Citation: CPT Pharmacometrics Syst. Pharmacol. (2018) 7, 90–102; doi:10.1002/psp4.12267

REVIEW

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*]

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 informatic 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 informatic approaches and facilitates the translational drug interaction research.

CPT Pharmacometrics Syst. Pharmacol. (2018) **7**, 90–102; doi:10.1002/psp4.12267; published online 9 January 2018.

Adverse drug events (ADEs), the unintended drug side effects, have led to major public health burden. In the United States alone, >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 evidence shows that the drug-drug interaction (DDI) is one of the leading causes of ADEs. With the increasing rate of polypharmacy, the risk of ADEs increased from 13% (2 drugs) to 58% (5 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 paroxetine in a clinical PK study.[4] *In vitro* metabolism studies revealed that this is due to paroxetine's strong inhibition of the tamoxifen biotransformation to endoxifen via the CYP2D6 pathway (5). In a follow-up pharmacogenetics study, patients with breast cancer 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 selective serotonin reuptake inhibitors 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 & 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] the FDA adverse event reporting system (FAERS) and electronic medical records (EMRs) were utilized to generate and validate drug-induced ADE and drug-drug-induced ADE associations. Duke *et al.*[10] proposed a text mining strategy for DDI molecular pharmacology evidence discovery from the public literature, which discovered 13,197 potential DDIs. In the follow-up *in vitro* study, Han *et al.*[11] validated the loratadine-simvastatin myotoxicity interaction, and its increased myopathy risk in both EMR and FAERS databases.

Driven by the emerging big data and novel computational models, there are three areas in which translational biomedical informatics and pharmacometrics are having a major impact on the drug interaction research. First, during the past 2 decades, federal regulatory agencies, hospitals, and research organizations maintained various patient databases, such as spontaneous reporting system (SRS), EMRs, and electronic health records (EHRs) for postmarketing 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, PKs 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 informatic approaches. For instance, the reviews by Harpaz *et al.*[12] and Koutkias & Jaulent[13] focused on

[1]Department of Biomedical Informatics, College of Medicine, the Ohio State University, Columbus, Ohio, 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, Indiana, 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, North Carolina, USA; [6]Department of Obstetrics and Gynecology, Indiana University, Indianapolis, Indiana, USA. *Correspondence: L Li (Lang.Li@osumc.edu)

[†]Pengyue Zhang and Heng-Yi Wu contributed equally to this work.

Received 7 November 2017; accepted 8 November 2017; published online on 9 January 2018. doi:10.1002/psp4.12267

computational models for SRS and EMR databases. Text and data mining techniques to detect ADE signals were reviewed by Karimi *et al.*[14] Jensen *et al.*[15] summarized available EMR/EHR databases and the obstacles for the EMR/EHR mining. However, these reviews did not focus 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 postmarketing surveillance data mining, PBPK modeling, and literature-based knowledge discovery, because these three approaches complement each other. The rest of this review is organized as follows: data mining methods for the postmarketing surveillance are shown; PBPK DDI models and databases are presented; literature-based DDI discovery approaches are presented; and the last section concludes this review.

DDI DATA MINING METHODS USING THE POSTMARKETING SURVEILLANCE DATA

A brief review of single drug ADE association analyses

Univariate disproportionality analyses. Disproportionality analyses (DPAs) are the pioneer approaches to quantify and prioritize single drug-ADE associations. For a drug-induced ADE pair, DPAs summarize data into a two-by-two 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-induced ADE pair is observed, and the expectation is the expected frequency of this drug-induced ADE pair under the assumption of no association. As its name implies, 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-induced ADE pairs of interest, or can conduct drug-wide and ADE-wide signal screening.

Proportional reporting ratios (PRRs) and reporting odds ratios (RORs) 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 ROR, are also often used for the signal detection. The 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 distribution assumptions. Later, Norén *et al.*[20] introduced a joint Dirichelet distribution prior and extended Bate's IC model. Like PRR and ROR, signal detection using IC can be based on its lower bound of the 95% confidence interval (IC_025). The 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 prespecified prior distribution. The 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 also applied to the PRR.[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 the local false discover rate 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 >1. Particularly, the second distribution characterizes the null hypothesis, whereas the third distribution characterizes the alternative hypothesis. Hence, the local false discover rate estimates the probability of the null distribution conditional on the data and the 3CMM.

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

Multiple logistic regression (MLR) and regulated logistic regression are two other approaches in analyzing drug-ADE associations. MLR is a traditional statistical approach to detect drug-induced 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-induced ADE signal detection by MLR may involve a large number of drugs than sample sizes, in which regulated logistic regression becomes a viable solution, such as ridge and Lasso regression models. An example of signal detections by lasso regression models includes Ola Caster *et al.*[26]

DPAs are less computationally expensive compared within other multivariate approaches.[12] Additionally, DPAs can be used either to analyze specific drug-induced ADE pairs of interest, or can conduct drug-wide and ADE-wide signal screening. Although the disproportionality measurements may suffer from confounding bias, evaluations by gold standard have shown DPAs to have decent performances (area under the concentration-time curve (AUC)).[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-induced ADE associations.

For logistic regression modeling, the number of predictors is usually less than 2,000, which is similar to the number of FDA approved drugs. For pharmacovigilance databases, the sample sizes are usually up to a few million. 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 50 GB 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.

Drug interaction signal detection

Some of the DPAs in this review 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.*[28] introduced an extended LRT method that can be used for detecting signals for multiple drugs (or ADEs) in a drug class (or in an ADE group). Likewise, an extended higher order IC method is proposed by Norén *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 cannot only be used for detecting the potential DDIs, but also can be used for detecting the association between a drug-induced ADE pair and another risk factor (e.g., age or gender). The examples for the extended EBGM can be found in Almenoff *et al.*[29] and DuMouchel & Pregibon.[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.

Norén *et al.*[31] proposed a novel model for detecting two-way DDIs. 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.*[33] proposed multiplicative and additive relationship to model the risks for single drugs and DDI pairs. 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 DDIs, 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 DDIs).

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, whereas the other pattern assumes that ADE risk increases like the dose (i.e., drug counts) response curve. This model then estimates a probability for each drug combination that 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 two-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 the tremendous amount of drug combinations, informatic approaches become a promising and practical solution. Two major informatic 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.*[35] proposed an FCI-filter approach that integrated FCI mining and uninformative association removal to mine multiple drug interactions from the FAERS. 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 *a priori* algorithm is utilized to mine the FAERS data. Their *a priori* 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.

PHARMACOKINETICS MODELING AND DATA SOURCES

In vitro in vivo drug interaction prediction using pharmacokinetics modeling

There are two ways to characterize PKs of drug. The top-down approach investigates clinical pharmacokinetic using clinical trial data, and it builds up a population PK model. The bottom-up approach, on the other hand, starts from PK 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 that can be scaled up (i.e., including potentially all drugs), and interfaced with informatic analyses.

The ratio of AUC (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 the AUCR based on unbound inhibitor concentration ([I]), inhibition rate constant for a drug (Ki), fraction of metabolism (fm) and fraction of renal clearance (fe) in Eq. 1.

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

All of these parameters can be obtained from various available data sources, except for fm. For example, the Metabolism and Transport Drug Interaction Database (DIDB) has a collection of drug Ki, and the book by Goodman and Gilman[41] has a collection of fe and drug maximum concentration (C_{max}), which can be used as [I]. There are several ways to estimate fm for a substrate. First, 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.*[42] 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 Eq. 2:

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

Second, pharmacogenetic studies can also be used to estimate fm through the fold-change in exposure of a substrate in extensive metabolizers compared to poor metabolizers.[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 Eq. 3:

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

Third, *in vitro* experiments also have been used to determine the contributions of several CYP pathways. Substrate depletion in the human liver microsomes 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 an 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 toward the total human liver microsome CLint via a scaling factor (RAF/ISEF) approach.[45] Recently, due to the success of the cryopreservation of human hepatocytes,[46] the 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 5 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 the absence and presence of ketoconazole as Eq. 4:

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

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

Adverse drug reactions databases and data sources

There are a number of drug-related databases that integrate bioinformatics, cheminformatics, and/or DDI knowledge, which 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] In addition, part of premarket drug development relies on the drug information and DDI knowledge to predict interactions between a new drug candidate and drugs currently on the market.

DDI-related database. The DrugBank[49] is a well-known comprehensive database that contains bioinformatics and chemo-informatic 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 of 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 databases were derived from data mining from health record data sources. For example, the OFF-SIDE database contains drug-event signals that are not

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 8,261 small molecule and biotech drugs including approved, withdrawn, 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 1,332 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 • Experimental conditions and results of DDI studies are all integrated	• No additive or synergic information 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	–

ADE, adverse drug event; AERS, adverse event reporting system; DDI, drug-drug interaction; DIDB, Drug Interaction Database; FDA, US Food and Drug Administration; PD, pharmacodynamic; PK, pharmacokinetic.

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, coprescriptions, and hidden covariates. Using the same method, DDI signals were further derived from FAERS data, which is called 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 reviews, the DIDB collects *in vitro* and *in vivo* data of PK 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.

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, and data types, and their integration with EHRs (**Table 2**,[54–64]). 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 (PheKB) website,[54] "an online environment supporting the workflow of building, sharing, and validating electronic phenotype algorithms," which offers algorithms using approaches, such as International Classification of Disease, 9th revision-Clinical Modification (ICD-9-CM) codes, medications, and NLP. The

Table 2 Sources for ADE-phenotyping

Source name	Level of evidence	Source description	Terminologies and datatypes	Integration into EHR
MedDRA[58]	Level III	A unified standard terminology for recording and reporting adverse drug events	From higher to lower levels: SOC, HLGT, HLT, PT, and LLT	• Used in structured data or unstructured clinical narratives
CPT[60]	Level III	A medical terminology to bill outpatient and office procedures	Category I, category II, and category III codes	• Used in structured data or unstructured clinical narratives
ICD[61]	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
LOINC[62]	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
SNOMED CT[63]	Level III	A multilingual clinical terminology to address the requirement for effective HER	Hierarchical representation of detailed clinical information (e.g., top level concepts, such as clinical finding, procedure, and substance)	• Used in structured data or unstructured clinical narratives
RxNorm[64]	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
CTCAE[59]	Level III	A comprehensive, multimodality grading system for reporting ADEs of cancer treatment	AEs terms associated with 5-point severity scale of ADE, and mapped to MedDRA LLTs to support standardization of ADEs terms in HER	• Used in structured data or unstructured clinical narratives • Severity scale of ADEs provides additional evidence
The SIDER database of drugs and side effects[57]	Level II	A computer-readable SE resource/database mined from FDA drug labels, contains about 1,430 drugs, 5,868 SEs, 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[56]	Level II	An evidence-based, physician-authored clinical decision support tool	Synthesized medical information, such as clinical guidelines, graded recommendation, and drug entries and interactions	• Used in structured data or unstructured clinical narratives • Evidence-based medical information and drug interactions assist in defining ADEs • Up-to-date clinical guidelines
OMOP[55]	Level II	Literature-based 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 laboratory values	• Used in structured data or unstructured clinical narratives • Broad and narrow definitions can be implemented directly into EHR based on users' needs
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

AE, adverse event; ADEs, adverse drug effects; CPT, Current Procedural Terminology; CTCAE, Common Terminology Criteria for Adverse Events; EHR, electronic health record; HLGT, high-level group term; HLT, high-level term; HOI, health outcome of interest; ICD, International Classification of Diseases; LLT, lowest level term; LOINC, Logical Observation Identifiers Names and Codes; MedDRA, Medical Dictionary for Regulatory Activities; OMOP, Observational Medical Outcomes Partnership; PheKB, Phenotype Knowledgebase website; PT, preferred term; SEs, side effects; SIDER, database of drugs and side effects; SNOMED CT, Systematized Nomenclature of Medicine; SOC, System Organ Class.

PheKB's main goals are to improve algorithm transportability and validity across institutions,[54] while achieving high positive predictive values.

Level II evidence provides ADE phenotypes that met criteria 3, and (2 and/or 4), but they have not been validated across institutions. For instance, the 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. The OMOP also recognized that

Table 3 Summary of integrated database

Database name	Data type in database	Data sources
DIKB	Mechanisms and PK DDIs information with confidence	● Retrospective studies ● clinical trials ● metabolic inhibition identification and 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 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

DDI, drug-drug interaction; DIKB, Drug Interaction Knowledge Base; KEGG, Kyoto Encyclopedia of Genes and Genomes; MeSH, medical subject heading; PDDI, potential drug-drug interaction; PK, pharmacokinetic.

literature has usually been inconsistent in defining and reporting ADEs, and sometimes lacking details of the exact codes and validations.[65] For example, acute liver injury has eight different definitions, such as laboratory-based versus diagnostic procedures.

In addition to the OMOP HOI library, UpToDate is another evidence-based physician-authored clinical guideline repository.[56] It provides 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, the Side Effect Resource (SIDER)[57] 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 ADEs,[58] and the other database include Common Terminology Criteria for Adverse Events (CTCAE).[59] Unlike MedDRA, the CTCAE also contains the severity of ADEs.

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 PK DDI 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 *et al.*[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 could be used to standardize the format and content of the observational databases, including medications, ADEs, symptoms, and indications. However, a complete solution for data integration is still unavailable. Second, the DDI information in the databases is limited. Particularly, information including the DDI type (e.g., additive/synergic, pharmacokinetic/pharmacodynamic (PK/PD)), mechanism, clinical impact and quantitative description should be included and improved in the future data collection.

KNOWLEDGE DISCOVERY FOR DRUG INTERACTION USING TEXT MINING TECHNOLOGIES

Literature-based knowledge discovery was pioneered by Don R. Swanson in 1986[69] and had been widespread for decades in the biomedical informatics domain. 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 medical subject heading (MeSH) terms in Medline.[71] Hristovski *et al.*[72] combined the outputs of two NLP systems to provide semantic prediction, which demonstrated the improvement for literature-based knowledge discovery. Tsuruoka & Tsujii[73] developed a search engine for Medline abstracts, called FACTA, which retrieves textual evidence of associations between the query terms and the concepts. Frijters *et al.*[74] developed the CoPub discovery tool to assess the significance of co-occurrence based on the mutual information measure and mine the new relationships between biomedical concepts. Finally, Yetisgen-Yildiz & Pratt[75] proposed an evaluation methodology allowing the comparisons across different systems.

Although 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, because not all co-occurring entities possess "meaningful" and "quality" relations. To retrieve explicit facts 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 DDIs. 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; and (3) linking the disjointed sets of facts from documents uncovers hidden links between drugs and generates novel hypotheses.

Drug interaction corpora

Great text mining methods rely on well-developed corpora. Corpora refers to manually annotated golden standard data. In the DDI text mining domain, DDI corpora developed in both DDI extraction challenge tasks in 2011 and 2013,[77,78] have guided a great number of supervised DDI text mining methodologies' development. The annotation strategies in corpus may differ subject to the purpose of text mining 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 makeup, part of speech tagging, and constituent or dependency of parsing trees[81]; and (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] The 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 the 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. The 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 is 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 PK PDDI text mining 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 ref. 83 was reused (i.e., object and precipitant for the role of interacting drugs or substances). In addition, the authors further categorized 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: (1) *in vivo* PK studies ($n = 56$); (2) *in vivo* pharmacogenetic studies ($n = 57$); (3) *in vivo* DDI studies ($n = 218$); and (4) *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, and PK parameters with their values and units, mechanisms, and 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 the 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 the Gonzalez laboratory was manually annotated for mining Twitter for adverse drug reactions (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 ADRs. In addition, another corpus, also created by the Gonzalez laboratory, 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]; and the PK corpus further differentiated PK DDI into *in vivo* and *in vitro* studies, and defines drug interactions using experimental evidence.[80] The corpora for social media were annotated differently from those in literature. Two corpora, created by the Gonzalez laboratory, 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 identifed the attributes for entities.[86]

Information retrieval and extraction for drug interaction and drug-related knowledge

In order to promote DDI text mining, DDI-Extraction challenges organized in 2011 and 2013 aimed for developing the text mining methodologies of the pharmacological substance recognition and DDI detection.[77,78] For the named entity recognition (NER) of pharmacological substances, the best results were achieved by WBI_NER. This NER approach is formulated as a sequence labeling 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 complexity in naming convention lead to the difficulty in name recognition. Another focus in the 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 an ensemble learning approach. Combined three different kernels (all-paths graph, shallow linguistic, and k-band shortest path spectrum kernels) with a case-based reasoning 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 first 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 nonlinear kernel-based methods outperformed linear supporting vector machine 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, a 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). The authors concluded that DDI representations in *in vitro* PK studies were more diverse than those in *in vivo* PK studies. 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.*[89] presented a graphic kernel-based approach to combine syntactic and semantic information for extracting PK drug interactions. 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 process, and clinical studies investigate the clinical effects (i.e., efficacy or side effects of DDIs). Recent work by Kolchinsky *et al.*[76] classified the *in vitro* and *in vivo* PK DDI evidences. More recently, Wu *et al.*[90] 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. 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. The 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 postmarketing 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 & Gonzalez[93] focused on the classification of sentences to detect ADR mentions utilizing features, including n-gram, Unified Medical Language System semantic types, Synset expansion, etc. 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 for 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. For information extraction, Carbonell *et al.*[94] analyzed the features using time series analysis, co-evaluated the mentions of drugs in Twitter within intervals of 30 minutes, and explored the potential drug effects and drug interactions. Another tool called ADRMine utilized a variety of features, including a new feature for modeling words' semantic similarities.[95] Using conditional random fields 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 ADR mentions.

Information discovery for novel drug interactions and ADEs

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 the 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 used. Both Tari *et al.*[97] and Percha *et al.*[98] are two typical examples of the close discovery method. Tari *et al.*[97] developed a method combining text mining and automated reasoning to infer DDIs with the support of enzyme and biological domain knowledge. 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 *et al.*[98] proposed a novel approach to predict novel DDIs by aggregating gene-drug interactions, which are extracted via the rule-based method. 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 the Tari *et al.*[97] and Percha *et al.*[98] methods, Duke *et al.*[10] proposed a literature discovery approach combined with analysis of EMRs and predicted 13,197 CYP-related DDIs. Based on literature data on *in vitro* drug metabolism and inhibitory potency, this translational approach finally identified five 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 *et al.*[99] proposed a literature-mining framework to enhance the prediction of DDIs and ADE types through integrating drug-gene interactions. Using the DDI features from DDI corpus, a supervised learning categorized ADEs into four types: adverse effect, effect at molecular level, effect related to PKs, 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.

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 postmarketing surveillance databases. Second, we present PK models for IVIVE 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 postmarketing 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 gaps. 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 will be used to stimulate novel and creative translational biomedical informatics methods for the drug interaction research.

Conflict of Interest. As an Associate Editor for *CPT: Pharmacometrics & Systems Pharmacology*, Lang Li was not involved in the review or decision process for this article. The authors declared no conflict of interest.

Source of Funding. No funding was received for this work.

1. US Food and Drug Administration (FDA). <https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070434.htm> (2015). Accessed 12 August 2017.
2. Lazarou, J., Pomeranz, B.H. & Corey, P.N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* **279**, 1200–1205 (1998).
3. Prybys, K.M., Melville, K.A. & Hanna, J.R. Polypharmacy in the elderly: clinical challenges in emergency practice. Part 1: overview, etiology, and drug interactions. *Emerg Med Rep.* **23**, 145–151 (2002).
4. Stearns, V. *et al.* Active tamoxifen metabolite plasma concentrations after coadministration of tamoxifen and the selective serotonin reuptake inhibitor paroxetine. *J. Natl. Cancer Inst.* **95**, 1758–1764 (2003).
5. Desta, Z., Ward, B.A., Soukhova, N.V. & Flockhart, D.A. Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system in vitro: prominent roles for CYP3A and CYP2D6. *J. Pharmacol. Exp. Ther.* **310**, 1062–1075 (2004).
6. Goetz, M.P. *et al.* Pharmacogenetics of tamoxifen biotransformation is associated with clinical outcomes of efficacy and hot flashes. *J. Clin. Oncol.* **23**, 9312–9318 (2005).
7. Henry, N.L., Stearns, V., Flockhart, D.A., Hayes, D.F. & Riba, M. Drug interactions and pharmacogenomics in the treatment of breast cancer and depression. *Am. J. Psychiatry* **165**, 1251–1255 (2008).
8. Hennessy, S. & Flockhart, D.A. The need for translational research on drug-drug interactions. *Clin. Pharmacol. Ther.* **91**, 771–773 (2012).
9. Tatonetti, N.P., Ye, P.P., Daneshjou, R. & Altman, R.B. Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* **4**, 125ra31 (2012).
10. Duke, J.D. *et al.* Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. *PLoS Comput. Biol.* **8**, e1002614 (2012).
11. Han, X. *et al.* Identification and mechanistic investigation of drug-drug interactions associated with myopathy: a translational approach. *Clin. Pharmacol. Ther.* **98**, 321–327 (2015).
12. Harpaz, R., DuMouchel, W., Shah, N.H., Madigan, D., Ryan, P. & Friedman, C. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin. Pharmacol. Ther.* **91**, 1010–1021 (2012).
13. Koutkias, V.G. & Jaulent, M.C. Computational approaches for pharmacovigilance signal detection: toward integrated and semantically-enriched frameworks. *Drug Saf.* **38**, 219–232 (2015).
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.* **47**, article 56 (2015).
15. Jensen, P.B., Jensen, L.J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405 (2012).
16. Evans, S.J., Waller, P.C. & Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol. Drug Saf.* **10**, 483–486 (2001).
17. van Puijenbroek, E.P., Bate, A., Leufkens, H.G., Lindquist, M., Orre, R. & Egberts, A.C. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol. Drug Saf.* **11**, 3–10 (2002).
18. Huang, L., Zalkikar, J. & Tiwari, R.C. A likelihood ratio test based method for signal detection with application to FDA's drug safety data. *J. Am. Stat. Assoc.* **106**, 1230–1241 (2011).
19. Bate, A. *et al.* A Bayesian neural network method for adverse drug reaction signal generation. *Eur. J. Clin. Pharmacol.* **54**, 315–321 (1998).
20. Norén, G.N., Bate, A., Orre, R. & Edwards, I.R. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Stat. Med.* **25**, 3740–3757 (2006).
21. DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am. Stat.* **53**, 177–190 (1999).
22. Ahmed, I. *et al.* Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. *Stat. Med.* **28**, 1774–1792 (2009).
23. Ahmed, I., Dalmasso, C., Haramburu, F., Thiessard, F., Broët, P. & Tubert-Bitter, P. False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics* **66**, 301–309 (2010).
24. Zhang, P. *Study Designs and Statistical Methods for Pharmacogenomics and Drug Interaction Studies*. Indiana University. <https://scholarworks.iupui.edu/handle/1805/11300> (2016).
25. Harpaz, R., Haerian, K., Chase, H.S. & Friedman, C. Mining electronic health records for adverse drug effects using regression based methods. 100–107. IHI '10 Proceedings of the 1st ACM International Health Informatics Symposium, Arlington, Virginia, 11–12 November 2010.
26. Ola Caster, G., Norén, G.N., Madigan, D. & Bate, A. Large-scale regression-based pattern discovery: the example of screening the WHO global drug safety database. *Stat. Anal. Data Min.* **3**, 197–208 (2010).
27. Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P. & Shah, N.H. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clin. Pharmacol. Ther.* **93**, 539–546 (2013).
28. Huang, L., Zalkikar, J. & Tiwari, R.C. Likelihood ratio test-based method for signal detection in drug classes using FDA's AERS database. *J. Biopharm. Stat.* **23**, 178–200 (2013).
29. Almenoff, J.S., DuMouchel, W., Kindman, L.A., Yang, X. & Fram, D. Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug

interactions in the post-marketing setting. *Pharmacoepidemiol. Drug Saf.* **12**, 517–521 (2003).

30. DuMouchel, W. & Pregibon, D. Empirical Bayes screening for multi-item associations. 67–76. KDD '01 Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, California, 26–29 August 2001.
31. Norén, G.N., Sundberg, R., Bate, A. & Edwards, I.R. A statistical methodology for drug-drug interaction surveillance. *Stat. Med.* **27**, 3057–3070 (2008).
32. Van Puijenbroek, E.P., Egberts, A.C., Meyboom, R.H. & Leufkens, H.G. Signalling possible drug-drug interactions in a spontaneous reporting system: delay of withdrawal bleeding during concomitant use of oral contraceptives and itraconazole. *Br. J. Clin. Pharmacol.* **47**, 689–693 (1999).
33. Thakrar, B.T., Grundschober, S.B. & Doessegger, L. Detecting signals of drug-drug interactions in a spontaneous reports database. *Br. J. Clin. Pharmacol.* **64**, 489–495 (2007).
34. Zhang, P. *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.* **4**, 474–480 (2015).
35. Xiang, Y. *et al.* Efficiently mining adverse event reporting system for multiple drug interactions. *AMIA Jt. Summits Transl. Sci. Proc.* **2014**, 120–125 (2014).
36. Harpaz, R., Chase, H.S. & Friedman, C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics* **11** Suppl 9, S7 (2010).
37. Jamei, M., Dickinson, G.L. & Rostami-Hodjegan, A. A framework for assessing interindividual 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. Pharmacokinet.* **24**, 53–75 (2009).
38. Yeo, K.R., Jamei, M. & Rostami-Hodjegan, A. Predicting drug-drug interactions: application of physiologically based pharmacokinetic models under a systems biology approach. *Expert Rev. Clin. Pharmacol.* **6**, 143–157 (2013).
39. Ito, K., Hallifax, D., Obach, R.S. & Houston, J.B. Impact of parallel pathways of drug elimination and multiple cytochrome P450 involvement on drug-drug interactions: CYP2D6 paradigm. *Drug Metab. Dispos.* **33**, 837–844 (2005).
40. Lu, C., Miwa, G.T., Prakash, S.R., Gan, L.S. & Balani, S.K. A novel model for the prediction of drug-drug interactions in humans based on in vitro cytochrome p450 phenotypic data. *Drug Metab. Dispos.* **35**, 79–85 (2007).
41. Brunton, L.L., Lazo, J.S. & Parker, K.L. *Goodman & Gilman's the Pharmacological Basis of Therapeutics, 11th Edition* (McGraw-Hill, New York, NY, 2006).
42. Yeung, C.K. *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.* **4**, 489–494 (2015).
43. McGourty, J.C., Silas, J.H., Lennard, M.S., Tucker, G.T. & Woods, H.F. Metoprolol metabolism and debrisoquine oxidation polymorphism–population and family studies. *Br. J. Clin. Pharmacol.* **20**, 555–566 (1985).
44. Li, Z.M., Guo, L.H. & Ren, X.M. Biotransformation of 8:2 fluorotelomer alcohol by recombinant human cytochrome P450s, human liver microsomes and human liver cytosol. *Environ. Sci. Process. Impacts* **18**, 538–546 (2016).
45. Bohnert, T. *et al.* Evaluation of a new molecular entity as a victim of metabolic drug-drug interactions-an industry perspective. *Drug Metab. Dispos.* **44**, 1399–1423 (2016).
46. Stéphenne, X., Najimi, M. & Sokal, E.M. Hepatocyte cryopreservation: is it time to change the strategy? *World J. Gastroenterol.* **16**, 1–14 (2010).
47. Mao, J., Mohutsky, M.A., Harrelson, J.P., Wrighton, S.A. & Hall, S.D. Prediction of CYP3A-mediated drug-drug interactions using human hepatocytes suspended in human plasma. *Drug Metab. Dispos.* **39**, 591–602 (2011).
48. Desbans, C. *et al.* Prediction of fraction metabolized via CYP3A in humans utilizing cryopreserved human hepatocytes from a set of 12 single donors. *Xenobiotica* **44**, 17–27 (2014).
49. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091–D1097 (2014).
50. Drugs.com. <www.drugs.com> (2000). Accessed 12 August 2017.
51. Kyoto Encyclopedia of Genes and Genomes (KEGG). <http://www.genome.jp/kegg/> (1995). Accessed 12 August 2017.
52. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).
53. Li, Q. *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.* **21**, 776–784 (2014).
54. Kirby, J.C. *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Inform. Assoc.* **23**, 1046–1052 (2016).
55. Fox, B.I., Hollingsworth, J.C., Gray, M.D., Hollingsworth, M.L., Gao, J. & Hansen, R.A. Developing an expert panel process to refine health outcome definitions in observational data. *J. Biomed. Inform.* **46**, 795–804 (2013).
56. UpToDate. <http://www.uptodate.com/contents/search> (2017). Accessed 12 August 2017.
57. Kuhn, M., Letunic, I., Jensen, L.J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075–D1079 (2016).
58. Brown, E.G., Wood, L. & Wood, S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf.* **20**, 109–117 (1999).
59. Chen, A.P. *et al.* Grading dermatologic adverse events of cancer treatments: the Common Terminology Criteria for Adverse Events version 4.0. *J. Am. Acad. Dermatol.* **67**, 1025–1039 (2012).
60. CPT - Current Procedural Terminology. <https://www.ama-assn.org/practice-management/cpt-current-procedural-terminology> (1995). Accessed 12 August 2017.
61. International Classification of Diseases (ICD) <http://www.who.int/classifications/icd/en/> (2017). Accessed 12 August 2017.
62. Wilson, P.S. & Scichilone, R.A. LOINC as a data standard: how LOINC can be used in electronic environments. *J. AHIMA* **82**, 44–47 (2011).
63. Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT). <http://www.ihtsdo.org/snomed-ct/> (2011). Accessed 12 August 2017.
64. RxNorm. <https://www.nlm.nih.gov/research/umls/rxnorm/> (2004). Accessed 12 August 2017.
65. Hansen, R.A., Gray, M.D., Fox, B.I., Hollingsworth, J.C., Gao, J. & Zeng, P. How well do various health outcome definitions identify appropriate cases in observational studies? *Drug Saf.* **36** Suppl 1, S27–S32 (2013).
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.* **42**, 979–989 (2009).
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.* **42**, 990–1003 (2009).
68. Ayvaz, S. *et al.* Toward a complete dataset of drug-drug interaction information from publicly available sources. *J. Biomed. Inform.* **55**, 206–217 (2015).
69. Bekhuis, T. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomed. Digit. Libr.* **3**, 2 (2006).
70. Weeber, M., Kors, J.A. & Mons, B. Online tools to support literature-based discovery in the life sciences. *Brief. Bioinform.* **6**, 277–286 (2005).
71. Srinivasan, P. Text mining: generating hypotheses from MEDLINE. *J. Assoc. Inf. Sci. Technol.* **55**, 396–413 (2004).
72. Hristovski, D., Friedman, C., Rindflesch, T.C. & Peterlin, B. Exploiting semantic relations for literature-based discovery. AMIA Annual Symposium proceedings AMIA Symposium. 349–353 (2006).
73. Tsuruoka, Y. & Tsujii, J. Improving the performance of dictionary-based approaches in protein name recognition. *J. Biomed. Inform.* **37**, 461–470 (2004).
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.* **6**, 9 (2010).
75. Yetisgen-Yildiz, M. & Pratt, W. A new evaluation methodology for literature-based discovery systems. *J. Biomed. Inform.* **42**, 633–643 (2009).
76. Kolchinsky, A., Louren?o, A., Wu, H.Y., Li, L. & Rocha, L.M. Extraction of pharmacokinetic evidence of drug-drug interactions from the literature. *PLoS One* **10**, e0122199 (2014).
77. Segura-Bedmar, I., Martínez, P. & Sánchez-Cisneros, D. *The 1st DDIExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts*. <http://ceur-ws.org/Vol-761/paper0.pdf> (2011).
78. Segura-Bedmar, I., Martinez, P. & Herrero-Zazo, M. Lessons learnt from the DDIExtraction-2013 shared task. *J. Biomed. Inform.* **51**, 152–164 (2014).
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.* **46**, 914–920 (2013).
80. Wu, H.Y. *et al.* An integrated pharmacokinetics ontology and corpus for text mining. *BMC Bioinformatics* **14**, 35 (2013).
81. Tateisi, Y., Yakushiji, A., Ohta, T., Tsujii, J.I., eds. *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, W.J., Rzhetsky, A. & Shatkay, H. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics* **7**, 356 (2006).
83. Boyce, R., Gardner, G. & Harkema, H., eds. *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. *A machine-learning approach for drug-drug interaction extraction from FDA structured product labels*. National Library of Medicine Training Conference, Pittsburgh, Pennsylvania.
85. Vilar, S., Friedman, C. & Hripcsak, G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. *Brief. Bioinform.* (2017).
86. Alvaro, N., Miyao, Y. & Collier, N. TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill.* **3**, e24 (2017).
87. Ginn, R. *et al. 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 Brief* **10**, 122–131 (2017).
89. Zhang, Y. *et al.* Leveraging syntactic and semantic graph kernels to extract pharmacokinetic drug drug interactions from biomedical literature. *BMC Syst. Biol.* **10** (suppl. 3), 67 (2016).
90. Wu, H.Y. *et al. Translational drug interaction evidence gap discovery using text mining.* 2017 Joint Summits, San Francisco, California, 27–30, March 2017.
91. Harpaz, R. *et al.* Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Saf.* **37**, 777–790 (2014).
92. Abbasi, A. *et al.* Social media analytics for smart health. *IEEE Intell. Syst.* **29**, 60–80 (2014).
93. Sarker, A. & Gonzalez, G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomed. Inform.* **53**, 196–207 (2015).
94. Carbonell, P., Mayer, M.A. & Bravo, A. Exploring brand-name drug mentions on Twitter for pharmacovigilance. *Stud. Health Technol. Inform.* **210**, 55–59 (2015).
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. *J. Am. Med. Inform. Assoc.* **22**, 671–681 (2015).
96. Wong, C.M., Ko, Y. & Chan, A. Clinically significant drug-drug interactions between oral anticancer agents and nonanticancer agents: profiling and comparison of two drug compendia. *Ann. Pharmacother.* **42**, 1737–1748 (2008).
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).* **26**, i547–i553 (2010).
98. Percha, B., Garten, Y. & Altman, R.B. Discovery and explanation of drug-drug interactions via text mining. *Pac. Symp. Biocomput.* 410–421 (2012).
99. Raja, K., Patrick, M., Elder, J.T. & Tsoi, L.C. Machine learning workflow to enhance predictions of adverse drug reactions (ADRs) through drug-gene interactions: application to drugs for cutaneous diseases. *Sci. Rep.* **7**, 3690 (2017).
100. Hamed, A.A., Wu, X., Erickson, R. & Fandy, T. Twitter K-H networks in action: advancing biomedical literature for drug search. *J. Biomed. Inform.* **56**, 157–168 (2015).