Translational Biomedical Informatics and Pharmacometrics Approaches in the Drug Interactions Research

Pengyue Zhang[1, #], Heng-Yi Wu[1, #], Chien-Wei Chiang[1], Lei Wang[1, 2], Samar Binkheder[3, 4], Xueying Wang[2], Donglin Zeng[5], Sara K. Quinney[6], and Lang Li[1, *]

1 Department of Biomedical Informatics, College of Medicine, the Ohio State University, Columbus, OH, USA;

2 Intelligent Systems and Bioinformatics Institute, College of Automation, Harbin Engineering University, Harbin, Heilongjiang, China;

3 Department of Biohealth Informatics, Indiana University School of Informatics and Computing, Indianapolis, IN, USA;

4 Medical Informatics Unit, College of Medicine, King Saud University, Riyadh, Saudi Arabia;

5 Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA;

6 Department of Obstetrics and Gynecology, Indiana University, Indianapolis, IN, USA.

# These two authors contribute equally.

* To whom correspondence should be addressed.

As an Associate Editor for *CPT: Pharmacometrics & Systems Pharmacology*, Lang Li was not involved in the review or decision process for this paper.

Funding: No funding was received for this work.

COI: The authors declared no competing interests for this work.

Introduction

Drug interaction is a leading cause of adverse drug events and a major obstacle for current clinical practice. Pharmacovigilance data mining, pharmacokinetic modeling, and text mining are computation and informatics tools on integrating drug interaction knowledge and generating drug interaction hypothesis. We provide a comprehensive overview of these translational biomedical informatics methodologies with related databases. We hope this review illustrates the complementary nature of these informatics approaches and facilitates the translational drug interaction research.

1. Background

Adverse drug events (ADEs), the unintended drug side effects, have led to the major public health burden. In US alone, more than 500,000 serious ADEs were reported annually to the US Food and Drug Administration (FDA) during the past 5 years (1). ADEs are one of the leading causes of morbidity and mortality. A meta-analysis of 39 prospective studies suggested that 6.7% of inpatients have severe ADEs and 0.32% have fatal drug reactions (2). Substantial evidences show that the drug-drug interaction (DDI) is one of the leading causes of ADEs. With the increasing rate of poly-pharmacy, the risk of ADEs increased from 13% (two drugs) to 58% (five drugs) (3). Hence, efficient and powerful computational approaches are needed in detecting the DDI-induced ADE signals, and investigating their molecular mechanisms.

In order to evaluate clinical effects and molecular mechanisms of DDIs, clinical pharmacokinetic (PK) studies, pharmaco-epidemiologic studies, and in-vitro PK experiments have been routinely utilized. One salient example is that of breast cancer hormonal therapy, tamoxifen. The formation of its active metabolite, endoxifen, was inhibited by co-administrated selective serotonin reuptake inhibitor (SSRI) paroxetine in a clinical pharmacokinetics study (4). In vitro metabolism studies revealed that this is due to paroxetine’s strong inhibition of the tamoxifen bio-transformation to endoxifen via the CYP2D6 pathway (5). In a follow-up pharmacogenetics study, breast cancer patients with CYP2D6 loss function variants had a higher risk of disease relapse and a lower incidence of hot flush (6). The clinical consequence of treating breast cancer and depression using tamoxifen and SSRIs was reviewed (7), and called for further investigation. This example clearly demonstrates that the translational significance of drug interaction studies relies on both clinical and molecular pharmacology evidences. As described by Hennessy and Flockhart (8), an integrated informatics, epidemiology, and pharmacology approach has the potential to accelerate the translational drug interaction studies. Pioneered by Tatonetti *et al*. (9), FDA adverse event reporting system (FAERS) and electronic medical records (EMR) were utilized to generate and validate Drug-ADE and drug-drug-ADE associations. Duke *et al.* proposed a text mining strategy for DDI molecular pharmacology evidence discovery from the public literature (10), which discovered 13,197 potential DDIs. In the follow up in vitro study, Han *et al*. validated the loratadine-simvastatin myotoxicity interaction, and its increased myopathy risk in both EMR and FAERS databases (11).

Driven by the emerging big data and novel computational models, there are three areas where translational biomedical informatics and pharmacometrics are having a major impact on the drug interaction research. First, during the past two decades, federal regulatory agencies, hospitals and research organizations maintained various patient databases such as spontaneous reporting system (SRS), electronic medical records (EMR) and electronic health records (EHR) for post-marketing surveillance and epidemiological studies. When these data are increasingly available to the research communities, computational models have been developed to identify and prioritize DDIs (12). Second, pharmacokinetics of DDIs have been well characterized and predicted with physiologically based pharmacokinetic (PBPK) models. Third, knowledge discovery through the literature has become a powerful approach for the DDI detection, in which the natural language processing (NLP) is the key computation technology.

A few reviews have highlighted some translational biomedical informatics approaches. For instance, the reviews by Koutias and Jaulent, and Harpaz *et al*. focused on computational models for SRS and EMR databases (12, 13). Text and data mining techniques to detect ADE signals were reviewed by Karimi *et al.* (14). Jensen *et al*. summarized available EMR/EHR databases and the obstacles for the EMR/EHR mining (15). However, these reviews did not focused on the translational nature in the ADE research, and none of them specifically addressed the DDI research. In this review, we focus on computational approaches for post marketing surveillance data mining, PBPK modeling, and literature based knowledge discovery, because these three approaches complement to each other. The rest of this review is organized as following: data mining methods for the post marketing surveillance are shown in section 2; PBPK DDI models and databases are presented in section 3; literature-based DDI discovery approaches are presented in section 4; and section 5 concludes this review.

2. DDI Data Mining Methods Using the Post-Marketing Surveillance Data

2.1 A Brief Review of Single Drug ADE Association Analyses

2.1.1 Univariate Disproportionality Analyses (DPAs)

DPAs are the pioneer approaches to quantify and prioritize single drug-ADE associations. For a drug-ADE pair, DPAs summarize data into a 2-by-2 contingency table, in which contains the frequencies classified by the usage of a drug (yes/no) and the occurrence of an ADE (yes/no). The outcome is the frequency that this drug-ADE pair is observed, and the expectation is the expected frequency of this drug-ADE pair under the assumption of no association. As its name, DPAs compare the outcomes to their expectations. DPAs can be classified as frequentist, Bayesian or empirical Bayesian. DPAs can be either used to analyze specific drug-ADE pairs of interest, or can conduct drug wide and ADE wide signal screening.

Proportional Reporting Ratio (PRR) and Reporting Odds Ratio (ROR) are frequentist DPAs (16, 17). ROR calculates the ratio of the ADE odds between the group of patients taking the drug and the other patients not taking the drug. PRR, on the other hand, calculates the ratio of two relative ADE risks between two patient groups. Practically, PRR_025 and ROR_025, the lower bound of 95% confidence intervals for PRR and POR, are often used for the signal detection, too. Likelihood Ratio Test (LRT) is another frequentist DPA (18). It assumed that the drug induced ADE frequency follows a Poisson distribution. Under the null hypothesis, this Poisson distribution had the same ADE rate as the background rate, i.e. the ADE rate for patients not taking the drug; and under the alternative hypothesis, they are not the same. The log-likelihood ratio statistics are then constructed to test this hypothesis. The LRT tests a drug and all ADEs at the same time, and the distribution of the maximum LRT can be calculated through the permutations.

Information Component (IC) is a Bayesian DPA (19). This approach assumes that the drug induced ADE frequency follows a binomial distribution itself; its expected frequency is calculated from the marginal drug frequency and ADE frequency; and the prior distribution of drug marginal frequency and ADE marginal frequency are assumed to be uniform distributions. The IC calculates the expected ratio between drug induced ADE frequency and its expected frequency under all these distributions assumptions. Later, Noren *et al.* introduced a joint Dirichelet distribution prior and extended Bate's IC model (20). Like PRR and ROR, signal detection using IC can be based on its lower bound of the 95% confidence interval (IC_025). Empirical Bayesian Geometric Mean (EBGM) is an empirical Bayesian DPA. Similar to the IC approach, EBGM calculates the expected ratio between drug induced ADE frequency and its expected frequency (21). However, different from the IC approach, a two-component mixture of gamma distributions was chosen to model the ratio, and this mixture model was further estimated from the data instead of pre-specified prior distribution. Bayesian False Discovery Rate (BFDR) is another empirical Bayesian DPA (22). For the above mentioned PRR, ROR and EBGM models, BFDR calculates the posterior probability for a predefined null hypothesis. For instance, BFDR was originally applied to the EBGM model (22); and later on, it was applied to the PRR too (23). BFDR itself can be used for signal detection.

Three-Component Mixture Model (3CMM) is an empirical Bayesian DPA developed by our group (24). Similar to the EBGM, 3CMM utilizes gamma-Poisson assumption as well. However, unlike EBGM, 3CMM has three distributions that characterize the ratio between drug-induced ADE frequency and its expected frequency; and local false discover rate (lfdr) is introduced for false positive control. Under 3CMM, the first distribution specifies the point mass distribution at 0 for the ratio; the second distribution has a mean ratio of 1; and the third one has a mean greater than 1. Particularly, the second distribution characterizes the null hypothesis, while the third distribution characterizes the alternative hypothesis. Hence, the lfdr estimates the probability of the null distribution conditional on the data and the 3CMM.

2.1.2 Multivariate Analyses

Univariate DPAs suffer from the confounding bias, which can be addressed in multivariate analysis. Tatonetti *et al.* assumes that confounding variables, such as co-morbidities, can be characterized by the co-medication variables (9). He applied logistic regression model first, and estimated the propensity score for each drug of interest. Then, in analyzing a drug-ADE association, this drug's propensity score was used to adjust the confounding variables.

Multiple logistic regression (MLR) and regulated logistic regression (RLR) are two other approaches in analyzing drug-ADE associations. MLR is a traditional statistical approach to detect drug-ADE association. It can be considered to be a multivariate extension of ROR. Usually, the MLR analyzes an ADE and all available drugs at the same time. Examples of applying MLR to EHR data can be found in Harpaz *et al*. (25). In certain situations, drug-ADE signal detection by MLR may involve a large number of drugs than sample sizes, where RLR becomes a viable solution, such as ridge and Lasso regression models. Example of signal detections by lasso regression models includes Caster *et al.* (26).

DPAs are less computationally expensive compared within other multivariate approaches (12). Additionally, DPAs can be either used to analyze specific drug-ADE pairs of interest, or can conduct drug wide and ADE wide signal screening. Though the disproportionality measurements may suffer from confounding bias, evaluations by gold standard shown DPAs to have decent performances (AUCs) (27). Hence, DPAs are routinely used for large scale hypothesis generation. Multivariate analyses, on the other hand, are typically observed in epidemiology studies to validate a few candidate drug ADE associations.

For logistic regression modeling, the number of predictors are usually less than two thousands, which is the similar to the number of FDA approved drugs. For pharmacovigilance databases, the sample sizes are usually up to a few millions. As a consequence, enhanced computational resources or smart techniques are required to handle the big data challenge. Our experiences indicate that a super computer with 50GB memory can handle MLR with a few hundred drugs and four million observations. With less powerful computational resources, bootstrap regression would be an ideal solution.

2.2 Drug interaction signal detection

Some of the DPAs in section 2.1 can be extended to detect drug interaction signals. By treating a drug combination as a new drug, the disproportionality measurements can be obtained, accordingly. For instance, Huang *et al*. introduced an extended LRT method which can be used for detecting signals for multiple drugs (or ADEs) in a drug class (or in an ADE group) (28). Likewise, an extended higher order IC method is proposed by Noren *et al.* (20). Higher order IC is based on the same model assumption as the traditional

IC, and its credibility interval can be derived similarly. They can not only be used for detecting the potential drug-drug interactions, but also can be used for detecting the association between a drug-ADE pair and another risk factor (e.g. age or gender). The examples for the extended EBGM can be found in Almenoff *et al*. and DuMouchel *et al.* (29, 30). Although these extended DPAs can be used for detecting the potential DDI signals, these approaches cannot distinguish the signals that are associated with drug interactions or just with independent drugs.

Noren *et al*. proposed a novel model for detecting two-way DDIs (31). In their model, a ratio of the DDI induced ADE risk and its expected ADE risk is calculated, and the expected ADE risk is calculated from the single drug induced ADE risk from both drugs and baseline ADE risk from neither drugs. Like the IC approach, a Bayes approach is taken to estimate the expected DDI risk ratio, and an uninformative prior was speculated for the prior. This prior has the advantage of shrinking the ratio toward 1 when the sample size is small.

The regression based method for detecting DDIs can avoid the confounding variable problems. Examples for the logistic regression model applied for detecting potential DDIs from SRS can be found in Van Puijenbroek *et al*. (32). Thakrar *et al*. proposed multiplicative and additive relationship to model the risks for single drugs and DDI pairs (33). The multiplicative model assumes that the risk associated with a drug multiplies with the background risk, and the additive model assumes that the risk associated with a drug is additive to the background risk. Their results show that the additive model is a more sensitive method for detecting signals and the multiplicative model may further help on qualifying the strength of the signals detected by the additive model. In addition to detecting the ADEs that were caused by the drug-drug interactions, the regression model can also be used for detecting the signals that one drug may reduce the ADEs of the other drug (i.e. beneficial effects of DDI).

2.3 High Dimensional Drug Interaction Detection

We recently developed a novel mixture drug-count response model (MDRM) to characterize and detect high dimensional drug interaction signals (34). MDRM is an empirical Bayesian method. This model assumes that the drug induced ADE follows two patterns: one pattern assumes a constant ADE risk regardless of the dimension of the drug combinations, while the other patter assumes that ADE risk increases like dose (i.e. drug counts) response curve. This model then estimates a probability for each drug combination who follows the drug-count response model. MDRM, for the first time, characterizes the pattern of high dimensional drug interactions and ADEs. Its innovation lies in the fact that MDRM allows different drug combinations to share the same drug-count response relationship, as the sample size of each drug combination goes very small when the dimension of the drug combination increases.

Currently, the amount of FDA approved drugs generate over millions of 2-way drug combinations; and as the dimension of drug combination increases, the amount of plausible drug combinations increases in a factorial speed. As traditional statistical models are insufficient to deal with tremendous amount of drug combinations, informatics approaches become a promising and practical solution. Two major informatics techniques to detect drug interaction signals include frequent closed itemset (FCI) mining and association rule mining. FCI is powerful on eliminating redundant drug combinations. For instance, if *drug A*, *drug B*, *ADE X*, *ADE Y* is a FCI, then its subsets (such as *drug A*, *ADE X*) are considered to be redundant. These redundant subsets can be removed unless such a subset appears in a record that does not contain all items of *drug A*, *drug B*, *ADE X*, *ADE Y*. Xiang *et al.* proposed a FCI-filter approach that integrated FCI mining and uninformative association removal to mine multiple drug interactions from the FAERS (35). Under their approach, potential itemsets are generated by FCI mining first; and uninformative itemsets are removed, if the itemsets and supporting transactions can be obtained from the interaction of other itemsets and their supporting transactions. An example of the application of association rule mining can be found in Harpaz *et al.* (36), in which Apriori algorithm is utilized to mine the FAERS data. Their Apriori configuration considers only itemsets that have a set of drugs in the antecedent and a set of ADEs in the consequent. Additionally, their prioritized itemsets are further filtered by the relative risks.

3. Pharmacokinetics Modeling and Data Sources

3.1 In vitro in vivo Drug Interaction Prediction Using Pharmacokinetics Modeling

There are two ways to characterize pharmacokinetics of drug. The top-down approach investigates clinical pharmacokinetic using clinical trial data, and it builds up a population pharmacokinetic model. The bottom-up approach, on the other hand, starts from pharmacokinetic data measured from in vitro studies, and extrapolates and predicts clinical drug exposure in humans. In this review, we will focus on one of the bottom-up approaches, steady state in vitro in vivo extrapolation (IVIVE) of drug interaction prediction. There are other great and comprehensive reviews on the bottom-up approach (37, 38). We select our focused IVIVE model because it is the one can be scaled up (i.e. including potentially all drugs), and interfaced with informatics analyses.

The ratio of area under the (substrate) concentration time curve (AUCR) in the present and absence of inhibitors is widely used to determine the severity of a DDI. Here, we focus on a static DDI model proposed by Ito *et al*. (39) and modified by Lu *et al*. (40), which calculates AUCR based on unbound inhibitor concentration ([I]), inhibition rate constant for a drug (Ki), fraction of metabolism (fm) and fraction of renal clearance (fe) [equation 1].

$$\mathrm{AUCR} = \frac{AUC(inhibited)}{AUC(uninhibited)} = \frac{1}{(1 - f_e)\sum_{i=1}^{n}\left[fm_i \times \frac{1}{1 + \left(\sum_{j=1}^{J}\frac{[I_j]}{ki_j}\right)}\right] + f_e} \qquad [1]$$

All these parameters can be obtained from various available data sources, except for fm. For example, Metabolism and Transport Drug Interaction Database (DIDB) has a collection of drug Ki, and Goodman and Gilman has a collection of fe and drug maximum concentration (Cmax) which can be used as [I] (41). There are several ways to estimate fm for a substrate. Firstly, change in AUC or clearance in the presence of a co-administered CYP inhibitor through a clinical PK study is used to determinate the contribution of the CYP for a drug. For example, Yeung *et al*. utilized clinical drug interaction studies, in which ketoconazole was used as the CYP3A4 probe inhibitor, and calculated a drug's fm in the CYP3A4 pathway using equation 2 (42):

$$fm_{3A4} = 1 - \frac{AUC(uninhibited)}{AUC(inhibited)} \qquad [2]$$

Secondly, pharmacogenetics studies can also be used to estimate fm through the fold-change in exposure of a substrate in extensive metabolizers (EMs) comparing to poor metabolizers (PMs) (39). A large population of patients were studied with respect to the metabolism of metoprolol, which was metabolized by CYP2D6 (43), and fm was calculated by equation 3:

$$fm_{2D6} = 1 - \frac{AUC(CYP2D6, EM, AVG)}{AUC(CYP2D6, PM, AVG)} = 1 - \frac{CL(CYP2D6, PM, AVG)}{CL(CYP2D6, EM, AVG)} \qquad [3]$$

Thirdly, in-vitro experiments also have been used to determine the contributions of several CYP pathways. Substrate depletion in the human liver microsomes (HLM) is one method that the drug is incubated with or without specific CYP selective inhibitors. The percent of inhibition can be calculated by comparing the metabolism rates with and without inhibitor. Substrate depletion can also be incubated with individual recombinant enzymes isoforms (44). Each isozyme contribution is estimated as the percent contribution of each CYP enzyme towards the total HLM CLint via a scaling factor (RAF/ISEF) approach (45). Recently, due to the success of the cryopreservation of human hepatocytes (46), hepatocyte suspension model (47) becomes a new method to estimate fm. Physiologically, cryopreserved human hepatocyte is closer to the human hepatic metabolism than the other in vitro system does. Desbans *et al*. (48) used cryopreserved human hepatocytes from 12 donors to estimate fm of CYP3A for five prototypical CYP3A substrates. After hepatocytes were incubated with test compounds and/or the inhibitor, the intrinsic clearance was estimated from the parent compound depletion profile. Then fmCYP3A was calculated from the ratio between CLint in absence and in presence of ketoconazole as equation 4:

$$fm_{3A} = 1 - \frac{CL_{int}\ (inhibited)}{CL_{int}\ (uninhibited)} \qquad [4]$$

Although there are several different methods successful to determine fm, there is no comprehensive database that systematically store fm for DDI research.

3.2 Adverse Drug Reactions Databases and Data Sources

There are a number of drug related databases, which integrate bioinformatics, cheminformatics and/or DDI knowledge, have been widely used for the drug interaction alerting in a large range of clinical decision support and electronic prescribing systems. Meanwhile, clinical signal-based databases can be helpful for understanding the mechanism of action for drugs (11). Also, part of pre-market drug development relies on the drug information and DDI knowledge to predict interactions between a new drug candidate and drugs currently on the market.

3.2.1 DDI related Database

DrugBank (49) is a well-known comprehensive database which contains bioinformatics and chemo-informatics resource of 9,591 drugs including molecule and biotech drugs. It combines detailed chemical, pharmacological and pharmaceutical information with comprehensive drug targets, such as sequence, structure or pathway information. All these can be useful for ADE research. In addition, DDI knowledge is included in the database. However, due to the simple description, additive and synergic interactions are hardly differentiated. Therefore, it is difficult to assure that an ADE is caused by a true interaction or simple dose increase. There are other similar comprehensive databases including Drugs.com (50) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (51). Some DDI knowledge database was derived from data mining from health record data sources. For example, the OFFSIDE database contains drug-event signals that are not listed on the FDA's official drug label (9). These signals were mined from FAERS data by a data-driven approach that removes many synthetic associations from indications, co-prescriptions, and hidden covariates. Using the same method, DDI signals were further derived from FAERS data, which is called as TWOSIDES (9).

Other databases, in which clinical information and mechanism knowledge are included, were derived from text mining and literature curating methods. By manually curating published literatures and FDA New Drug Application (NDA) reviews, DIDB collects in vitro and in vivo data of pharmacokinetics drug interactions. Unlike DrugBank, experimental conditions and results of DDI studies, which are crucial DDI factors, are all integrated in the DIDB database. Another important database is PharmGKB (52). It is one of the largest databases collecting associations among genes, drugs and diseases published in the literature. PharmGKB is well regarded as a reliable resource for personalized medicine and pathway-oriented DDI research (52). **Table 1** provides summarized main features of DDI related databases.

3.2.2. ADE-phenotyping sources

Our focused ADE-phenotype refers to an EHR-based patient cohort definition, which experiences an ADE (53). Here we provide ADE-phenotype sources, level of evidence, terminologies & data types, and their integration with EHR (**Table 2**). Four criteria of ADE evidence are given as the following: ADE definition algorithm validation (*criterion 1*), comprehensiveness of the ADE definition algorithm (*criterion 2*), literature and/or ADE-related evidence (*criterion 3*), and terminological support for structured and unstructured data (*criterion 4*). Using these four criteria, we define the three levels of ADE evidence.

Level I Evidence provides the highest reliable and precise ADE phenotypes. They met *criteria 1, 2*, and *(3 and/or 4).* For instance, the Phenotype Knowledgebase website (PheKB) (54), "an online environment supporting the workflow of building, sharing, and validating electronic phenotype algorithms" (54), which offers algorithms using approaches, such as ICD-9-CM codes, medications, and NLP. The PheKB's main goals are to improve algorithm transportability and validity across institutions (54) while achieving high positive predictive values (PPVs).

Level II Evidence ADE phenotypes met *criteria 3, and (2 and/or 4)*, but they have not been validated across institutions. For instance, Observational Medical Outcomes Partnership (OMOP) (55) has a library based on systematic literature review of a number of health outcomes of interest (HOIs) definitions to improve observational studies' reproducibility. OMOP also recognized that literature has usually been inconsistent in defining and reporting ADEs, and sometimes lacked of details of the exact codes and validations (56). For example, acute liver injury has eight different definitions, such as laboratory-based versus diagnostic procedures.

In addition to OMOP HOI library, UpToDate is another evidence-based physician-authored clinical guideline repository (57). It provides an evidence-based and manually curated clinical guidelines for ADEs. Although it does not directly define the ADE using the EHR data, it certainly can assist in defining ADEs. Furthermore, SIDER (58) is also a reliable source for ADE definitions.

Level III Evidence refers to terminology and vocabularies based data sources (*criterion 4*). For example, the medical dictionary for regulatory activities (MedDRA) is a key database for ADE (59), and the other database include CTCAE (60). Unlike MedDRA, CTCAE also contains the severity of ADE.

3.2.3. Database Integration

On the basis of these drug/DDI/ADE databases, some integrated databases combine them together, and form a complete dataset. The Drug Interaction Knowledge Base (DIKB), an evidence-based observed and predicted knowledge base, contains mechanisms and pharmacokinetic drug-drug interactions information for over 60 psychotropic and HMG-CoA reductase inhibitors (66). A rule-based metabolic DDI prediction was conducted with DIKB to determine the most optimal set of predictions (67). Further, Ayvaz (68) constructed an integrated potential DDI (PDDI) source by combining clinical-oriented information sources, natural language processing corpora, and bioinformatics/pharmacovigilance information sources by analyzing the overlap between the data sources and mapping drug entity to DrugBank ID. This dataset can benefit NLP corpora and lead to a better synthesis of PDDI knowledge. The merged data sources in the integrated database are descripted in **Table 3.**

From the translational research perspective, there are some limitations in these data sources. First, there are as yet few means to integrate different databases conveniently and economically. In PDDI database, DrugBank ID was used for medication standardization. Additionally, OMOP Common Data Model (CDM) could be used to standardize the format and content of the observational databases including medication, ADE, symptom and indication. However, complete solution for data integration is unavailable yet. Second, the DDI information in the databases are limited. Particularly, information including the DDI type (e.g. additive/synergic, Pharmacokinetics/Pharmacodynamics (PK/PD)), mechanism, clinical impact and quantitative description should be included and improved in the future data collection.

4. Knowledge discovery for drug interaction using text mining technologies

Literature-based knowledge discovery was pioneered by Don R. Swanson in 1986 and had been widespread for decades in the biomedical informatics domain (69). This technique bridges new relationships between existing knowledge by exploring the co-occurrence of words or phrases from different literature articles. Following this lead, many "open discovery" and "close discovery" methods were developed to discover interesting associations among a large set of data items. To distinguish open and close discovery, we take the relationship between a disease and treatments as an example, the open method can generate a hypothesis to find the underlying pathological mechanisms of a disease. It starts with a disease, discovers the mechanisms of the disease from literatures, and finally finds a drug that may interact with those mechanisms (intermediates). Differently, a close discovery method can verify and elaborate an initial hypothesis. Its searching process starts simultaneously from a disease and a drug. Their overlapping mechanisms (intermediates) can demonstrate the relationship between a disease and a drug (70). Based on these two concepts, in the last decade, several discovery systems were developed. Srinivasan presented both open and close algorithms to automatically discover a small set of interesting hypotheses from a suitable text collection using MeSH terms in Medline (71). Hristovski combined the outputs of two NLP systems to provide semantic prediction, which demonstrated the improvement for literature-based knowledge discovery (72). Tsuruoka developed a searching engine for Medline abstracts, called FACTA, which retrieves textual evidence of associations between the query terms and the concepts (73). Frijters developed CoPub discovery tool to assess the significance of co-occurrence based on the mutual information measure and mine the new relationships between biomedical concepts (74). Finally, Yetisgen-Yildiz proposed an evaluation methodology allowing the comparisons cross different systems (75).

While there have been many discovery methods developed, most of them often mined co-occurring entities from free-text in documents or data fields. The co-occurrences method has a critical drawback, since not all co-occurring entities possess "meaningful" and "quality" relations. To retrieve explicit fact from documents as efficiently as possible, text mining technologies facilitate quality discovery from biomedical literature, EHR, or Social media. Information Retrieval (IR) is the quality control process, which enables the identification of relevant documents and provides the quality of data resource for knowledge discovery. For example, the DDI IR step identifies higher quality DDI articles from PubMed (76). Information Extraction (IE) is the task of extracting information from unstructured text. The scope of extractions can be as simple as the predefined entities, such as the names of proteins, genes as well as drugs, or can be as complicated as the "true" associations between entities, such as drug-gene interactions or drug-drug interactions. Instead of co-occurrence-based knowledge, those applications automatically scrutinize the phase of generating quality information and potentially empower extracted information into truly novel hypotheses for open discovery or solid validations for close discovery.

In this section, our review will focus on how text mining technologies assist on the drug interaction discovery in three aspects: (1) The manually curated corpus facilitates text analysis by providing syntactic and semantic pharmacological knowledge for retrieving and extracting DDI. (2) The IR and IE technologies help aggregate quality data extensively, thereby providing the potential to perform hypothesis generations and validations. (3) Linking the dis-jointed sets of facts from documents uncovers hidden links between drugs and generates novel hypotheses.

4.1 Drug Interaction Corpora

Great TM methods rely on well-developed corpora. Corpora refers to manually annotated golden standard data. In the DDI TM domain, DDI corpora developed in both DDI Extraction challenge tasks in 2011 and 2013 (77, 78), have guided a great number of supervised DDI TM methodologies' development. The annotation strategies in corpus may differ subject to the purpose of TM tasks. There are three types of annotations in corpus. *1. Semantic annotation* creates semantic labels for terminologies or relationships (79, 80). *2. Syntactic annotation* includes structural make-up, part-of-speech tagging, and constituent or dependency parsing trees (81). *3. Fragment annotation* characterizes the properties of scientific text in specific measurements. Different from semantic and syntactic annotations, it provides sufficient generality to transcend the subject area. Fragment annotation was first designed to characterize text using five qualitative dimensions: focus, polarity, certainty, evidence, and directionality (82).

Although many corpora are available, only a few focus on the topic of DDIs (77-80, 83, 84). DDI Corpus 2011 and 2013 were built as reference standards for 2011 and 2013 DDI Extraction Challenges, respectively (77-79). These two corpora, consisting of 792 texts selected from Drugbank database and 233 Medline abstracts, were annotated with pharmacological substances and DDI relationships, including both PK and PD DDIs. The annotation schema includes drug entities (e.g. drug, brand, chemical agents, and drug groups) and DDI relationships (e.g. effect, mechanism, advice, or interaction). Another two corpora, PK DDI Corpus (83) and NLM CV DDI Corpus (84), were built up using drug product labels. PK DDI corpus comprises 64 labels. Two characteristics (type and role) are utilized to classify drug entities, and two properties (observed effect and experimental statement) are provided to model each PK DDI relationship. The types of drugs are active ingredient, drug product, or metabolite; and the roles of drugs are object and precipitant. The relationship between two co-administrated drugs are either positive or negative modality. The stated qualitative experimental data can also be used to identify drug interactions. NLM CV DDI Corpus of 180 cardiovascular drug product labels was developed, and acted as a reference standard for pharmacokinetic PDDI TM in product labeling. The annotation schema contains drug entities and DDI roles. Pharmacologic substances, including drugs, drug classes, and other substances (e.g. food) are annotated as entities. For the roles of drugs in the interaction, the schema from (83) was reused (i.e. object and precipitant for the role of interacting drugs or substances). In addition, authors further categorize interactions into “increase” and “decrease” classes. The final corpus, called PK corpus (80), was developed in our group. It was constructed to present four classes of PK abstracts: in vivo PK studies (n=56), in vivo pharmacogenetics studies (n=57), in vivo DDI studies (n=218), and in vitro DDI studies (n=210). A hierarchical three-level annotation schema was proposed to annotate key terms, drug interaction sentences, and drug interaction pairs. Except for drug names, this PK corpus was different from the other corpus, including enzyme, drug dosage, PK parameters with their values and units, mechanisms, action terms reflecting interactions are annotated. With regard to the relationship, DDIs were not only annotated based on their narrative descriptions, but also were judged using their quantitative and qualitative evidences. The fold change (FC) in PK parameters (e.g. FC > 1.5 or FC < 0.67 in AUC) or statistical measurement (e.g. P-value < 0.05) specifies the numeric rule to define DDI quantitatively. The significance statement (e.g. significantly, moderately, or probably) specifies the language expression pattern for DDI relationship qualitatively.

Other than the data recourses from biomedical literature or drug labels, social media, such as blog, forum, or Twitter, provide huge potential in the identification of ADEs and DDIs (85). In the past few years, corpora obtained from social media texts started emerging (86). A corpus of 10,822 tweets by Gonzalez lab was manually annotated for mining Twitter for ADRs (87). The annotation mainly focuses on drug names and ADRs. Different from the annotations in biomedical literature or drug labels, this corpus was sought to annotate not only the presence or absence of drug names or ADRs, but also to identify the span of expressions conveying individual ADR In addition, another corpus, also created by Gonzalez lab, consists of 267,215 Twitter posts. In this corpus, two sets of language models were created to encapsulate “semantic properties” by presenting word tokens as dense vectors and “n-gram sequences” by capturing sequential patterns (88). Moreover, TwiMed is one of the most recent corpus, which comprises 1,000 tweets and 1,000 PubMed sentences (86). The annotations covered entities (drug, symptom, and disease) and their relations (outcome-negative, outcome-positive, and reason-to-use). Similar to fragment annotation, their attributes for entities are further annotated to provide their characteristics (polarity, person, modality, exemplification, duration, severity, status, and sentiment).

In sum, all aforementioned corpora characterize different aspects of DDI studies. DDI corpus focused on the distinction in drug type and DDI effect (77, 79); PK DDI corpus and NLM CV DDI Corpus annotated package inserts as the data sources and identified the roles of drugs in DDI relationships (83, 84); PK corpus further differentiated PK DDI into in vivo and in vitro studies, and define drug interactions using experimental evidences (80). The corpora for social media were annotated differently from those in literature. Two corpora, created by Gonzalez lab, were annotated in different scopes (87, 88). One focused on entity level and another focused on language models. TwiMed not only annotated with both entity and relation levels but also identify the attributes for entities (86).

4.2 Information Retrieval and Extraction for Drug Interaction and Drug-related Knowledge

In order to promote DDI TM, DDI-Extraction challenges organized in 2011 and 2013 aimed for developing the TM methodologies of the pharmacological substance recognition and DDI detection (77, 78). For the named entity recognition of pharmacological substances, the best results were achieved by WBI_NER. This NER approach is formulated as a sequence labelling task (IOB format). Using domain-independent features from ChemSpot, Jochem, and ChEBI ontology, linear-chain conditional random field model was implemented to predict the sequences of name entities. The second best method (NLM LHC) utilized dictionaries from multiple biomedical resources, such as Drugbank, ATC system, or MeSH headings. In this challenge, most approaches can perform well on the recognition of generic or brand names, but not drug-n category (substances not approved for human use). The great variation and complex in naming convention lead to the difficulty in name recognition. Another focus in DDI-Extraction 2011 challenge is to identify true DDIs from all possible DDI pairs from the biomedical text in Medline abstracts and Drugbank. Among 10 participation computational algorithms, the best performance (F-measure=0.657) was achieved by the system (WBI) using ensemble learning approach. Combined three different kernels (all-paths graph, shallow linguistic, and k-band shortest path spectrum kernels) with a case-based reasoning (CBR) called MOARA, a majority voting ensemble of constructing machine learning methods was built for binary prediction. The DDI-Extraction 2011 concluded that approaches using kernel-based methods achieved better performances than the feature-based methods. In addition, most systems used primarily syntactic information, but not much semantic information. Different from the 2011 challenge, DDI-Extraction 2013 not only aimed to detect DDI pairs, but also classified them into one of the following four types: advice, effect, mechanism, and interaction statement. In the 2013 challenge, FBK-irst achieved best performance and yielded an F-score of 0.80 for DDI detection and an F-score of 0.65 for DDI detection and classification. It applied a hybrid kernel based method and exploited the scope of negations and semantic roles for filtering negative instances. The 2013 challenge concluded that the systems using non-linear kernel-based methods outperformed linear SVM systems.

Other than DDI Corpus in previous two challenges, PK corpus (80) was also utilized for developing DDI extraction tools. The extraction tasks were implemented in the in vivo and in vitro DDI corpus separately using the approach with all paths graph kernel. Interestingly, huge discrepancy on the performance was found between two sub-corpora in the PK corpus. The reported F-measure of in vivo DDI corpus, 0.76, is much higher than that of in vitro DDI corpus (0.52). Authors concluded that DDI representations in in vitro

PK study were more diverse than those in in vivo PK study. It usually contains more drugs and PK parameters to describe DDI evidences, and it compares their inhibition/induction capability in a long sentence. Using the same dataset (PK corpus), Zhang *et al*. presented a graphic kernel based approach to combine syntactic and semantic information for extracting pharmacokinetic drug interaction (89). Compared with the previous all paths graph kernel methods (80), this new method further utilized semantic annotations from PK corpus and the F-measures were improved from 75.91 % to 81.94 % on the in vivo dataset and from 51.50 % to 69.34 % on the in vitro dataset, respectively.

Learned from the previous works (80, 89), clearly, the performance of extracting PK DDI evidences would be varied if their experiment methods were different. For achieving better performances, it is important for a text mining system to treat DDI evidences differently according to their study types. In vitro studies investigate whether a drug is a substrate, inhibitor, or inducer of metabolizing enzymes or transporters; in vivo PK studies investigate the kinetics of drug metabolism involved in absorption, distribution, metabolism, and excretion (ADME) process, and clinical studies investigate the clinical effects, i.e. efficacy or side effects of DDIs. Recent work by Kolchinsky (76) classified the in vitro and in vivo PK DDI evidences. More recently, Wu developed a suite of text mining tools to explore and distinguish three different types of DDI evidences, namely in vitro PK, in vivo PK and clinical PD (90). A large-scale mining from 25 million abstracts in PubMed (1975-2015) was accomplished to retrieve DDI relevant abstracts and identify DDI pairs for each study. The result shows that 986 DDI pairs with all three types of evidences have their clinical usages. 2,157 DIDs with known clinical PK/PD DDI evidences and 13,012 DDIs with only clinical PD evidence have enormous research potentials. This result pointed out knowledge gaps and potentially gives an impetus to translational drug interaction research.

Besides data mining using the post-marketing surveillance data or text mining using the scientific literature, social media provide different promising resources for identifying DDIs and ADEs. Social media databases are based on direct experiences from drug users. Thus, they provide up-to-date and timely messages conveying drug related information (91). Due to the unique issues of social media content, including credibility, uniqueness, frequency, and salience of the data (92), the existing IR and IE techniques for scientific literature may not be effective for social media data. To this end, many works were developed in the past few years. Sarker focused on the classification of sentences to detect ADR mentions utilizing features, including n-gram, UMLS semantic types, Synset expansion, etc. (93). By the same authors, the distribution word representations were generated to capture different types of semantic information and an n-gram sequential language model was used to capture sequential word occurrence probability. Utilizing both information facilitates the text classification and text normalization for drug-related knowledge (88). Except for the commonly used features extracted from narratives, sentiment analysis features is valuable for improving the performance of detection. Korkontzelos demonstrated that the features extracted from Twitter data using sentiment analysis can achieve a statistical F-measure increase. For information extraction, Carbonell analyzed the features using time series analysis, co-evaluate the mentions of drugs in Twitter within intervals of 30 minutes, and explore the potential drug effects and drug interactions (94). Another tool called ADRMine utilized a variety of features, including a new feature for modeling words' semantic similarities (95). Using conditional random fields (CRF) classifier, the similarities are modeled by clustering words based on word representation vectors (embeddings) generated from unlabeled user posts in social media. This work proved that word cluster features can significantly improve extraction performance for mining adverse drug reaction mentions.

4.3 Information discovery for novel drug interaction and ADE

Information retrieval and extraction for drug interaction evidence from biomedical literature lend an impetus to the generation of “meaningful” and “quality” evidences, which helps on aggregating DDIs and improving the coverage of DDI databases. However, an overlapping analysis between Drugbank and Micromedex showed that there are around 25% of disagreements (96). The lack of scientific evidences complicates the process of verifying the discrepancies. Therefore, to explore the mechanism behind drug interaction, it is crucial to supply the necessary scientific evidence to validate DDIs.

To discover novel drug interactions and explore their mechanisms, knowledge discovery strategy had been widely employed. Both Tari and Percha are two typical examples of close discovery method. Tari developed a method combining text mining and automated reasoning to infer DDIs with the support of enzyme and biological domain knowledge (97). By representing the general knowledge related to the metabolism (drug-gene) and biological interaction (protein-protein) with the logic rules, DDIs were predicted in the reasoning phase. In a different paper, Percha proposed a novel approach to predict novel DDIs by aggregating gene-drug interactions which are extracted via rule-based method (98). Using the established DDIs as the training set, a supervised classifier was trained to score potential DDIs based on the normalized drug-gene assertions extracted from the literature that relate two drugs to a gene product. More significantly, a semantic network built based on the extracted drug-gene assertions were implemented to explain the pharmacological mechanisms for newly-predicted DDIs. Different from Tari and Percha's methods, Duke proposed a literature discovery approach combined with analysis of electronic medical records (EMRs) and predicted 13,197 CYP-related DDIs (10). Based on literature data on in vitro drug metabolism and inhibitory potency, this translational approach finally identified 5 novel drug interactions that synergistically increased the risk of myopathy.

Other than DDI prediction, identifying ADEs caused by DDIs using text mining approaches draws more and more attention. Recent approaches utilized the features that drug interaction with the same gene targets may lead to ADEs and drugs with similar structures for ADE predictions. In this fashion, Raja proposed a literature-mining framework to enhance the prediction of DDIs and ADE types through integrating drug-gene interactions (99). Using the DDI features from DDI corpus, a supervised learning categorized ADEs into four types: adverse effect, effect at molecular level, effect related to pharmacokinetics, and DDI without known ADEs. This tool was applied to predict DDIs and ADE types related to cutaneous diseases and successfully identify promising new ADEs.

Interestingly, an example of Twitter applicability in knowledge discovery for drug interactions is proposed by Hamed *et al.* (100). This tool called HashPairMiner majorly used hashtags in computational analysis to discover novel DDI pairs. Based on the computation of associations for co-occurred keywords in the same tweets and associations between keywords and hashtags that also appeared in the same tweet, a new network mining algorithm was created to detect connections between pairs of drugs. This work demonstrated how hashtags can connect information and synthesize new knowledge.

5. Conclusion

In this article, we review three essential computation and informatics approaches for the translational drug interaction research. First, we provide an overview for computational models for mining drug interaction signals from post-marketing surveillance databases. Second, we present PK models for in vitro in vivo extrapolation in DDI prediction. We particularly emphasize the value of fm in the DDI prediction. We also review and summarize available DDI related databases, ADE-phenotyping sources and integrated DDI databases. Third, we show diverse text mining techniques to discover ADEs and drug interactions from literatures and social media. Signals identified by each approach can serve as potential drug interaction hypotheses. Although significant progresses and achievements have been made for each of these approaches separately, researchers rarely utilize them jointly for drug interaction hypothesis generation and knowledge discovery. In the real world, these three approaches are naturally complementary to each other. On one hand, drug interactions shall or may initially manifest in clinical practices and reported to the clinical databases, and consequently can be detected by the post-marketing surveillance data mining. On the other hand, in vitro experiments together with in vitro in vivo models are well established to evaluate drug interaction PK evidence and validate their mechanisms. Nevertheless, findings of clinical drug interaction signals and in vitro drug interaction mechanisms are published in the research community. Effective literature-based knowledge discovery approaches will enhance drug interaction research by providing both clinical and in vitro drug interaction knowledge, or identify DDI knowledge gap. This review shall help scientists to integrate all these translational biomedical informatics analyses for an improved translational drug interaction research. Most importantly, we hope this review to stimulate novel and creative translational biomedical informatics methods for the drug interaction research.

References:

1. FDA. https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070434.htm (2015). Accessed 12 August 2017.
2. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA. 1998;279(15):1200-5.
3. Prybys K, Gee A. Polypharmacy in the elderly: clinical challenges in emergency practice. Part. 2002;1:145-51.
4. Stearns V, Johnson MD, Rae JM, Morocho A, Novielli A, Bhargava P, et al. Active tamoxifen metabolite plasma concentrations after coadministration of tamoxifen and the selective serotonin reuptake inhibitor paroxetine. Journal of the National Cancer Institute. 2003;95(23):1758-64.
5. Desta Z, Ward BA, Soukhova NV, Flockhart DA. Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system in vitro: prominent roles for CYP3A and CYP2D6. The Journal of pharmacology and experimental therapeutics. 2004;310(3):1062-75.
6. Goetz MP, Rae JM, Suman VJ, Safgren SL, Ames MM, Visscher DW, et al. Pharmacogenetics of tamoxifen biotransformation is associated with clinical outcomes of efficacy and hot flashes. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2005;23(36):9312-8.
7. Henry NL, Stearns V, Flockhart DA, Hayes DF, Riba M. Drug interactions and pharmacogenomics in the treatment of breast cancer and depression. The American journal of psychiatry. 2008;165(10):1251-5.
8. Hennessy S, Flockhart DA. The need for translational research on drug-drug interactions. Clinical pharmacology and therapeutics. 2012;91(5):771-3.
9. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med. 2012;4(125):125ra31.
10. Duke JD, Han X, Wang Z, Subhadarshini A, Karnik SD, Li X, et al. Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. PLoS Comput Biol. 2012;8(8):e1002614.
11. Han X, Quinney SK, Wang Z, Zhang P, Duke J, Desta Z, et al. Identification and Mechanistic Investigation of Drug–Drug Interactions Associated With Myopathy: A Translational Approach. Clinical Pharmacology & Therapeutics. 2015;98(3):321-7.
12. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical pharmacology and therapeutics. 2012;91(6):1010-21.
13. Koutkias VG, Jaulent MC. Computational approaches for pharmacovigilance signal detection: toward integrated and semantically-enriched frameworks. Drug Saf. 2015;38(3):219-32.
14. Karimi S, Wang C, Metke-Jimenez A, Gaire R, Paris C. Text and Data Mining Techniques in Adverse Drug Reaction Detection. Acm Comput Surv. 2015;47(4).
15. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395-405.
16. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidem Dr S. 2001;10(6):483-6.
17. van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidem Dr S. 2002;11(1):3-10.
18. Huang L, Zalkikar J, Tiwari RC. A Likelihood Ratio Test Based Method for Signal Detection With Application to FDA's Drug Safety Data. J Am Stat Assoc. 2011;106(496):1230-41.
19. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54(4):315-21.
20. Noren GN, Bate A, Orre R, Edwards IR. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. Stat Med. 2006;25(21):3740-57.

21. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat. 1999;53(3):177-90.

22. Ahmed I, Haramburu F, Fourrier-Reglat A, Thiessard F, Kreft-Jais C, Miremont-Salame G, et al. Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. Stat Med. 2009;28(13):1774-92.

23. Ahmed I, Dalmasso C, Haramburu F, Thiessard F, Broet P, Tubert-Bitter P. False Discovery Rate Estimation for Frequentist Pharmacovigilance Signal Detection Methods. Biometrics. 2010;66(1):301-9.

24. Zhang P. Study Designs and Statistical Methods for Pharmacogenomics and Drug Interaction Studies: Indiana University; 2016.

25. Harpaz R, Haerian K, Chase HS, Friedman C. Mining electronic health records for adverse drug effects using regression based methods. IHI '10 Proceedings of the 1st ACM International Health Informatics Symposium. 2010:Pages 100-7.

26. Ola Caster, G. Niklas Norén, David Madigan, Bate A. Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database. Statistical Analysis and Data Mining. 2010;3:197–208.

27. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System. Clinical Pharmacology & Therapeutics. 2013;93(6):539-46.

28. Huang L, Zalkikar J, Tiwari RC. Likelihood Ratio Test-Based Method for Signal Detection in Drug Classes Using FDA's AERS Database. J Biopharm Stat. 2013;23(1):178-200.

29. Almenoff JS, DuMouchel W, Kindman LA, Yang XH, Fram D. Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. Pharmacoepidem Dr S. 2003;12(6):517-21.

30. DuMouchel William, Pregibon D. Empirical bayes screening for multi-item associations. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. 2001.

31. Noren GN, Sundberg R, Bate A, Edwards IR. A statistical methodology for drug-drug interaction surveillance. Stat Med. 2008;27(16):3057-70.

32. van Puijenbroek EP, Egberts ACG, Meyboom RHB, Leufkens HGM. Signalling possible drug-drug interactions in a spontaneous reporting system: delay of withdrawal bleeding during concomitant use of oral contraceptives and itraconazole. British journal of clinical pharmacology. 1999;47(6):689-93.

33. Thakrar BT, Grundschober SB, Doessegger L. Detecting signals of drug-drug interactions in a spontaneous reports database. British journal of clinical pharmacology. 2007;64(4):489-95.

34. Zhang P, Du L, Wang L, Liu M, Cheng L, Chiang CW, et al. A Mixture Dose-Response Model for Identifying High-Dimensional Drug Interaction Effects on Myopathy Using Electronic Medical Record Databases. CPT Pharmacometrics Syst Pharmacol. 2015;4(8):474-80.

35. Xiang Y, Albin A, Ren K, Zhang P, Etter JP, Lin S, et al. Efficiently mining Adverse Event Reporting System for multiple drug interactions. AMIA Jt Summits Transl Sci Proc. 2014;2014:120-5.

36. Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC bioinformatics. 2010;11 Suppl 9:S7.

37. Jamei M, Dickinson GL, Rostami-Hodjegan A. A Framework for Assessing Inter-individual Variability in Pharmacokinetics Using Virtual Human Populations and Integrating General Knowledge of Physical Chemistry, Biology, Anatomy, Physiology and Genetics: A Tale of 'Bottom-Up' vs 'Top-Down' Recognition of Covariates. Drug Metab Pharmacok. 2009;24(1):53-75.

38. Yeo KR, Jamei M, Rostami-Hodjegan A. Predicting drug-drug interactions: application of physiologically based pharmacokinetic models under a systems biology approach. Expert Rev Clin Pharmacol. 2013;6(2):143-57.

39. Ito K, Hallifax D, Obach RS, Houston JB. Impact of parallel pathways of drug elimination and multiple cytochrome P450 involvement on drug-drug interactions: CYP2D6 paradigm. Drug metabolism and disposition: the biological fate of chemicals. 2005;33(6):837-44.

40. Lu C, Miwa GT, Prakash SR, Gan LS, Balani SK. A novel model for the prediction of drug-drug interactions in humans based on in vitro cytochrome p450 phenotypic data. Drug metabolism and disposition: the biological fate of chemicals. 2007;35(1):79-85.
41. Goodman LS, Gilman A, Brunton LL, Lazo JS, Parker KL. Goodman & Gilman's the pharmacological basis of therapeutics. 11th ed. New York: McGraw-Hill; 2006. xxiii, 2021 p. p.
42. Yeung CK, Yoshida K, Kusama M, Zhang H, Ragueneau-Majlessi I, Argon S, et al. Organ Impairment-Drug-Drug Interaction Database: A Tool for Evaluating the Impact of Renal or Hepatic Impairment and Pharmacologic Inhibition on the Systemic Exposure of Drugs. CPT Pharmacometrics Syst Pharmacol. 2015;4(8):489-94.
43. McGourty JC, Silas JH, Lennard MS, Tucker GT, Woods HF. Metoprolol metabolism and debrisoquine oxidation polymorphism--population and family studies. British journal of clinical pharmacology. 1985;20(6):555-66.
44. Li ZM, Guo LH, Ren XM. Biotransformation of 8:2 fluorotelomer alcohol by recombinant human cytochrome P450s, human liver microsomes and human liver cytosol. Environ Sci Process Impacts. 2016;18(5):538-46.
45. Bohnert T, Patel A, Templeton I, Chen Y, Lu C, Lai G, et al. Evaluation of a New Molecular Entity as a Victim of Metabolic Drug-Drug Interactions-an Industry Perspective. Drug metabolism and disposition: the biological fate of chemicals. 2016;44(8):1399-423.
46. Stephenne X, Najimi M, Sokal EM. Hepatocyte cryopreservation: is it time to change the strategy? World J Gastroenterol. 2010;16(1):1-14.
47. Mao J, Mohutsky MA, Harrelson JP, Wrighton SA, Hall SD. Prediction of CYP3A-mediated drug-drug interactions using human hepatocytes suspended in human plasma. Drug metabolism and disposition: the biological fate of chemicals. 2011;39(4):591-602.
48. Desbans C, Hilgendorf C, Lutz M, Bachellier P, Zacharias T, Weber JC, et al. Prediction of fraction metabolized via CYP3A in humans utilizing cryopreserved human hepatocytes from a set of 12 single donors. Xenobiotica. 2014;44(1):17-27.
49. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic acids research. 2013;42(D1):D1091-D7.
50. Drugs.com. www.drugs.com (2000). Accessed 12 August 2017.
51. KEGG. http://www.genome.jp/kegg/ (1995). Accessed 12 August 2017.
52. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clinical pharmacology and therapeutics. 2012;92(4):414-7.
53. Li Q, Melton K, Lingren T, Kirkendall ES, Hall E, Zhai H, et al. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. J Am Med Inform Assoc. 2014;21(5):776-84.
54. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016;23(6):1046-52.
55. Fox BI, Hollingsworth JC, Gray MD, Hollingsworth ML, Gao J, Hansen RA. Developing an expert panel process to refine health outcome definitions in observational data. J Biomed Inform. 2013;46(5):795-804.
56. Hansen RA, Gray MD, Fox BI, Hollingsworth JC, Gao J, Zeng P. How well do various health outcome definitions identify appropriate cases in observational studies? Drug Saf. 2013;36 Suppl 1:S27-32.
57. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, Kiefer R, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. Journal of the American Medical Informatics Association. 2015;22(6):1220-30.
58. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic acids research. 2016;44(D1):D1075-9.

59. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf. 1999;20(2):109-17.

60. Chen AP, Setser A, Anadkat MJ, Cotliar J, Olsen EA, Garden BC, et al. Grading dermatologic adverse events of cancer treatments: the Common Terminology Criteria for Adverse Events Version 4.0. J Am Acad Dermatol. 2012;67(5):1025-39.

61. Chen Y, Ghosh J, Bejan CA, Gunter CA, Gupta S, Kho A, et al. Building bridges across electronic health record systems through inferred phenotypic topics. J Biomed Inform. 2015;55:82-93.

62. Hsu W, Gonzalez NR, Chien A, Pablo Villablanca J, Pajukanta P, Vinuela F, et al. An integrated, ontology-driven approach to constructing observational databases for research. J Biomed Inform. 2015;55:132-42.

63. Wilson PS, Scichilone RA. LOINC as a data standard: how LOINC can be used in electronic environments. J Ahima. 2011;82(7):44-7.

64. Systematized Nomenclature Of Medicine-Clinical Terms (SNOMED CT). http://www.ihtsdo.org/snomed-ct/ (2011). Accessed 12 August 2017.

65. Lin C, Karlson EW, Dligach D, Ramirez MP, Miller TA, Mo H, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. Journal of the American Medical Informatics Association. 2015;22(e1):e151-61.

66. Boyce R, Collins C, Horn J, Kalet I. Computing with evidence Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment. J Biomed Inform. 2009;42(6):979-89.

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

68. Ayvaz S, Horn J, Hassanzadeh O, Zhu Q, Stan J, Tatonetti NP, et al. Toward a complete dataset of drug-drug interaction information from publicly available sources. J Biomed Inform. 2015;55:206-17.

69. Bekhuis T. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. Biomedical Digital Libraries. 2006;3:2-.

70. Weeber M, Kors JA, Mons B. Online tools to support literature-based discovery in the life sciences. Briefings in bioinformatics. 2005;6(3):277-86.

71. Srinivasan P. Text mining: generating hypotheses from MEDLINE. Journal of the Association for Information Science and Technology. 2004;55(5):396-413.

72. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annual Symposium proceedings AMIA Symposium. 2006:349-53.

73. Tsuruoka Y, Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition. J Biomed Inform. 2004;37(6):461-70.

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

75. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. J Biomed Inform. 2009;42(4):633-43.

76. Kolchinsky A, Lourenço A, Wu HY, Li L, Rocha LM. Extraction of Pharmacokinetic Evidence of Drug-drug Interactions from the literature. PLoS Comput Biol. 2014:Submitted.

77. Segura-Bedmar I, Martınez P, Sánchez-Cisneros D, editors. The 1st DDIExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011; 2011; Spain.

78. Segura-Bedmar I, Martinez P, Herrero-Zazo M. Lessons learnt from the DDIExtraction-2013 Shared Task. J Biomed Inform. 2014;51:152-64.

79. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. J Biomed Inform. 2013;46(5):914-20.

80. Wu HY, Karnik S, Subhadarshini A, Wang Z, Philips S, Han X, et al. An integrated pharmacokinetics ontology and corpus for text mining. BMC bioinformatics. 2013;14:35.

81. Tateisi Y, Yakushiji A, Ohta T, Tsujii Ji, editors. Syntax Annotation for the GENIA corpus. Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005), Jeju Island, Korea, October; 2005.
82. Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC bioinformatics. 2006;7(1):356.
83. Boyce R, Gardner G, Harkema H, editors. Using natural language processing to extract drug-drug interaction information from package inserts. BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing.
84. Stan J, editor A machine-learning approach for drug-drug interaction extraction from FDA structured product labels. National Library of Medicine Training Conference, Pittsburgh PA, USA.
85. Vilar S, Friedman C, Hripcsak G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Briefings in bioinformatics. 2017.
86. Alvaro N, Miyao Y, Collier N. TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations. JMIR public health and surveillance. 2017;3(2):e24.
87. Ginn R, Pimpalkhute P, Nikfarjam A, Patki A, O’Connor K, Sarker A, et al., editors. Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing; 2014.
88. Sarker A, Gonzalez G. A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. Data in brief. 2017;10:122-31.
89. Zhang Y, Wu HY, Xu J, Wang J, Soysal E, Li L, et al. Leveraging syntactic and semantic graph kernels to extract pharmacokinetic drug drug interactions from biomedical literature. BMC systems biology. 2016;10 Suppl 3:67.
90. Wu H-Y, Zhang S, Desta Z, Quinney S, Li L, editors. TRANSLATIONAL DRUG INTERACTION EVIDENCE GAP DISCOVERY USING TEXT MINING. CLINICAL PHARMACOLOGY & THERAPEUTICS; 2017: WILEY-BLACKWELL 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.
91. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. Drug safety. 2014;37(10):777-90.
92. Abbasi A, Adjeroh D, Dredze M, Paul MJ, Zahedi FM, Zhao H, et al. Social media analytics for smart health. IEEE Intelligent Systems. 2014;29(2):60-80.
93. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform. 2015;53:196-207.
94. Carbonell P, Mayer MA, Bravo A. Exploring brand-name drug mentions on Twitter for pharmacovigilance. Studies in health technology and informatics. 2015;210:55-9.
95. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association : JAMIA. 2015;22(3):671-81.
96. Wong CM, Ko Y, Chan A. Clinically significant drug-drug interactions between oral anticancer agents and nonanticancer agents: profiling and comparison of two drug compendia. The Annals of pharmacotherapy. 2008;42(12):1737-48.
97. Tari L, Anwar S, Liang S, Cai J, Baral C. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics (Oxford, England). 2010;26(18):i547-53.
98. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2012:410-21.
99. Raja K, Patrick M, Elder JT, Tsoi LC. Machine learning workflow to enhance predictions of Adverse Drug Reactions (ADRs) through drug-gene interactions: application to drugs for cutaneous diseases. Scientific reports. 2017;7(1):3690.
100. Hamed AA, Wu X, Erickson R, Fandy T. Twitter K-H networks in action: Advancing biomedical literature for drug search. J Biomed Inform. 2015;56:157-68.

Table 1. Summary of Drug/DDI-based databases

Database Name	**Data Type in Database**	**Data Sources**	**Main feature(s)**	**DDI related Shortcoming(s)**
DrugBank	Bioinformatics/ Cheminformatics/DDI	Manual Search/merged with many other databases	▪ DrugBank collects 8261 small molecule and biotech drugs including approved, withdraw and experimental drugs ▪ Chemical, pharmacological, pharmaceutical information and DDI knowledge are combined in the database	▪ Simple details in DDI ▪ No additive or synergic information for DDI
OFFSIDES	Drug-ADE relationship	Signal Detection in AERS	▪ OFFSIDES database contains 438,801 drug-event signals connecting 1332 drugs and 10,097 adverse events ▪ These effects are not listed on the FDA's official drug label ▪ Confidence is signed for each relationship	-
TWOSIDES	DDI-ADE relationship	Signal Detection in AERS	▪ 868,221 significant associations are included ▪ Associations are limited to new-found ones ▪ PD DDI and PK DDI are included	▪ No additive or synergic information for DDI ▪ PK and PD DDI are not classified
DIDB	In vitro and in vivo data of PK DDI	Manually curating published literatures	▪ DIDB collects in vitro and in vivo data of PK DDI.	▪ No additive or synergic information

			▪ Experimental conditions and results of DDI studies are all integrated	for DDI ▪ Only PK DDI are included
PharmGKB	Pharmacogenetics and pharmacogenomics knowledge	Literature and drug label reviews	▪ PharmGKB is one of the largest databases in pharmacogenetics and pharmacogenomics knowledge ▪ Gene-drug associations, drug-centered pathway and gene-drug-disease relationships are included via literature and drug label reviews	-

Table 2. Sources for ADE-phenotyping

Source Name	**Level of evidence**	**Source Description**	**Terminologies and datatypes**	**Integration into EHR**
Medical Dictionary for Regulatory Activities (MedDRA) (59)	***Level III***	A unified standard terminology for recording and reporting adverse drug events.	From higher to lower levels: System Organ Class (SOC), High-Level Group Terms (HLGT), High-Level Terms (HLT), Preferred Terms (PT), and Lowest Level Terms (LLT).	• Used in structured data or unstructured clinical narratives.
Current Procedural Terminology (CPT) (61)	***Level III***	A medical terminology to bill outpatient & office procedures.	Category I, Category II, and Category III codes	• Used in structured data or unstructured clinical narratives.
International Classification of Diseases (ICD) (62)	***Level III***	An international diagnostic classification standard codes for clinical, and research purposes.	Hierarchical comprehensive classification of diseases, signs, symptoms, and procedures	• Used in structured data or unstructured clinical narratives.
Logical Observation Identifiers Names and Codes (LOINC) (63)	***Level III***	A common language for identifying health measurements, observations, and documents.	Set of identifiers, names, and codes. Mostly used for laboratory tests concepts.	• Used in structured data or unstructured clinical narratives.
The Systematized Nomenclature of Medicine (SNOMED	***Level III***	A multilingual clinical terminology to address the requirement for effective Electronic Health Record	Hierarchical representation of detailed clinical information, e.g. top level concepts, such as	• Used in structured data or unstructured clinical narratives.

CT) (64)		(EHR).	clinical finding, procedure, and substance.	
RxNorm (65)	***Level III***	A normalized naming system for generic and branded drugs that supports interoperability between clinical systems.	Normalized names and unique identifiers for medicines and drugs linked to their ingredients, strength, and dose forms.	• Used in structured data or unstructured clinical narratives.
Common Terminology Criteria for Adverse Events (CTCAE) (60)	***Level III***	A comprehensive, multimodality grading system for reporting adverse drug effects (ADEs) of cancer treatment.	AEs terms associated with 5-point severity scale of ADE, and mapped to MedDRA Lowest Level Term (LLT) to supports standardization of ADEs terms in EHR.	• Used in structured data or unstructured clinical narratives. • Severity scale of ADEs provides additional evidence.
The SIDER database of drugs and side effects (SIDER) (58)	***Level II***	A computer-readable side effect resource/database mined from FDA drug labels, contains about 1,430 drugs, 5,868 side effects (SE), and 139,756 drug-SE pairs.	Connects drugs to their recorded ADEs terms, provides frequency information, occurrence of ADEs, and drug indications. ADEs are mapped to MedDRA-preferred terms.	• Used in structured data or unstructured clinical narratives. • Used for mapping drugs to ADEs.
UpToDate (57)	***Level II***	An evidence-based, physician-authored clinical decision support tool.	Synthesized medical information, such as clinical guidelines, graded recommendation, and drug	• Used in structured data or unstructured clinical narratives. • Evidence-based medical information and drug

			entries and interactions.	interactions assist in defining ADEs. • An up-to-date clinical guidelines
Observational Medical Outcomes Partnership (OMOP) (55)	***Level II***	Literature-based Health Outcome of Interest (HOI) definitions library of conditions that have relevant to drug toxicities, medical significance, and/or public health.	ICD, CPT, SNOMED CT, LOINC, diagnostic or therapeutic procedures, and lab values	• Used in structured data or unstructured clinical narratives. • Broad and narrow definitions can be implemented directly into EHR based on users' needs
The Phenotype Knowledgebase website (PheKB) (54)	***Level I***	A collaborative environment to build and validate phenotyping algorithms.	ICD, CPT, Laboratories, Medications, Natural Language Processing, Vital Signs	• Used in structured data or unstructured clinical narratives. • Comprehensive validated definitions and/or algorithms can be implemented into EHR based on users' needs

Table 3. Summary of Integrated database

Database Name	Data Type in Database	Data Sources
DIKB	Mechanisms and pharmacokinetic drug-drug interactions information with confidence	▪ Retrospective studies ▪ clinical trials ▪ metabolic inhibition identification & inhibition catalysis identification ▪ statements, reviews and observational reports ▪ phenotyping definition including MeSH, WordNet and NCI Thesaurus
Merged PDDI	Potential DDI	a. 5 clinically-oriented information sources ▪ CredibleMeds ▪ VA NDF- RT ▪ ONC High Priority ▪ ONC Non-interruptive ▪ OSCAR b. 4 Natural Language Processing (NLP) Corpora ▪ DDI Corpus 2011 ▪ DDI Corpus 2013 ▪ PK DDI Corpus ▪ NLM CV DDI Corpus c. 5 Bioinformatics/Pharmacovigilance information sources ▪ KEGG DDI ▪ TWOSIDES ▪ DrugBank ▪ DIKB ▪ SemMedDB