indicates the potential of our method in PK parameter mining from literature.

Chapter 5.

CONCLUSIONS AND FUTURE WORK

5.1 PK Data Repository

One of the main purposes of this dissertation is to provide parameter data support for drug PK modeling. However, the mined numerical PK parameter data is mostly from clinical trials, which cannot be applied directly in drug modeling. Thus we proposed a meta-analysis approach [169] to transform PK parameters from TM to be modeling compatible, which makes TM really meaningful for drug studies. It was performed with a multivariate nonlinear mixed model.

By combining all the steps above, we developed a literature mining framework for PK parameter data extraction as shown in Figure 5.1. It is a pipeline made up of four automatic components: (1) information retrieval, which applies both ontology based name entity recognition (NER) and machine learning methods to classify PubMed search results; (2) article retrieval, which downloads full PDF articles through PubMed external links; (3) information extraction, which extracts PK data from both tables and free text of articles; and (4) PK data repository, which provides storage and query of the mined and transformed PK data.

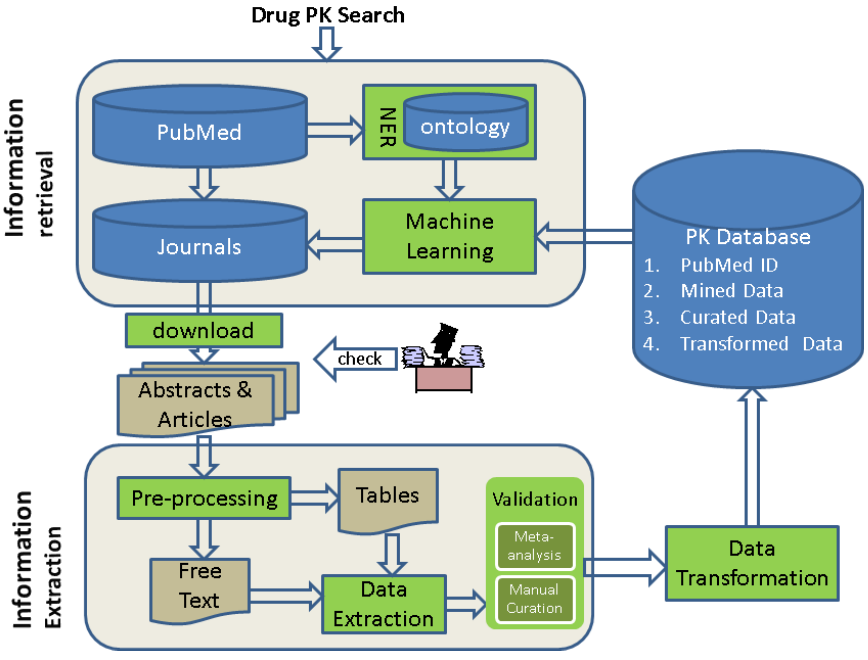


Figure 5.1 : PK Text Mining Framework

5.2 Contributions

This thesis focused on applied computer science methods, mainly in the field of text mining and machine learning, for drug pharmacokinetics data collection. We compared the performance of a machine learning, ~~based~~ fully automated TM pipeline with a manually-created template ~~based TM pipeline~~one. The template~~-~~-based TM shows a superior performance on most aspects. However, this method highly depends on human expertise to manually construct the template which fits specifically for a certain mining task---in this case, developed for the Midazolam~~n~~ drug. Meanwhile, the ~~SVM based~~ML mining strategy also shows a competitive performance, especially on precision, given the assumption that a positive and negative training data set is

Comment [LMR2]: Again, add something about how a fully automated pipeline compares with the pipeline you prefer which includes a manually-created template. What do you think the work shows about this comparison? What is the "value" of the expert data vs. a fully automated method?

available. The machine learning method can thus be used to extract a set of target data really fast which is especially valuable when a quick reference of the data range is needed. We explored the feasibility of TM on the extraction of drug PK parameter data from scientific abstracts and then articles. In abstract TM, our entity template based IR (F=0.78) and IE (F=0.90) method shows a superior overall performance when compared with the SVM based IR (F=0.68) and IE (F=0.79) method. In spite of overall performance, one method can be preferred over another depending on specific needs. In full-text TM, we compared the difference of PK data distribution in abstracts and articles. Then we proposed an effective solution for PK data extraction from tables of PDF articles.

The contributions of our work also include a pipeline of numerical data extraction from both abstracts and full-text articles:

- PK ontology for entity template construction
- Comparison/Combination of NLP and machine learning algorithms for PK information retrieval
- Full-text literature collection from open resources and publications
- Tabular data extraction
- Full text based PK information extraction

We have demonstrated the efficiency of our method to collect PK data from the literature using test drug midazolam. Through a comparison with information from a manually constructed PK database, DiDB, our method has achieved higher precision and richer information content.

~~Though~~ One limitation of this method is the comprehensive TM performance evaluation using more drugs, partly because of the lack of more gold standard data. Thus most tests in this dissertation were based on single test drug midazolam. Though~~,~~ we ~~should~~ expect similar performance when applying our TM method on other drugs because most drug studies follow a similar way to present PK data, which makes the template and rules developed for drug midazolam potentially portable for other drugs~~,~~. W~~w~~e still need more annotated literature data ~~accumulation with~~for more drugs to test this hypothesis, which is very time consuming considering the size of PubMed publications for individual drugs~~.~~ ~~Also,~~ Another hypothesis we need to test is that the template ~~was~~ constructed on top of our PK ontology and corpus~~, which~~ can summarize~~s~~ most key elements in drug studies. In abstract TM, addition test drugs were actually used in PK clearance data mining, which showed various information increases comparing with DiDB.

Comment [LMR3]: If it is so easy, why don't you show more results with other drugs? This statement is testable, so you need to say why you did not test it on the dissertation.

~~One~~ Another limit of our method is the coverage of additional PK parameters. We have focused on PK clearance data in abstract TM, as well as testing of four additional PK parameters (AUC, t1/2, Cmax, Tmax) in full-text TM. However, to achieve optimal data extraction performance for a specific PK parameter, the extraction rules may need to be adjusted accordingly. This requires some manual work. Also, for some drugs that are not very well studied, the published data can be very limited. Thus the missed articles from IR step and missed PK data from IE step, even though in a very small number, are unaffordable. For such drugs, a direct extraction (*Table* **3.4**) by skipping entity template and SVM should be more effective.

5.3 Future Work

In the full-text TM study, we have presented the results for test drug midazolam. To better demonstrate the power of our TM method, we will include more drugs to test its performance of IR, article retrieval, table processing, and IE.

To our knowledge, this is the first study to develop a literature mining based framework to extract numerical data from full text articles, which processes both plain text and tabular data. With the assistance of our mining pipeline, we should be able to construct a drug PK database effectively. We have built a website for our mined PK data and present it together with our existing drug information platform http://rweb.biostat.iupui.edu/DrugInteraction/. The PK data on this website is still quite sparse, we will apply our TM method to extract PK data for all the drugbank [9] drugs and promote it as a valuable tool for drug modeling studies. -As shown in the TM pipeline, we have tested all major parts and challenges in the pipeline. However, for additional drugs and PK parameters, we may need to improve the template and extraction rules accordingly. Also, ~~we are recruiting more~~with help from additional curators, we plan to validate results from each step of the TM pipeline, including IR, article retrieval, and data extraction. The validated data can be referred to design a new fully-automated machine learning method and improve our current template based method. On top of the PK parameter data extraction, we also need to collect key elements of a drug study, e.g. clinical trial design, using TM. Such information is very important to better justify the mined PK data.

In the full text documentation based TM section, we adjusted our IR standards to eliminate DDI type of abstracts because the PK parameter data extracted from DDI studies are usually biased by interactions. However, the PK parameter data from DDI studies are also very useful for drug modeling, for example, in drug DDI prediction and validation. We have developed and applied a set of DDI rules in full text IR step. Such rules can be referred by DDI TM but still have rooms to improve. On the other hand, most DDI studies contain unbiased PK parameters in control sets. Such PK parameters are qualified data for our PK collection and should also be included in our PK database. Furthermore, the data classification and extraction of DDI studies also depend on the training data accumulation and TM method improvements similarly as what we did in the PK data TM research. Thus the TM study presented in this dissertation sets a good base for our future DDI TM research.

Appendix I

```
$PROB transformation from non-comp to two-comp
$INPUT ID CF Dsn DV TIME DOSE MDV
; the code is also available at
; http://rweb.biostat.iupui.edu/nc2tc.ctl
; data file is available at
; http://rweb.biostat.iupui.edu/metanonmem.txt
; TIME data item is used to index non-compartment parameters
$ABBREV DERIV2=NO
$DATA metanonmem.txt IGNORE=C
$PRED
; two compartment parameters
TVV1=EXP(THETA(1))
V1=TVV1*EXP(ETA(1))
TVka=EXP(THETA(2))
ka=TVka*EXP(ETA(2))
TVke=EXP(THETA(3))
ke=TVke*EXP(ETA(3))
TVk12=EXP(THETA(4))
k12=TVk12
TVk21=EXP(THETA(5))
k21=TVk21

; Please refer to equation 5-8 in the paper for the transformation
; lambda1 and lambda2
ins = (ke+k21+k12)*(ke+k21+k12) - 4*ke*k21
lam1 = .5*(ke+k21+k12+SQRT(ins))
lam2 = .5*(ke+k21+k12-SQRT(ins))

; Cmax for PO
ka1 = ka-lam1
ka2 = ka-lam2
lam12 = lam1-lam2
Tmax = LOG(ka/lam1)/(ka-lam1)
coef = DOSE*CF*ka/V1
term1 = (k21-lam1)*EXP(-lam1*Tmax)/(ka1*(-lam12))
term2 = (k21-lam2)*EXP(-lam2*Tmax)/(ka2*lam12)
term3 = (ka-k21)*EXP(-ka*Tmax)/(ka1*ka2)
PRED0 = coef*(term1+term2-term3)

; Cmax for IV
PRED1 = DOSE/V1

; AUC
PRED2 = DOSE*CF/(V1*ke)

; Tmax
PRED3 = LOG(ka/lam1)/(ka-lam1)
```

```
; Tf (T_1/2, fast)
PRED4 = LOG(2)/lam1

; Ts (T_1/2, slow)
PRED5 = LOG(2)/lam2
; CL
PRED6 = V1*ke
; Vd
PRED7 = ke*V1/lam2

; T1=0, cmax for PO; 5, cmax for IV; 1, auc; 2, tmax; 3, tf; 4, ts; 6, cl; 7,
vd
T1=TIME
IF (T1.EQ.0) IPRE = PRED0; Cmax for PO
IF (T1.EQ.5) IPRE = PRED1; Cmax for IV
IF (T1.EQ.1) IPRE = PRED2; AUC
IF (T1.EQ.2) IPRE = PRED3; Tmax
IF (T1.EQ.3) IPRE = PRED4; Tf
IF (T1.EQ.4) IPRE = PRED5; Ts
IF (T1.EQ.6) IPRE = PRED6; CL
IF (T1.EQ.7) IPRE = PRED7; Vd

Y=IPRE*EXP(EPS(1))
IF (ins.LT.0)  Y = 0
IF (IPRE.LE.0) Y = 0

$THETA
(0,3.38,10) ;V1
(-10,0.494,5) ;ka
(-10,-0.323,5) ;ke
(-10,-1.3,5) ;k12
(-10,-1.4,5) ;k21
$OMEGA
.0367
.374
.025
$SIGMA
0.0486

; FOCE
$EST METHOD=1 NOABORT SIG=5 MAX=9999 PRINT=10 POSTHOC
```

Reference

1. Oberholzer-Gee, F. and S.N. Inamdar, *Merck's recall of rofecoxib--a strategic perspective.* N Engl J Med, 2004. **351**(21): p. 2147-9.
2. *http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.*
3. Chang, M., et al., *Innovative approaches in drug development.* J Biopharm Stat, 2007. **17**(5): p. 775-89.
4. Chien, J.Y., et al., *Pharmacokinetics/Pharmacodynamics and the stages of drug development: role of modeling and simulation.* AAPS J, 2005. **7**(3): p. E544-59.
5. Lalonde, R.L., et al., *Model-based drug development.* Clin Pharmacol Ther, 2007. **82**(1): p. 21-32.
6. O'Neill, R.T., *FDA's critical path initiative: a perspective on contributions of biostatistics.* Biom J, 2006. **48**(4): p. 559-64.
7. Yao, L., J.A. Evans, and A. Rzhetsky, *Novel opportunities for computational biology and sociology in drug discovery.* Trends Biotechnol, 2010. **28**(4): p. 161-70.
8. DiDB, *http://www.druginteractioninfo.org/.*
9. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets.* Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.
10. Thorn, C.F., T.E. Klein, and R.B. Altman, *Pharmacogenomics and bioinformatics: PharmGKB.* Pharmacogenomics, 2010. **11**(4): p. 501-5.
11. DailyMed, *http://dailymed.nlm.nih.gov/dailymed/about.cfm.*
12. PubPK, *http://www.pubpk.org.*
13. Moda, T.L., et al., *PK/DB: database for pharmacokinetic properties and predictive in silico ADME models.* Bioinformatics, 2008. **24**(19): p. 2270-1.
14. Hunter, L. and K.B. Cohen, *Biomedical language processing: what's beyond PubMed?* Mol Cell, 2006. **21**(5): p. 589-94.
15. Mitchell P. Marcus , B.S., Mary Ann Marcinkiewicz *Building a Large Annotated Corpus of English: The Penn Treebank.* 1993. **19**(2): p. 313-330.
16. Baumgartner, W.A., Jr., et al., *Manual curation is not sufficient for annotation of genomic databases.* Bioinformatics, 2007. **23**(13): p. i41-8.
17. Aronson, A.R., et al., *The NLM Indexing Initiative's Medical Text Indexer.* Stud Health Technol Inform, 2004. **107**(Pt 1): p. 268-72.
18. Leitner, F. and A. Valencia, *A text-mining perspective on the requirements for electronically annotated abstracts.* FEBS Lett, 2008. **582**(8): p. 1178-81.
19. Altman, R.B., et al., *Text mining for biology--the way forward: opinions from leading scientists.* Genome Biol, 2008. **9 Suppl 2**: p. S7.
20. Siadaty, M.S., J. Shu, and W.A. Knaus, *Relemed: sentence-level search engine with relevance score for the MEDLINE database of biomedical articles.* BMC Med Inform Decis Mak, 2007. **7**: p. 1.
21. *http://www.pubmedreader.com.*
22. Eaton, A.D., *HubMed: a web-based biomedical literature search interface.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W745-7.

23. Lewis, J., et al., *Text similarity: an alternative way to search MEDLINE.* Bioinformatics, 2006. **22**(18): p. 2298-304.
24. Poulter, G.L., et al., *MScanner: a classifier for retrieving Medline citations.* BMC Bioinformatics, 2008. **9**: p. 108.
25. Andrade, M.A. and A. Valencia, *Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families.* Bioinformatics, 1998. **14**(7): p. 600-7.
26. Dumais, S.T., *Enhancing performance in latent semantic (LSI) indexing.* Behavior ResearchMethods, Instruments and Computers, 1990. **23**(2): p. 229–236.
27. Hagit Shatkay , W.J.W., *Finding Themes in Medline Documents - Probabilistic Similarity Search.* Proc. IEEE Conf. on Advances in Digital Libraries, 2000: p. 183-192.
28. Hofmann, T., *Probabilistic Latent Semantic indexing.* Proc. 22nd ACM Int. Conf. on Research and Development in Information Retrieval, 1999.
29. Vapnik, V., *The Nature of Statistical Learning Theory.* Springer-Verlag, NY, 1995.
30. David R. H. Miller, T.L., Richard M. Schwartz, *A Hidden Markov Model. Information Retrieval System.* Proceedings of the 22nd annual international ACM SIGIR conference, 1999.
31. H.C. Wu, R.W.P.L., K.F. Wong, K.L. Kwok *Interpreting TF-IDF term weights as making relevance decisions.* ACM Transactions on Information Systems, 2008. **26**(3).
32. Wall, M.E., Andreas Rechtsteiner, Luis M. Rocha, *Singular value decomposition and principal component analysis.* A Practical Approach to Microarray Data Analysis. D.P. Berrar, W. Dubitzky, M. Granzow, eds. Kluwer: Norwell, MA, 2003: p. 91-109.
33. A. Lourenço, M.C., A. Wong, A. Nematzadeh, F. Pan, H. Shatkay, and L.M. Rocha, *A Linear Classifier Based on Entity Recognition Tools and a Statistical Approach to Method Extraction in the Protein-Protein Interaction Literature.* BMC Bioinformatics, 2011. **In Press**.
34. Van Landeghem, S., et al., *Discriminative and informative features for biomolecular text mining with ensemble feature selection.* Bioinformatics, 2010. **26**(18): p. i554-60.
35. Chen, L., H. Liu, and C. Friedman, *Gene name ambiguity of eukaryotic nomenclatures.* Bioinformatics, 2005. **21**(2): p. 248-56.
36. Mons, B., *Which gene did you mean?* BMC Bioinformatics, 2005. **6**: p. 142.
37. Jiang J, Z.C., *An empirical study of tokenization strategies for biomedical information retrieval.* Inform Retr, 2007. **10**: p. 341-363.
38. Tomanek, K., J. Wermter, and U. Hahn, *A reappraisal of sentence and token splitting for life sciences documents.* Stud Health Technol Inform, 2007. **129**(Pt 1): p. 524-8.
39. Gaudan, S., H. Kirsch, and D. Rebholz-Schuhmann, *Resolving abbreviations to their senses in Medline.* Bioinformatics, 2005. **21**(18): p. 3658-64.
40. Zhou, W., V.I. Torvik, and N.R. Smalheiser, *ADAM: another database of abbreviations in MEDLINE.* Bioinformatics, 2006. **22**(22): p. 2813-8.
41. Chang, J.T., H. Schutze, and R.B. Altman, *Creating an online dictionary of abbreviations from MEDLINE.* J Am Med Inform Assoc, 2002. **9**(6): p. 612-20.
42. Okazaki, N., S. Ananiadou, and J. Tsujii, *Building a high-quality sense inventory for improved abbreviation disambiguation.* Bioinformatics, 2010. **26**(9): p. 1246-53.
43. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2010.
44. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2010.
45. Segura-Bedmar, I., et al., *Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents.* BMC Bioinformatics, 2010. **11 Suppl 2**: p. S1.
46. Tamames, J. and A. Valencia, *The success (or not) of HUGO nomenclature.* Genome Biol, 2006. **7**(5): p. 402.

47. Gerner, M., G. Nenadic, and C.M. Bergman, *LINNAEUS: a species name identification system for biomedical literature.* BMC Bioinformatics, 2010. **11**: p. 85.
48. http://www.nlm.nih.gov/research/umls/.
49. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
50. Liu, H., et al., *BioThesaurus: a web-based thesaurus of protein and gene names.* Bioinformatics, 2006. **22**(1): p. 103-5.
51. Fellbaum, C., *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press, 1998.
52. Huang, K.C., et al., *Using WordNet synonym substitution to enhance UMLS source integration.* Artif Intell Med, 2009. **46**(2): p. 97-109.
53. Katerina Frantzi, S.A., Hideki Mima, *Automatic recognition of multiword terms.* International Journal on Digital Libraries, 2000. **3**(2): p. 115-130.
54. Blaschke, C. and A. Valencia, *Automatic ontology construction from the literature.* Genome Inform, 2002. **13**: p. 201-13.
55. Philipp, C., *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.* New York, USA:Springer, ScienceþBusiness Media, LLC, 2006.
56. Ryu P-M, C.K.-S., *Taxonomy learning using term specificity and similarity.* Proceedings of the 2ndWorkshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, 2006: p. 41-8.
57. Winnenburg, R., et al., *Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?* Brief Bioinform, 2008. **9**(6): p. 466-78.
58. Smith, L., et al., *Overview of BioCreative II gene mention recognition.* Genome Biol, 2008. **9 Suppl 2**: p. S2.
59. Jin, Y., et al., *Automated recognition of malignancy mentions in biomedical literature.* BMC Bioinformatics, 2006. **7**: p. 492.
60. Yeh, A., et al., *BioCreAtIvE task 1A: gene mention finding evaluation.* BMC Bioinformatics, 2005. **6 Suppl 1**: p. S2.
61. Huang, M., J. Liu, and X. Zhu, *GeneTUKit: a software for document-level gene normalization.* Bioinformatics, 2011.
62. Kuhn, M., et al., *STITCH: interaction networks of chemicals and proteins.* Nucleic Acids Res, 2008. **36**(Database issue): p. D684-8.
63. Bjorne, J., et al., *Complex event extraction at PubMed scale.* Bioinformatics, 2010. **26**(12): p. i382-90.
64. Miwa, M., et al., *Event extraction with complex event classification using rich features.* J Bioinform Comput Biol, 2010. **8**(1): p. 131-46.
65. Jin D. Kim, T.O., Sampo Pyysalo, Yoshinobu Kano, Jun'ichi Tsujii, *Overview of bionlp'09 shared task on event extraction.* Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task, Association for Computational Linguistics, Boulder, Colorado, 2009: p. 1-9.
66. M, P., *An algorithm for suffix stripping.* Program, 1980. **14**: p. 130-137.
67. Smalheiser, N.R., W. Zhou, and V.I. Torvik, *Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results.* J Biomed Discov Collab, 2008. **3**: p. 2.
68. Harmston, N., W. Filsell, and M.P. Stumpf, *What the papers say: Text mining for genomics and systems biology.* Hum Genomics, 2010. **5**(1): p. 17-29.
69. Smith, L., T. Rindflesch, and W.J. Wilbur, *MedPost: a part-of-speech tagger for bioMedical text.* Bioinformatics, 2004. **20**(14): p. 2320-1.
70. Divita, G., A.C. Browne, and R. Loane, *dTagger: a POS tagger.* AMIA Annu Symp Proc, 2006: p. 200-3.

71. Baumgartner, W.A., Jr., et al., *Concept recognition for extracting protein interaction relations from biomedical text.* Genome Biol, 2008. **9 Suppl 2**: p. S9.

72. Settles, B., *ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text.* Bioinformatics, 2005. **21**(14): p. 3191-2.

73. B, C., *Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval.* Proceedings of the 13th Annual Text Retrieval Conference, 2004.

74. Razvan Bunescu, R.M., Razvan Bunescu, Raymond Mooney, *Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from MEDLINE.* Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL, 2006: p. 49–56.

75. Gaizauskas, R., et al., *Protein structures and information extraction from biological texts: the PASTA system.* Bioinformatics, 2003. **19**(1): p. 135-43.

76. Lau, W.W., C.A. Johnson, and K.G. Becker, *Rule-based human gene normalization in biomedical text with confidence estimation.* Comput Syst Bioinformatics Conf, 2007. **6**: p. 371-9.

77. Blaschke, C. and A. Valencia, *The potential use of SUISEKI as a protein interaction discovery tool.* Genome Inform, 2001. **12**: p. 123-34.

78. Niu, Y., D. Otasek, and I. Jurisica, *Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D.* Bioinformatics, 2010. **26**(1): p. 111-9.

79. Fayruzov, T., et al., *Linguistic feature analysis for protein interaction extraction.* BMC Bioinformatics, 2009. **10**: p. 374.

80. Fundel, K., R. Kuffner, and R. Zimmer, *RelEx--relation extraction using dependency parse trees.* Bioinformatics, 2007. **23**(3): p. 365-71.

81. Bethard, S., et al., *Semantic role labeling for protein transport predicates.* BMC Bioinformatics, 2008. **9**: p. 277.

82. Sanchez-Graillet, O. and M. Poesio, *Negation of protein-protein interactions: analysis and extraction.* Bioinformatics, 2007. **23**(13): p. i424-32.

83. Krallinger, M., R. Malik, and A. Valencia, *Text mining and protein annotations: the construction and use of protein description sentences.* Genome Inform, 2006. **17**(2): p. 121-30.

84. Hastie T, T.R., Friedman J. , *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer; New York: 2009. **Chapter 16: Random Forests.**

85. Hosmer, D.W.L., Stanley *Applied Logistic Regression (2nd ed.).* Wiley, 2000.

86. Galitsky, B.A., S.O. Kuznetsov, and D.V. Vinogradov, *Applying hybrid reasoning to mine for associative features in biological data.* J Biomed Inform, 2007. **40**(3): p. 203-20.

87. Swanson, D.R., *Fish oil, Raynaud's syndrome, and undiscovered public knowledge.* Perspect Biol Med, 1986. **30**(1): p. 7-18.

88. DiGiacomo, R.A., J.M. Kremer, and D.M. Shah, *Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study.* Am J Med, 1989. **86**(2): p. 158-64.

89. Srinivasan, P. and B. Libbus, *Mining MEDLINE for implicit links between dietary substances and diseases.* Bioinformatics, 2004. **20 Suppl 1**: p. i290-6.

90. Hristovski, D., et al., *Using literature-based discovery to identify disease candidate genes.* Int J Med Inform, 2005. **74**(2-4): p. 289-98.

91. van Haagen, H.H., et al., *Novel protein-protein interactions inferred from literature context.* PLoS One, 2009. **4**(11): p. e7894.

92. Jensen, L.J., J. Saric, and P. Bork, *Literature mining for the biologist: from information retrieval to biological discovery.* Nat Rev Genet, 2006. **7**(2): p. 119-29.

93. Barbosa-Silva, A., et al., *LAITOR--Literature Assistant for Identification of Terms co-Occurrences and Relationships.* BMC Bioinformatics, 2010. **11**: p. 70.
94. Andreopoulos, B., D. Alexopoulou, and M. Schroeder, *Word Sense Disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering.* Int J Data Min Bioinform, 2008. **2**(3): p. 193-215.
95. Jenssen, T.K., et al., *A literature network of human genes for high-throughput analysis of gene expression.* Nat Genet, 2001. **28**(1): p. 21-8.
96. Alako, B.T., et al., *CoPub Mapper: mining MEDLINE based on search term co-publication.* BMC Bioinformatics, 2005. **6**: p. 51.
97. Kim, J.J., P. Pezik, and D. Rebholz-Schuhmann, *MedEvi: retrieving textual evidence of relations between biomedical concepts from Medline.* Bioinformatics, 2008. **24**(11): p. 1410-2.
98. Leitner, F., et al., *An Overview of BioCreative II.5.* IEEE/ACM Trans Comput Biol Bioinform, 2010. **7**(3): p. 385-99.
99. Arighi, C.N., et al., *Overview of the BioCreative III Workshop.* BMC Bioinformatics, 2011. **12 Suppl 8**: p. S1.
100. Bui, Q.C., S. Katrenko, and P.M. Sloot, *A hybrid approach to extract protein-protein interactions.* Bioinformatics, 2010.
101. Li, Y., et al., *Learning an enriched representation from unlabeled data for protein-protein interaction extraction.* BMC Bioinformatics, 2010. **11 Suppl 2**: p. S7.
102. Brady, S. and H. Shatkay, *EpiLoc: a (working) text-based system for predicting protein subcellular location.* Pac Symp Biocomput, 2008: p. 604-15.
103. Caporaso, J.G., et al., *MutationFinder: a high-performance system for extracting point mutation mentions from text.* Bioinformatics, 2007. **23**(14): p. 1862-5.
104. Xuan, W., et al., *Medline search engine for finding genetic markers with biological significance.* Bioinformatics, 2007. **23**(18): p. 2477-84.
105. Tamames, J. and V. de Lorenzo, *EnvMine: a text-mining system for the automatic extraction of contextual information.* BMC Bioinformatics, 2010. **11**: p. 294.
106. Naeem, H., et al., *miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature.* BMC Bioinformatics, 2010. **11**: p. 135.
107. Li, X., et al., *Global mapping of gene/protein interactions in PubMed abstracts: a framework and an experiment with P53 interactions.* J Biomed Inform, 2007. **40**(5): p. 453-64.
108. Xu, H., et al., *MedEx: a medication information extraction system for clinical narratives.* J Am Med Inform Assoc, 2010. **17**(1): p. 19-24.
109. Ongenaert, M., et al., *PubMeth: a cancer methylation database combining text-mining and expert annotation.* Nucleic Acids Res, 2008. **36**(Database issue): p. D842-6.
110. Fang, Y.C., H.C. Huang, and H.F. Juan, *MeInfoText: associated gene methylation and cancer information from text mining.* BMC Bioinformatics, 2008. **9**: p. 22.
111. Li, J., X. Zhu, and J.Y. Chen, *Discovering breast cancer drug candidates from biomedical literature.* Int J Data Min Bioinform, 2010. **4**(3): p. 241-55.
112. Xu, S. and M. Krauthammer, *A new pivoting and iterative text detection algorithm for biomedical images.* J Biomed Inform, 2010. **43**(6): p. 924-31.
113. Sneiderman, C.A., et al., *Knowledge-based methods to help clinicians find answers in MEDLINE.* J Am Med Inform Assoc, 2007. **14**(6): p. 772-80.
114. Dina Demner-Fushman, J.L., *Answering clinical questions with knowledge-based and statistical techniques.* Comput Linguist, 2007. **33**: p. 63-103.
115. Xu, H., et al., *A natural language processing (NLP) tool to assist in the curation of the laboratory Mouse Tumor Biology Database.* AMIA Annu Symp Proc, 2006: p. 1150.

116. Narayanaswamy, M., K.E. Ravikumar, and K. Vijay-Shanker, *Beyond the clause: extraction of phosphorylation information from medline abstracts.* Bioinformatics, 2005. **21 Suppl 1**: p. i319-27.

117. Karamanis, N., et al., *Integrating natural language processing with FlyBase curation.* Pac Symp Biocomput, 2007: p. 245-56.

118. Muller, H.M., E.E. Kenny, and P.W. Sternberg, *Textpresso: an ontology-based information retrieval and extraction system for biological literature.* PLoS Biol, 2004. **2**(11): p. e309.

119. Couto, F.M., et al., *GOAnnotator: linking protein GO annotations to evidence text.* J Biomed Discov Collab, 2006. **1**: p. 19.

120. Donaldson, I., et al., *PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine.* BMC Bioinformatics, 2003. **4**: p. 11.

121. Burkhardt, K., B. Schneider, and J. Ory, *A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank.* PLoS Comput Biol, 2006. **2**(10): p. e99.

122. BioNLP, *http://zope.bioinfo.cnio.es/bionlp_tools/*.

123. Tscherne, H. and B. Wippermann, *[Functional therapy in treatment of fractures and joint injuries].* Zentralbl Chir, 1990. **115**(16): p. 997-1005.

124. Dina Demner-Fushman, S.M.H., Nicholas C. Ide, Russell F. Loane, Patrick Ruch, Miguel E. Ruiz, Lawrence H. Smith, Lorraine K. Tanabe, W. John Wilbur, Alan R. Aronson, *Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort.* TREC 2006.

125. Jaeger, S., et al., *Integrating protein-protein interactions and text mining for protein function prediction.* BMC Bioinformatics, 2008. **9 Suppl 8**: p. S2.

126. Shatkay, H., et al., *SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data.* Bioinformatics, 2007. **23**(11): p. 1410-7.

127. von Mering, C., et al., *STRING 7--recent developments in the integration and prediction of protein interactions.* Nucleic Acids Res, 2007. **35**(Database issue): p. D358-62.

128. Cohen, A.M. and W.R. Hersh, *A survey of current work in biomedical text mining.* Brief Bioinform, 2005. **6**(1): p. 57-71.

129. Blaschke, C. and A. Valencia, *Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study.* Comp Funct Genomics, 2001. **2**(4): p. 196-206.

130. Cohen, K.B., et al., *The structural and content aspects of abstracts versus bodies of full text journal articles are different.* BMC Bioinformatics, 2010. **11**: p. 492.

131. Divoli, A., M.A. Wooldridge, and M.A. Hearst, *Full text and figure display improves bioscience literature search.* PLoS One, 2010. **5**(4): p. e9619.

132. McEntyre, J.R., et al., *UKPMC: a full text article resource for the life sciences.* Nucleic Acids Res, 2010.

133. Shah, P.K., et al., *Information extraction from full text scientific articles: where are the keywords?* BMC Bioinformatics, 2003. **4**: p. 20.

134. Schuemie, M.J., et al., *Distribution of information in biomedical abstracts and full-text publications.* Bioinformatics, 2004. **20**(16): p. 2597-604.

135. OTMI, *http://opentextmining.org/wiki/Main_Page*.

136. Corney, D.P., et al., *BioRAT: extracting biological information from full-length papers.* Bioinformatics, 2004. **20**(17): p. 3206-13.

137. Lourenco, A., et al., *@Note: a workbench for biomedical text mining.* J Biomed Inform, 2009. **42**(4): p. 710-20.

138. Garten, Y. and R.B. Altman, *Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text.* BMC Bioinformatics, 2009. **10 Suppl 2**: p. S6.
139. Spasic, I., et al., *KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways.* Bioinformatics, 2009. **25**(11): p. 1404-11.
140. BioCreative'12, *http://www.biocreative.org/events/bc-workshop-2012/workshop/*.
141. Weeber, M., et al., *Text-based discovery in biomedicine: the architecture of the DAD-system.* Proc AMIA Symp, 2000: p. 903-7.
142. Roberts, P.M. and W.S. Hayes, *Information needs and the role of text mining in drug development.* Pac Symp Biocomput, 2008: p. 592-603.
143. Garten, Y., A. Coulet, and R.B. Altman, *Recent progress in automatically extracting information from the pharmacogenomic literature.* Pharmacogenomics, 2010. **11**(10): p. 1467-89.
144. Cohen, K.B., et al., *MINING THE PHARMACOGENOMICS LITERATURE - Workshop Introduction.* Pac Symp Biocomput, 2011: p. 362-3.
145. Percha, B., Y. Garten, and R.B. Altman, *Discovery and explanation of drug-drug interactions via text mining.* Pac Symp Biocomput, 2012: p. 410-21.
146. Tari, L., et al., *Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism.* Bioinformatics, 2010. **26**(18): p. i547-53.
147. Segura-Bedmar, I., P. Martinez, and C. de Pablo-Sanchez, *A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents.* BMC Bioinformatics, 2011. **12 Suppl 2**: p. S1.
148. Aronson, A.R., *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.* Proc AMIA Symp, 2001: p. 17-21.
149. Segura-Bedmar, I., P. Martinez, and C. de Pablo-Sanchez, *Using a shallow linguistic kernel for drug-drug interaction extraction.* J Biomed Inform, 2011. **44**(5): p. 789-804.
150. Giuliano C, L.A., Romano L., *Exploiting shallow linguistic information for relation extraction from biomedical literature.* In: Proceedings of the eleventh conference of the European chapter of the association for computational linguistics (EACL-2006), 2006: p. 5-7.
151. Boyce, R., et al., *Computing with evidence Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment.* J Biomed Inform, 2009. **42**(6): p. 979-89.
152. Boyce, R., et al., *Computing with evidence Part II: An evidential approach to predicting metabolic drug-drug interactions.* J Biomed Inform, 2009. **42**(6): p. 990-1003.
153. Hakenberg, J., et al., *Finding kinetic parameters using text mining.* OMICS, 2004. **8**(2): p. 131-52.
154. NOONBURG, D.B., *PDFTOTEXT [On-line]. Available: http://www.aimnet.com/*derekn/xpdf/.* 1996.
155. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment.* Nucleic Acids Res, 2008. **36**(Database issue): p. D480-4.
156. Blake, J.A. and M.A. Harris, *The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis.* Curr Protoc Bioinformatics, 2008. **Chapter 7**: p. Unit 7 2.
157. Le Novere, N., *Model storage, exchange and integration. .* BMC Neurosci, 2006. **7**(S11).
158. Jyh-Jong Tsay, B.-L.W., Chang-Ching Hsieh, *Automatic Extraction of Kinetic Information from Biochemical Literatures.* FSKD, 2009. **5**: p. 28-32.
159. Heinen, S., B. Thielen, and D. Schomburg, *KID--an algorithm for fast and efficient text mining used to automatically generate a database containing kinetic information of enzymes.* BMC Bioinformatics, 2010. **11**: p. 375.
160. Cheung, K.H., et al., *Structured digital tables on the Semantic Web: toward a structured digital literature.* Mol Syst Biol, 2010. **6**: p. 403.
161. Protégé, *http://protege.stanford.edu*.

162. PKO, *http://bioportal.bioontology.org/visualize/40917.*

163. Wang, Z., et al., *Literature mining on pharmacokinetics numerical data: a feasibility study.* J Biomed Inform, 2009. **42**(4): p. 726-35.

164. Cristian Duda, G.F., Donald Kossmann, Chong Zhou, *AJAXSearch: crawling, indexing and searching web 2.0 applications.* PVLDB, 2008. **1**(2): p. 1440-1443.

165. Padmini Srinivasan , J.M., Olivier Bodenreider , Gautam Pant , Filippo Menczer *Web Crawling Agents for Retrieving Biomedical Information.* Proc. of the International Workshop on Bioinformatics and Multi-Agent Systems, 2002.

166. Hearst, M.A., et al., *BioText Search Engine: beyond abstract search.* Bioinformatics, 2007. **23**(16): p. 2196-7.

167. Claire Nedellec, M.O.A.V., Philippe Bessières, *Sentence filtering for information extraction in genomics, a classification problem.* PKDD 2001: p. 326-337.

168. Fernandez, J.M., R. Hoffmann, and A. Valencia, *iHOP web services.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W21-6.

169. Wang, Z., et al., *Non-compartment model to compartment model pharmacokinetics transformation meta-analysis--a multivariate nonlinear mixed model.* BMC Syst Biol, 2010. **4 Suppl 1**: p. S8.

170. Segel, H.I., *Enzyme Kinetics – Behavior and analysis of rapid equilibrium and steady state enzyme systems.* John Wiley & Sons, Inc. New York., 1975.

171. Consortium, T.I.T., *Membrane transporters in drug development.* Nature Review Drug Discovery, 2010. **9**: p. 215-236.

172. Rostami-Hodjegan, A., and Tucker, G., *"In Silico" simulations to assess the "in vivo" consequences of "in vitro" metabolic drug-drug interactions.* Drug Discovery Today: Technologies, 2004. **1**: p. 441-448.

173. Rowland, M., and Tozer, T. N., *Clinical Pharmacokinetics Concept and Applications*. Third ed1995, London: Lippincott Williams & Wilkins.

174. Gibaldi, M., and Perrier, D., *Pharmacokinetics, 2nd edition.* Dekker, 1982.

175. Huang, S.M., Temple, R., Throckmorton, D.C., and Lesko, L. J. , *Drug interaction studies: study design, data analysis, and implications for dosing and labeling.* Clinical Pharmacology and Therapeutics, 2007. **81**(2): p. 298-304.

176. Guengerich, F.P., *Cytochrome p450 and chemical toxicology.* Chem Res Toxicol, 2008. **21**(1): p. 70-83.

177. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., et al., *DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.* Nucleic Acids Res., 2011. **39**(D1035-1041).

178. Rubin, D.L., N.F. Noy, and M.A. Musen, *Protege: a tool for managing and using terminology in radiology applications.* J Digit Imaging, 2007. **20 Suppl 1**: p. 34-46.

179. Borges, S., et al., *Composite functional genetic and comedication CYP2D6 activity score in predicting tamoxifen drug exposure among breast cancer patients.* J Clin Pharmacol. **50**(4): p. 450-8.

180. Chien, J.Y., et al., *Stochastic prediction of CYP3A-mediated inhibition of midazolam clearance by ketoconazole*. Drug Metab Dispos, 2006. **34**(7): p. 1208-19.

181. Williams, J.A., et al., *Comparative metabolic capabilities of CYP3A4, CYP3A5, and CYP3A7.* Drug Metab Dispos, 2002. **30**(8): p. 883-91.

182. Brunton, L.L., Chabner, B.A., Knollmann, B.C., *Goodman & Gilman's The Pharmacological Basis of Therapeutics.* **12**.

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

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

185. Airola, A., Pyysalo, S., Bjorne, J., Pahikkala, T., Ginter, F., Salakoski, T, *All- paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning.* BMC Bioinformatics, 2008. **9**(suppl 11): p. S2.

186. De Marneffe, M., MacCartney, B., Manning, C., *Generating typed dependency parses from phrase structure parses.* Proceedings of LREC., 2006. **6**: p. 449-454.

187. Karnik, S., Subhadarshini, A., Wang, Z., Rocha, L.M., and Li, L., *Extraction of drug-drug interactions using all paths graph kernel.* Proc. of the 1st Challenge task on Drug Drug Interaction Extraction, 2011: p. 83-88.

188. opennlp, http://opennlp.sourceforge.net/index.html.

189. Jeffrey C. Reynar, A.R., *A Maximum Entropy Approach to Identifying Sentence Boundaries. .* In Proceedings of the Fifth Conference on Applied Natural Language Processing., 1997.

190. Porter., M.F., *An algorithm for suffix stripping.* Program., 1980. **14**(3): p. 130–137.

191. Lang., Y.M.K.S.W.Z.H.S.L., *A Bayesian meta-analysis on published sample mean and variance pharmacokinetic data with application to drug-drug interaction prediction.* Journal of biopharmaceutical statistics. **18**(6): p. 1063-83.

192. Joachims, T., *Making large-Scale SVM Learning Practical.* Advances in Kernel Methods - Support Vector Learning. B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press., 1999.

193. MetaMap, *http://metamap.nlm.nih.gov/.*

194. McCray, A.T., S. Srinivasan, and A.C. Browne, *Lexical methods for managing variation in biomedical terminologies.* Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care, 1994: p. 235-9.

195. Peters, L., J.E. Kapusnik-Uner, and O. Bodenreider, *Methods for managing variation in clinical drug names.* AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2010. **2010**: p. 637-41.

196. Fleishaker, J.C., et al., *Hormonal effects on tirilazad clearance in women: assessment of the role of CYP3A.* J Clin Pharmacol, 1999. **39**(3): p. 260-7.