Briefings in Bioinformatics, 2018, 1–14

doi: 10.1093/bib/bbx169
Paper

Drug knowledge bases and their applications in biomedical informatics research

Yongjun Zhu, Olivier Elemento, Jyotishman Pathak and Fei Wang

Corresponding author: Fei Wang, Division of Health Informatics, Department of Healthcare Policy and Research at Weill Cornell Medicine at Cornell University, 425 East 61st Street, Suite 301, DV-308, New York, NY 10065, USA. E-mail: few2001@med.cornell.edu

Abstract

Recent advances in biomedical research have generated a large volume of drug-related data. To effectively handle this flood of data, many initiatives have been taken to help researchers make good use of them. As the results of these initiatives, many drug knowledge bases have been constructed. They range from simple ones with specific focuses to comprehensive ones that contain information on almost every aspect of a drug. These curated drug knowledge bases have made significant contributions to the development of efficient and effective health information technologies for better health-care service delivery. Understanding and comparing existing drug knowledge bases and how they are applied in various biomedical studies will help us recognize the state of the art and design better knowledge bases in the future. In addition, researchers can get insights on novel applications of the drug knowledge bases through a review of successful use cases. In this study, we provide a review of existing popular drug knowledge bases and their applications in drug-related studies. We discuss challenges in constructing and using drug knowledge bases as well as future research directions toward a better ecosystem of drug knowledge bases.

Key words: drug knowledge bases; biomedical text mining; drug repositioning; adverse drug reaction analysis; pharmacogenomic analysis

Introduction

In recent years, because of the rapid development of computer technologies, extensive drug-related data, such as drugs, diseases, genes and proteins, have been generated [1]. The availability of such data has greatly facilitated drug-related research, such as network medicine [2], pharmacogenomics [3] and personalized medicine [4]. The complexity and absence of major standard regarding nomenclature and experimental condition have given rise to a need for curation of these data. Drug knowledge base, as an organic way of massaging and curating those different aspects of drug data, has been a popular research topic in biomedical informatics recently. A variety of drug knowledge bases have been developed to provide curated drug-related data.

There are two primary goals on developing knowledge bases: systematic curation of existing knowledge and efficient discovery of new knowledge. These two aspects are intertwined because (1) new knowledge is typically discovered based on existing knowledge; (2) after the new knowledge has been validated, it will be curated and inserted to the existing knowledge bases again. With these two aspects alternating each other, the drug knowledge bases will keep on growing and becoming more and more comprehensive.

Yongjun Zhu, PhD, is a Postdoctoral Associate in the Division of Health Informatics, Department of Healthcare Policy and Research at Weill Cornell Medicine at Cornell University, New York, NY.
Olivier Elemento, PhD, is an Associate Professor at Department of Physiology and Biophysics, Weill Cornell Medicine. He is also the acting director of Caryl and Israel Englander Institute for Precision Medicine, and associate director of HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine at Weill Cornell Medicine at Cornell University, New York, NY.
Jyotishman Pathak, PhD, is the Frances and John L. Loeb Professor of Medical Informatics and Chief of Division of Health Informatics, Department of Healthcare Policy and Research, at Weill Cornell Medicine at Cornell University, New York, NY.
Fei Wang, PhD, is an Assistant Professor in the Division of Health Informatics, Department of Healthcare Policy and Research at Weill Cornell Medicine at Cornell University, New York, NY.
Submitted: 8 August 2017; **Received (in revised form):** 15 November 2017

To achieve the two goals, it is important to understand and compare existing popular drug knowledge bases, as they not only provided us the curated information but also help us recognize the pros and cons of the state of the art and identify novel opportunities. Moreover, it would also be helpful to investigate how they were applied in different applications, which can help us gain better insights on the demands from the users.

Despite the existing drug knowledge bases and various applications using them, to the best of our knowledge, there is no detailed review comparing the existing drug knowledge bases and discussing any insights derived from those applications. A review about high-throughput methods for combinatorial drug discovery by Sun *et al.* [5] introduced publicly available drug-related data sources, but it was specifically focusing on combinatorial drug design. In this review, we aim to fill in this gap by providing an overview of existing, widely used drug knowledge bases with various foci and how they were applied in biomedical research. In addition, we discuss challenges in constructing and using drug knowledge bases as well as future research directions toward a better drug knowledge ecosystem.

Existing drug knowledge bases

In this section, we survey existing popular drug knowledge bases that are publicly available. There are many biomedical knowledge bases that contain drug information. In this article, we will just focus on the ones that are dedicated to approved, clinical trial and experimental drugs. For example, BindingDB [6], which is a database of interactions of proteins and ligands, is not included in the survey because it is not specifically about drugs, but ligands. ChEMBL [7] is also excluded because it contains information on thousands of compounds, few of which are approved drugs used for patients. In addition, drug knowledge bases that are not publicly available because of various reasons (e.g. being not open to public or service being suspended) are not included in this survey. Drug terminology databases such as RxNorm [8] and Veterans Health Administration National Drug File-Reference Terminology (NDF-RT) [9] are excluded as well.

Table 1 provides an overview of the surveyed drug knowledge bases in this article. We complied the list based on our own domain knowledge and online search. Specifically, we first come up with an independent list of popular drug knowledge bases by each of us. Then, we search by Google with search terms 'Drug Knowledge Base' and 'Drug Database' and compile another list. Finally, we take an intersection of these five lists to obtain the ultimate list. It presents a nonexhaustive list of popular drug knowledge bases that have been frequently mentioned in research studies. In the following, we will briefly review them.

Pharmacogenomics Knowledge Base (PharmGKB) [10–12] contains gene–drug relations. Released in 2000, it is one of the first knowledge bases about drugs. The primary source of the information is literature, and information from other gene and drug knowledge bases such as dbSNP [42] and DrugBank [20–23] is also included. Information on diseases, genetic variants, pathways and drug dosing guidelines is provided. In addition to drug–gene relations, drug–drug relations and drug–disease relations are also available. PharmGKB does not have its own ID system for drugs, but drugs are linked to entries in many other drug knowledge bases such as Therapeutic Target Database (TTD) [13–17], DrugBank [25–28], DailyMed [18] and KEGG DRUG [19–24].

TTD [13–17] includes information of drugs and their therapeutic targets. Released in 2002, in addition to drug–target relations, it provides five types of drug combination effects (in total 118 interactions), such as synergistic, additive, antagonistic, potentiative and reductive. The database was created by consulting textbooks, journal articles, catalogs of FDA approved drugs, reports from pharmaceutical companies and US patent databases. Standardized IDs were assigned to targets and drug entries, and these IDs are associated with IDs in other drug databases, such as DrugBank [25–28] and SuperDrug [43]. The latest version (10 September 2015) contains 31 614 drugs including approved, clinical trial and experimental drugs as well as 2589 targets.

DailyMed [18], created in 2005, provides information about marketed drugs in United Sates. It is maintained by the National Library of Medicine (NLM) and provides FDA label information. It contains 96 955 drug listings. Data entries in DailyMed are connected with RxNorm entries. Indication information is available through drug labels.

KEGG DRUG [19–24] is a component of KEGG (Kyoto Encyclopedia of Genes and Genomes), which is a comprehensive knowledge base of four categories, including systems information (e.g. KEGG PATHWAY), genomic information (e.g. KEGG GENES), chemical information (e.g. KEGG COMPOUND) and health information (e.g. KEGG DRUG). KEGG DRUG is a database for approved drugs in Japan, the United States and Europe. Information on drugs is extracted from drug labels (package inserts). Drug–drug interaction information is extracted from drug labels of all prescription drugs in Japan. Therefore, the database only includes information on approved drugs. The current version (22 December 2016) contains information on over 4000 drugs and 200 000 interactions. One advantage of using KEGG DRUG is that it is connected to other KEGG components such as KEGG pathway to provide integrated information. Drugs in the database are identified with the database's own ID system (called D number) and are linked to other knowledge bases, such as DrugBank [25–28] and DailyMed [18].

DrugBank [25–28], released in 2006, is one of the most comprehensive databases for drugs. The main information sources for DrugBank are textbooks and journal articles. It provides information on two types of drugs (i.e. small molecule and biotech) categorized into six groups (i.e. approved, vet approved, nutraceutical, illicit, withdrawn, investigational and experimental). DrugBank also provides information on targets, indications and pathways. The latest version, which is version 5.0, contains 8261 drug entries and 4388 nonredundant protein (i.e. drug target/enzyme/transporter/carrier) sequences that are linked to these drug entries. In addition, the database contains 275 directly studied drug effects [adverse drug reactions (ADRs) and general effects) extracted from literature and various online resources as well as 5789 effects inferred based on drug metabolism and known enzyme polymorphisms. While the first version of the database was created manually, the curations of the subsequent versions were assisted by a set of automated tools. All data acquired from automated processes were manually inspected and verified. DrugBank is updated on a daily basis, while the downloads are released quarterly.

SuperTarget [29, 30] is a database of drug–target relations that contains information on drugs, targets and side effects. Released in 2008, it was created by leveraging many other knowledge bases and the literature indexed by PubMed. Started from the drug information collected in previous work—SuperDrug [43], sentences containing potential drug–target relations from article abstracts in PubMed were extracted. Manual

Table 1. Drug knowledge bases surveyed in this study

KB	Owner	Link	Release	Source	Entities	Relations	Drug ID	Cross-reference
PharmGKB [10–12]	Stanford University	https://www.pharmgkb.org/	2000	Literature, other KBs	Gene, Drug	Drug–gene, drug–drug, drug–disease	PharmGKB Accession Id	TTD, DrugBank, KEGG Drug, DailyMed
TTD [13–17]	National University of Singapore	http://bidd.nus.edu.sg/group/cjttd/	2002	Literature, public documents, US patent databases	Target, Drug	Drug–target, drug–disease, drug–drug	TTD Drug ID	DrugBank, SuperDrug
DailyMed [18]	NLM	https://dailymed.nlm.nih.gov/dailymed/	2005	FDA drug labels	Drug	Drug–disease	NDC	RxNorm
KEGG DRUG [19–24]	Kyoto University	http://www.genome.jp/kegg/drug/	2005	Drug labels	Drug	Drug–target, drug–drug	D number	DrugBank
DrugBank [20–23]	University of Alberta	https://www.drugbank.ca/	2006	Literature, other KBs	Drug, Target	Drug–drug, drug–target, drug–disease	DrugBank ID	TTD, PharmGKB, KEGG DRUG
SuperTarget [29, 30]	Charité – Universitätsmedizin Berlin	http://insilico.charite.de/supertarget/	2008	Literature, other KBs	Target, Drug	Drug–target, drug–side effect	PubChem ID	NA
DIKB [31, 32]	University of Pittsburgh	https://dbmi-icode-01.dbmi.pitt.edu/dikb-evidence/front-page.html	2009	Literature, drug labels	Drug	Drug–drug	NA	DrugBank
SIDER [33, 34]	European Molecular Biology Laboratory	http://sideeffects.embl.de/	2010	Public documents, Package inserts	Side effect, Drug	Drug–side effect, drug–disease	PubChem ID	NA
DGIdb [35, 36]	Washington University in St. Louis	http://dgidb.genome.wustl.edu/	2013	Literature, other KBs	Gene, Drug	Drug–gene	PubChem ID	NA
DrOn [37, 38]	University of Arkansas for Medical Sciences	https://ontology.atlassian.net/wiki/spaces/DRON/overview	2013	RxNorm, National Drug Code	Drug	NA	NA	NA
DINTO [39]	Universidad Carlos III de Madrid	http://labda.inf.uc3m.es/doku.php?id=es:labda_dinto	2013	Literature, other KBs	Drug	Drug–drug	NA	NA
Merged-PDDI [40]	University of Pittsburgh	https://www.dikb.org/Merged-PDDI/	2015	Literature, other KBs	Drug	Drug–drug	NA	DrugBank, DIKB
DID [41]	Merck Research Laboratories	https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-016-0110-0	2017	Other KBs	Drug	Drug–target	NA	NA

Note: Columns represent, in consecutive order, names, owners, links, release years, information sources, entities, entity relations, ID systems and cross-references of the drug knowledge bases.

curation was performed to guarantee the validity of the information. Drug–target relations were also collected from drug knowledges, such as DrugBank [20–23] and TTD [13–17]. These relations were manually confirmed by reviewing relevant literature. Relations that cannot be confirmed by literature were annotated with sources where these retaliations had been extracted. ADRs (side effects) were extracted from Canadian Adverse Reaction Monitoring Program [44]. Drugs in SuperTarget are identified using IDs from PubChem [45, 46], which is a database of chemical molecules. SuperTarget does not provide crosslinks to other drug knowledge bases. The latest version (2011) contains information on 6219 targets, 195 770 drugs and 332 828 drug–target relations.

Drug Interaction Knowledge Base (DIKB) [31, 32] is an ontology about drug-mechanism evidence. The goal of this ontology is to associate assertions about drugs' mechanistic properties with supporting and refuting evidences. Over 30 evidence types from seven groups (i.e. retrospective studies, clinical trials, metabolic inhibition identification, metabolic catalysis identification, statements, reviews and observational reports) are used to support or refute assertions. Journal articles, drug labels and authoritative statements were used to collect evidences. Based on the provided evidences, individuals can make their own judgments. The ontology is downloadable at NCBO BioPortal [47], and the current version (May 2015) contains 360 classes and 140 properties.

Side Effect Resource (SIDER) [33, 34] is a database for drugs and their reported ADRs (side effects). The information was obtained from public documents and package inserts. Information on indications of drugs is also contained in the database. It was released in 2010, and the latest version (21 October 2015) contains 1430 drugs, 5868 side effects and 139 756 pairs of drug and side effects. About 40% of the pairs have frequency information. For each drug–side effect pair, label sources are available for review.

Drug–Gene Interaction database (DGIdb) [35, 36], released in 2013, provides two types of information: known drug–gene interactions and druggable genes that have not been targeted therapeutically. Information was obtained from literature and over 20 publically available sources such as PharmGKB [10–12], TTD [13–17], DrugBank [20–23], PubChem [45, 46], Gene Ontology [48, 49] and many other databases. Along with each drug–gene interaction, information on the number of sources and PubMed references that support the interaction is provided. DGIdb includes 39 categories of druggable genes and 35 interaction types (inhibitors, activators, cofactors, etc.). The latest version of DGIdb contains over 40 000 genes, 10 000 drugs and 15 000 drug–gene interactions.

Drug Ontology (DrOn) [37, 38] is an ontology of drugs. RxNorm [8] is the primary source of the ontology. DrOn provides a historically comprehensive list of National Drug Codes (NDC) [50]. RxNorm was used because its historical versions contain rich information on historical NDCs by associating its Concept Unique Identifier (RxCUI) with NDC. Drug ingredients in RxNorm are mapped to Chemical Entities of Biological Interest (ChEBI) ontology [51]. DrOn does not contain relational information between drugs and other entities such as targets. The latest version (September 2016) contains 434 663 drugs and 20 properties.

Drug Interaction Ontology (DINTO) [39] is an ontology for drug–drug interactions. The goal of DINTO is to integrate existing resources on drug–drug interactions and provide a comprehensive ontology of different types of drug–drug interactions. DINTO systematically integrates resources from the DDI corpora [52], ChEBI [51], DrugBank [20–23], Ontology of Adverse Events (OAE) [53] and SIDER [33, 34]. Drug–drug interactions are classified based on their clinical relevance, type of consequence or effect and preceding mechanisms. The latest version (August 2015) has 28 178 classes and 90 properties.

Merged-PDDI [40] is a comprehensive database of ∼100 000 drug–drug interactions created by integrating 14 publicly available sources. They obtained information on drug–drug interactions from journal articles, trustful websites, DDI corpora [52] developed for Natural Language Processing (NLP) challenges and other drug-related databases, such as DrugBank [20–23], KEGG [19–24] and DIKB [31, 32]. A simple PDDI data model was created to combine data entries from various sources.

Drug-Indication Database (DID) [41] is a resource for drug indications. Information was collected from 12 publicly or commercially available sources. DID contains 29 964 drugs, 10 938 targets and 192 008 drug-indication pairs. Information is available as triples of drug, indication and indication subtype. Indication subtype hierarchy is available so that users can select the level of granularity for their specific use cases. DDI nonproprietary subset is available for download.

From the above descriptions, we see that the main information source for constructing these knowledge bases is scientific literature (in total eight knowledge bases in Table 1 have used information from literature). Extracting useful information from literature to construct those knowledge bases is a time- and labor-intensive task. Although lots of intelligent NLP techniques have been developed and leveraged, manual efforts are still heavily involved to control the quality of the information extraction process. After the first several knowledge bases have been constructed, later on newer knowledge bases were also built on existing knowledge bases in addition to literature. Figure 1 summarizes the linkages between those existing drug knowledge bases.

Applications of drug knowledge bases

As we stated in the introduction, drug knowledge bases and the applications using them are two important aspects. In this section, we will review the biomedical application that uses existing drug knowledge bases. Using the names of the drug knowledge bases in Table 1 as search terms on PubMed, we have retrieved a set of studies that mention those names in titles or abstracts. We then manually reviewed abstracts one by one to filter out the studies that mention drug knowledge bases but did not use them. Based on our review of these application studies, we categorized them into five categories based on their tasks: biomedical text mining, drug repositioning, ADR analysis, pharmacogenomics analysis and others, which are presented in Table 2. Drug knowledge bases were mainly used for methodology developments and validations in those tasks. In the following, we will review those applications and how drug knowledge bases were used in more detail.

Biomedical text mining

Biomedical text mining is a process of extracting knowledge from biomedical textual data such as clinical notes and scientific literature. Information extraction and document summarization are the two major tasks here with a few advanced tasks such as question answering [97]. The information extraction step mainly involves two tasks: named entity recognition (NER) and relation extraction. Specifically, for this survey, we focus on

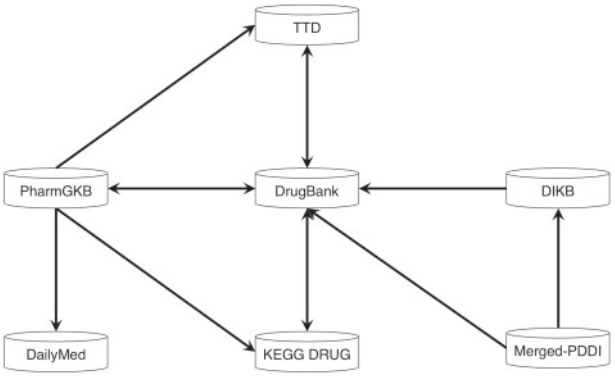

Figure 1. Cross-references among existing drug knowledge bases. Given a pair of drug knowledge bases, a directed line segment linking them indicates one of them provides links to the other.

the techniques recognizing drug names and/or extracting relations between drugs as well as between drugs and other biomedical entities such as diseases, targets and genes from biomedical literature.

Drug NER

He *et al.* [54] proposed a machine learning algorithm to identify drug names from biomedical texts. Drug names in DrugBank [20–23] were used to extract drug names from PubMed abstracts by applying a context pattern induction method [98]. The extracted drug names were used as a dictionary, and a method called feature coupling generalization [99] was used to fitter the dictionary. As the final step, conditional random fields [58] were used with the dictionary to identify drug names. Korkontzelos *et al.* [100] tackled drug NER by combining results of multiple named entity recognizers through a voting system. Drug names extracted from DrugBank were used as the dictionary to annotate text to construct a training set.

Drug relation extraction

Theobald *et al.* [56] used relational information on drugs, diseases and genes in PharmGKB [10–12] to explore conditional dependencies among these entities from PubMed based on their co-occurrence. Derived conditional probabilities can lead to new hypotheses, inferences and personalized medicine. Clematide *et al.* [57] proposed a method of ranking candidate relations between diseases, drugs and genes extracted from PubMed abstracts. PubMed abstracts that had been used when constructing knowledge bases, such as PharmGKB and CTD [101], were selected to test the proposed method. A simple frequency-based text mining approach was used to develop the ranking method. Percha *et al.* [58] proposed a method that combines supervised and unsupervised machine learning to extract biomedical relationships from unstructured text. Drugs and gene lexicons were obtained from PharmGKB, and their dependency paths were extracted from PubMed abstracts. Extracted dependency paths connecting drug–gene pairs were used as features for clustering. Known relationships from PharmGKB were used in a supervised manner to extract unknown relationships.

Drug repositioning

Drug repositioning aims to reposition existing drugs for new indications to save drug development cost and increase productivity [102]. Drug knowledge bases are a valuable source to perform drug–disease and drug–target interaction analyses to achieve the goal. The reviewed studies tried to predict novel drug–target and drug–disease interactions by leveraging rich information in the drug knowledge bases.

Drug–target interaction prediction

Li *et al.* [59] retrieved drugs, targets and their interactions from DrugBank [20–23] to perform prediction of drug–target interactions using large-scale molecular docking. Known drug–target interactions from DrugBank were tested against their docking method to find out a subset of protein targets whose interactions with drugs are predictable using the method. Predicted interactions between the subset of targets and drugs in DrugBank were presented. In [60], a method of identifying new drug indications by combining literature mining and knowledge bases was proposed. Drug–target interactions were extracted from DrugBank, while gene–disease and protein–protein relations were extracted from the literature. Combined information was used to perform reasoning to identify new drug indications. The reasoning was based on a set of predefined rules on drug mechanism. Cobanoglu *et al.* [61] proposed a probabilistic matrix factorization method to predict drug–target interactions. Known drug–target interactions were extracted from DrugBank and used to learn hidden structures and predict new drug–target interactions. Yamanishi *et al.* [62] proposed DINIES, a system that predicts drug–target interactions based on known drug–target interactions obtained from drug knowledge bases such as DrugBank and TTD [13–17]. Two prediction approaches (i.e. chemogenomic and pharmacogenomics approaches) were proposed. The former approach used chemical structure and protein sequence, while the latter approach used side effect and protein sequence. Tao *et al.* [63] combined ontology-based inference and network analysis to predict drug targets. Drugs, targets and other entities relate to colorectal cancer (CRC) were extracted from DrugBank and PharmGKB [10–12] to construct an ontology. CRC disease genes were collected and a set of inference rules was defined to identify CRC potential drug target genes. Inferred genes were ranked based on their relationships with CRC disease genes in a protein–protein interaction network [103]. Zhang *et al.* [64] combined information obtained from previous genome-wide association studies, proteomics and metabolomics studies with drug–target information obtained from TTD to systematically narrow down the list of druggable proteins. Drugs of the identified protein targets were compared with existing diabetic drugs by considering gene expression patterns of cells treated by the two set of drugs. A subset of drugs was identified as candidate diabetic drugs. Seal *et al.* [65] explored the effectiveness of a network-based approach (i.e. random walk with restart [104]) on the prediction of drug–target interaction. The approach was applied to a heterogeneous network comprising drug–drug, drug–target and target–target networks complied from DrugBank. Yuan *et al.* [66] proposed a method that combines two machine learning approaches: similarity- and feature-based methods to predict drug–target interaction. Drug–target interactions were obtained from DrugBank. Outputs of six similarity-based methods were used as features of the learning to rank method to rank targets (or drugs) given a drug (or a target).

Drug–disease interaction prediction

Yang and Agarwal [67] used known side effects of drugs as features to build a prediction model that predicts indications of

Table 2. Application studies surveyed in this review

Task	Subtask	Studies	KBs used	Summary
Biomedical text mining	Drug NER	He *et al.* [54], Korkontzelos *et al.* [55]	DrugBank [20–23]	Names extracted from DrugBank were used as a dictionary for drug name extraction or a training set of machine learning algorithms
	Drug relation extraction	Theobald *et al.* [56], Clematide *et al.* [57], Percha *et al.* [58]	PharmGKB [10–12]	Curated relational information on drugs, diseases and genes in PharmGKB was used to generate features for machine leaning algorithms and evaluate automated relation extraction methods
Drug repositioning	Drug–target interaction prediction	Li *et al.* [59], Tari *et al.* [60], Cobanoglu *et al.* [61], Yamanishi *et al.* [62], Tao *et al.* [63], Zhang *et al.* [64], Seal *et al.* [65], Yuan *et al.* [66]	PharmGKB [10–12], DrugBank [20–23], TTD [13–17]	Known drug–target interactions extracted from PharmGKB, DrugBank and TTD were combined with domain knowledge to predict novel interactions using methods, such as molecular docking, ontology-based reasoning, network-based approach and machine learning
	Drug–disease interaction prediction	Yang & Agarwal [67], Bisgin *et al.* [68]	PharmGKB [10–12], SIDER [33, 34]	New drug–disease interactions were predicted by combining drug–disease and drug–side effect relations extracted from PharmGKB and SIDER
ADR analysis	Drug side effect exploration	Wang *et al.* [69], Bresso *et al.* [70]	DrugBank [20–23], SIDER [33, 34]	Known drugs and their side effects obtained from DrugBank and SIDER were used to understand existing side effects by exploring their relations with drug targets and genes as well as clustering them
	Drug side effect prediction	Pauwels *et al.* [71], Jahid and Ruan [72], LaBute *et al.* [73], Eshleman and Singh [74], Jamal *et al.* [75]	DrugBank [20–23], SIDER [33, 34]	Information obtained from DrugBank and SIDER was integrated with other information from UniProt [76], PDB [77], PubChem [45, 46] and twitter to predict new drug side effects using various drug properties, such as chemical structures and target information
	Drug–drug interaction prediction	Vilar *et al.* [78], He *et al.* [79], Cheng *et al.* [80], Hameed *et al.* [81]	DrugBank [20–23], TTD [13–17], SIDER [33, 34]	Drug-related information from DrugBank, TTD and SIDER was used to derive various similarity measures (e.g. structural, therapeutic, and genomic similarity) to predict new drug–drug interactions
Pharmacogenomic analysis		Rance *et al.* [82], Rasmussen and Dahmcke [83], Pakhomov *et al.* [84]	PharmGKB [10–12], DrugBank [20–23], TTD [13–17]	Known information about drugs and genes in PharmGKB, DrugBank and TTD was combined with other knowledge bases such as DGV [85] to predict drug–gene interactions or used as reference standards to evaluate results extracted from literature
Others	Drug classification	Re and Valentini [86], Lötsch and Ultsch [87]	DrugBank [20–23]	Information from DrugBank was combined with information from other databases (STITCH [88] and Gene Ontology [48, 49]) to predict therapeutic categories of drugs
	Drug clustering	Udrescu *et al.* [89], Papanikolaou *et al.* [90]	DrugBank [20–23]	Drug properties in DrugBank were used to derive similarity measures or build a network to cluster drugs
	Drug–target network analysis	Sun *et al.* [91], Sun *et al.* [92]	PharmGKB [10–12], DrugBank [20–23]	Information from PharmGKB and DrugBank was used to construct drug–target, drug–gene and drug–drug interaction networks to explore network features of entities and visualize them
	Potential addictive drug prediction	Sun *et al.* [93]	DrugBank [20–23]	Addictive drugs as well as nonaddictive drugs that share targets with the addictive drugs were obtained from DrugBank to explore useful information for the prediction of potential addictive drugs
	Beneficial drug combination prediction	Iwata *et al.* [94]	DrugBank [20–23], TTD [13–17], KEGG DRUG [19–24]	Information about drug–target interactions and known beneficial drug combinations obtained from DrugBank, TTD and KEGG DRUG was combined with ATC Classification System [95] to predict novel beneficial drug combinations
	Compound–target interaction prediction	Keum *et al.* [96]	DrugBank [20–23]	Information about compounds and targets obtained from DrugBank was used to calculate chemical similarities among compounds and genomic similarities among targets to predict their new interactions

Note: The first column represents five major drug-related biomedical tasks. The second column lists subtasks of the five major tasks. The third column lists studies that belong to each subtask. The fourth column presents knowledge bases used in each subtask. The fifth column provides brief summaries of the subtasks.

drugs. The basic hypothesis is that if two drugs are associated with the same side effect, then the drugs may share indications. Disease–side effect associations were constructed by extracting drug–side effect relations from SIDER [33, 34] and drug–disease relations from PharmGKB [10–12]. Bisgin *et al.* [68] applied latent Dirichlet allocation (LDA) [105] to phenome information to identify new indications of drugs. Side effects and indications extracted from SIDER [33, 34] were associated with drugs to construct a phenome matrix, which was later used to identify probabilistic associations between drugs and phonotypes using LDA.

ADR analysis

An ADR is an undesirable effect that caused by a drug beyond its anticipated therapeutic effects [106]. Identifying potential ADRs is a complicated process that is time-consuming and expensive. Computational methods that leverage existing drug knowledge bases have been proposed to perform ADR analysis. The reviewed studies performed three types of task: drug side effect exploration, i.e. exploring known drug side effects to get insights; drug side effect prediction, i.e. predicting unknown drug side effects; and drug–drug interaction prediction, i.e. predicting drug pairs that potentially cause ADRs by interacting with each other.

Drug side effect exploration

Wang *et al.* [69] explored the relationships between drug design and drug side effects through the analysis of human signaling network. Information on drugs and drug side effects was obtained from DrugBank [20–23] and SIDER [33, 34]. Network distances between drug targets and disease genes were used to explore how the network distances are associated with drug side effects. Bresso *et al.* [70] explored groups of drug side effects obtained by clustering individual side effects obtained from SIDER based on semantic similarity among terms describing the side effects. Drugs and targets were extracted from DrugBank to explore their properties that lead to a given group of side effects.

Drug side effect prediction

Pauwels *et al.* [71] proposed a method of predicting potential side effects of drugs based on their shared chemical structures that are likely to have side effects. The method was evaluated by predicting known side effects of drugs in SIDER [33, 34]. The method was then applied to drugs in DrugBank [20–23] to predict unknown side effects. Jahid and Ruan [72] used an ensemble approach to predict drug side effects based on information extracted from SIDER. The approach combined multiple machine learning classifiers based on the assumption that drugs with similar chemical structures may have similar side effects where each classifier was developed from drugs with similar chemical structures. The approach was applied to drugs in DrugBank to predict unknown side effects of drugs. LaBute *et al.* [73] combined molecular docking and a machine learning method to predict ADRs. Drug targets extracted from DrugBank were combined with protein information from UniProt [76] and Protein Data Bank (PDB) [77] to compute docking scores. The results of molecular docking were combined with the information of drug side effects extracted from SIDER. Logistic regression was applied to the combined information to predict ADRs. Eshleman and Singh [74] used a supervised machine learning approach to predict drug side effects using two sources: SIDER and twitter. MetaMap biomedical annotator [107] was used to extract drugs and side effects from twitter. Extracted drugs and side effects were connected if they had appeared in a user's history within a fixed window of time. Random forest [108] was applied to the sources to classify edges between drugs and side effects as either adverse or nonadverse. Jamal *et al.* [75] used machine learning algorithms to predict neurological ADR-based biological (i.e. targets, transporters and enzymes), chemical (i.e. substructure fingerprints) and phenotypic (i.e. side effects and therapeutic indications) properties extracted from DrugBank, PubChem [45, 46] and SIDER, respectively. A feature selection algorithm (i.e. Relief [109]) and Support Vector Machine (SVM) [110] were applied to predict ADRs.

Drug–drug interaction prediction

Vilar *et al.* [78] predicted drug–drug interactions based on molecular structural similarity between drugs involved in known drug–drug interactions and other drugs. The assumption was that if two drugs interact to produce a specific biological effect, drugs that are structurally similar to one of the two drugs are likely to produce the same effect by interacting with the other drug. Known drug–drug interactions were collected from DrugBank [20–23], and structural similarities of all drug pairs from DrugBank were computed to predict new interactions. He *et al.* [79] used a machine learning approach that combines different types of features to predict drug–drug interactions from biomedical literature. Three fields (i.e. indication, pharmacology and description) of drug entries in DrugBank were used to calculate similarities between drugs, and the information was later used as one feature of the proposed approach. Cheng *et al.* [80] used four drug–drug similarity measures as features of machine learning algorithms to predict drug–drug interactions. The four similarity measures are phenotypic similarity, therapeutic similarity, chemical structural similarity and genomic similarity. DrugBank [20–23], TTD [13–17] and SIDER [33, 34] were used as partial sources for deriving similarity measures. DrugBank was also used for evaluation. Hameed *et al.* [81] proposed a method to predict drug–drug interactions that can be used when negative samples for training are insufficient. The proposed method applied Growing Self Organizing Map (GSOM) [111] to infer negatives from unlabeled data set and used SVM [110] to infer drug–drug interactions. Drug–drug interactions from DrugBank were used as positive examples.

Pharmacogenomic analysis

Pharmacogenomics is a core area of precision medicine, which is a concept that considers individual variability in disease prevention and treatment [112, 113]. Genes are an important component of pharmacogenomic analysis and previous studies used drug knowledge bases to analyze interactions between drugs and genes/genetic variants. Rance *et al.* [82] proposed a mutation-centric approach to identify relations between drugs and genetic variants in PubMed abstracts. MetaMap biomedical annotator [107] and RxNorm [8] were used to extract and filter drug names. Co-occurrence between drug mentions and genetic variants was used to associate them. The results were evaluated against known relations between drugs and genetic variants in PharmGKB [10–12]. In [83], structural variants of drug target-encoding genes and their relations with drugs were explored. Drugs and targets extracted from DrugBank [20–23] and TTD [13–17] were combined with genomic variants extracted from

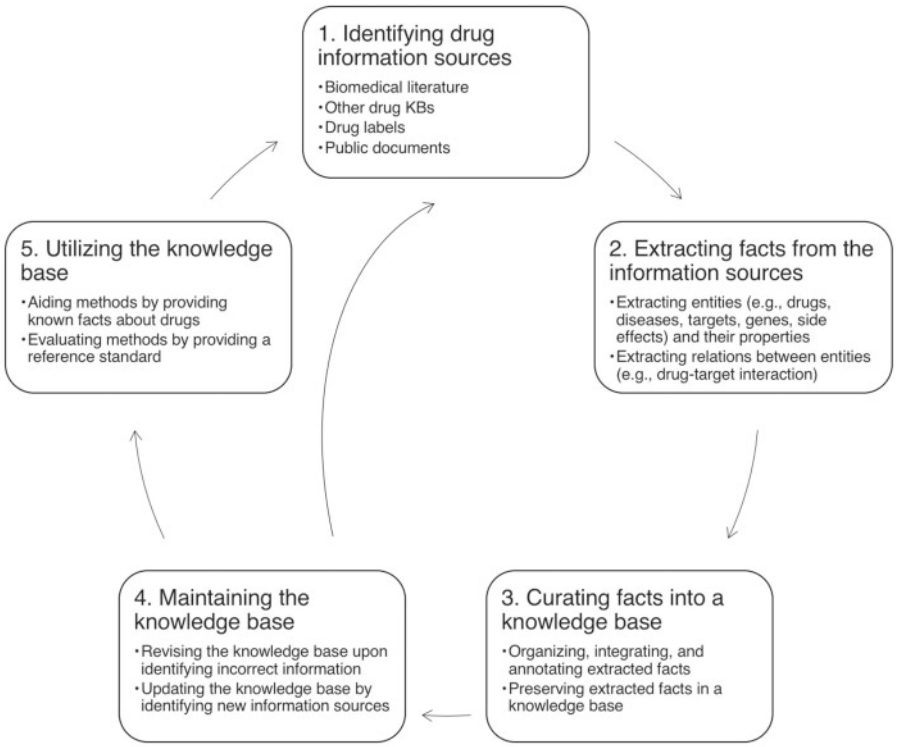

Figure 2. Drug knowledge base life cycle. Construction of a drug knowledge base starts from identifying drug-related information sources. From the information sources, drug-related entities, their properties and relations between the entities are extracted. The extracted facts are curated in a knowledge base. Curated knowledge base needs to be revised if incorrect information is identified and updated if new drug information sources are available. Constructed knowledge base is either used to aid or evaluate informatics methods.

the Database of Genomic Variants (DGV) [85]. Drugs associated with targets that are subject to genomic structural variation were identified. Pakhomov *et al.* [84] applied a machine learning approach to PubMed abstracts to identify drug–gene interactions. The text of PubMed abstracts was used as features, and drug–gene relations extracted from PharmGKB were used to supervise the model to identify whether the mentioned drugs and genes in an abstract are related.

Others

Several studies that do not fall into the aforementioned categories tackle various drug-related problems, such as drug classification [86, 87], drug clustering [89, 90], drug–target network analysis [91, 92], potential addictive drug prediction [93], beneficial drug combination prediction [94] and compound–target interaction prediction [96]. Re and Valentini [86] used annotations obtained from DrugBank [20–23] to predict therapeutic categories of drugs for drug repositioning. The study treated the problem as a drug ranking problem, given specific DrugBank therapeutic categories. Different pharmacological networks obtained from DrugBank and STITCH [88] were integrated to be used in drug ranking. Lötsch and Ultsch [87] investigated the functional genomics-based approach (as opposed to the approach of using drug targets) in drug classification. A drug target–biological process matrix was constructed by combining a drug–gene matrix and a gene–biological process matrix obtained from DrugBank and Gene Ontology [48, 49]. The matrices were projected onto a toroid grid using an artificial neural network approach (i.e. self-organizing map [114]), which was later visualized to identify clusters. Udrescu *et al.* [89] applied community detection algorithms to a drug–drug interaction network to cluster drugs. The drug–drug interaction network was constructed using information from an earlier version of DrugBank. A later version of DrugBank and other sources were used to interpret the clustering results. Papanikolaou *et al.* [90] applied text mining approaches to cluster DrugBank records and discover drug associations. Text fields of DrugBank such as 'Description', 'Indication', 'Pharmacodynamics' and 'Mechanism of Action' were extracted and terms of the fields were used to apply a set of similarity measures and clustering algorithms to generate clusters. Sun *et al.* [91] proposed a Web-based tool that constructs drug–target interaction networks. Drugs and related information were extracted from DrugBank [20–23] and PharmGKB [10–12] and integrated into a central database. The tool receives user queries and then searches the database to construct networks that reflect the queries to aid network-based analyses and visualizations. Sun *et al.* [92] constructed a drug–gene interaction network and a drug–drug interaction networks based on information from DrugBank. Network analyses were performed to examine functional and network features of targets as well as characteristics of specific groups of drugs. Sun *et al.* [93] used a network-based approach to explore potential drugs for addiction based on known addictive drugs. Addictive drugs, their targets and nonaddictive drugs that share at least one target with any addictive drug were extracted from DrugBank to construct a drug–target interaction network. The network was explored to identify useful information for predicting potential addictive drugs. Iwata *et al.* [94] proposed a machine learning approach to predict beneficial drug combinations based on information about drug–target interactions and the Anatomical Therapeutic Chemical (ATC) Classification

System [95]. Drug information was obtained from KEGG DRUG [19–24], DrugBank [20–23], TTD [13–17] and other databases. Known beneficial drug combinations were obtained from KEGG DRUG, and each drug–drug pair was represented by combining two drug profiles. Logistic regression was used as the prediction model. Keum *et al.* [96] used drug and protein information obtained from DrugBank [20–23] to predict compound–target interactions of natural products. Compounds, target proteins and their interactions were obtained from DrugBank. Chemical similarities among compounds and genomic similarities among target proteins were computed. Computed similarities were combined with herbal compound data to predict interactions between compounds of herbs and target proteins using Bipartite Local Model [115] and SVM [110].

Challenges in drug knowledge base life cycle

So far, we have reviewed the popular drug knowledge bases and the biomedical applications using them. In this section, we will define a drug knowledge base life cycle and discuss challenges associated with it. Figure 2 illustrates five major steps of the drug knowledge base life cycle. The first step of constructing a drug knowledge base is identifying drug-related information sources such as scientific literature. From the identified information sources, drug-related entities (e.g. drugs; diseases; targets; genes; side effects), their properties (e.g. drug categories, drug brands) and relations between the entities are extracted. These extracted facts are then curated into a knowledge base. Curation includes organization, integration, annotation, and preservation. Curated knowledge base needs to be revised if incorrect information is identified and updated if new drug information sources are available. Constructed knowledge base is either used to aid or evaluate informatics methods.

There are a few challenges across the life cycle, ranging from knowledge extraction from information sources to effective utilization of drug knowledge bases. In the following, we discuss those major challenges one by one.

Semiautomation of knowledge extraction process

Knowledge base construction is a time-consuming process that involves a significant amount of human efforts. Information collected from different sources is gone through one or more manual review processes before being curated into knowledge bases. While automated methods are being used to assist the curation, they are of a basic level, e.g. retrieving documents of certain topics or extracting sentences that contain certain drug names. Although these automated methods are of great help, human curators still need to invest much efforts to extract facts. While the full automation of the entire process is not attainable with current technologies, there is a need for semiautomated methods that can assist human curators in a more upgraded manner. For example, given a drug side effect, a semiautomated method can retrieve two sets of literature that have contradictory assertions, so that human curators save the time of manually identifying different assertions across the retrieved literature and make more informed decisions.

Integration of drug knowledge bases

A large amount of efforts has been devoted to construct drug knowledge bases. These knowledge bases were constructed with different goals, and they have different foci and coverage. Many studies we reviewed integrated information collected from different drug knowledge bases to solve problems. While these integrations were performed at a limited scale (i.e. not integrating knowledge bases as a whole), there are a few challenges when performing integration at a larger scale. The first challenge is entity matching (or mapping) between two or more knowledge bases. While a few drug knowledge bases provide crosslinks to others for entity matching, the information is not complete, i.e. not all entities have cross-links, and there are still many drug knowledge bases that do not provide such service. Another challenge is integrating relations of entities. It is not a problem if there is only one type of relation between two entities. However, if more than one types of relation exist between two entities, we first need to identify all possible relations between the two entities and review how these relations are represented in different drug knowledge bases. The integration becomes more difficult if the relations are represented as free text without structured format.

Keeping drug knowledge bases up to date

The main source of drug knowledge bases is scientific literature, and one of the most important tasks of creating drug knowledge bases is extracting facts from relevant literature. Ever-increasing biomedical literature presents two challenges to the maintenance of drug knowledge bases. First, because the volume of the literature that needs to be reviewed is increasing, more human involvement is required. Second, revisions are needed in case later studies present new evidences that refute assertions made in previous studies. These challenges, together with other issues, make the maintenance of drug knowledge bases difficult. Many existing drug knowledge bases either have not been updated since their first release or being updated intermittently (e.g. at an interval of 2 years). Keeping drug knowledge bases up to date with the latest literature is a demanding task.

Drug knowledge bases as training and test sets

Many studies reviewed in the previous section used machine learning approaches to solve many drug-related problems. A common point of these studies is that they use facts extracted from drug knowledge bases as positive training sets while treating others that are not curated in the knowledge bases as negative training sets. For example, in the task of drug–target interaction prediction, known drug–target interactions extracted from drug knowledge bases were used as positive training sets, while randomly generated, unknown drug–target interactions were treated as negative training sets. This approach has its own justification; however, it is not always true because there are so many facts that we have not discovered, and thus, have not been curated. More informed ways of using drug knowledge bases as training sets are needed. Many studies also used drug knowledge bases for evaluation. For example, SIDER was used to test machine learning models of predicting drug side effects. Although many studies proposed models that address the same task, cross-validations were impossible because they used different parts (also varying in size) of drug knowledge bases for evaluation. A community-wise effort is needed to explicitly define several guidelines of using drug knowledge bases for machine learning tasks, so that cross-validation is available and many efforts are centralized.

Future directions

Until now, we have introduced the existing popular drug knowledge bases, the biomedical applications using them, as well as the various challenges on a drug knowledge base life cycle. In this section, we discuss a few future research directions toward a better ecosystem of drug knowledge bases.

Drug knowledge base integration

Integration of drug knowledge bases, despite its challenges, has many potentials. It is reported that existing drug knowledge bases has little overlap [40]. The little overlap is largely because of different foci of the drug knowledge bases. For example, some drug knowledge bases such as PharmGKB [10–12] mainly focus on drug–gene relations, whereas others have different foci (e.g. drug side effects of SIDER [33, 34]). Integrating existing drug knowledge bases enables much more granular analyses by providing more comprehensive views with richer information. For example, current precision medicine studies [82–85] mainly used drug–gene interactions to understand how drugs affect people with different genes. By taking other relations such as drug–disease and drug–side effect relations into consideration, we can perform more detailed analyses such as identifying what drugs affect people with certain diseases and gene variants to cause what side effects. Such a detailed analysis is only possible when we have an integrated hub of drug knowledge bases. Owing to its importance, an open-source community effort has been established to develop an integrated knowledge base of drugs and health outcomes of interest [116].

Drug knowledge base implementation

Existing drug knowledge bases are presented with two types: drug databases and drug ontologies. The former type usually provides Web search interfaces for users to perform simple queries. It also supports downloads of raw data files for researchers to perform various analyses. The latter type, drug ontologies, uses W3C's Semantic Web technologies such as RDF [117] and OWL [118] to manage knowledge. A difficulty of using downloaded raw files is that it is not trivial to understand the structures of the data files. Because data are in the tabular format, it is not intuitive to understand relational information among entities. Compare with drug databases, drug ontologies have more explicit descriptions about entities and their relations. While the data in drug ontologies can be viewed using tools such as Protégé [119], for systematic and bulk accesses to the data, a specialized ontology query language (i.e. SPARQL [120]) is needed, which is a factor that affects their active utilization. As our knowledge on drugs is continuously increasing, scalability of drug knowledge bases is an important issue. An alternative approach of implementing drug knowledges bases is graph databases [121]. A graph database uses graph structures to manage data, which makes it an ideal solution for managing biomedical entities, their properties and their relations. It has been reported that graph databases scale better and load data faster than RDF stores, which are physical implementations of ontologies [122]. As a member of NoSQL databases [123], which are designed to address challenges in big data management, graph databases scale well and are a good data management system for drug knowledge bases as well as a drug knowledge graph.

Improving drug knowledge bases with predicted results

Many studies we reviewed tackle prediction tasks, such as drug–target interaction prediction [59–66, 103, 104], drug–disease interaction prediction [67, 68], drug side effect prediction [74–77, 107, 108] and drug–drug interaction prediction [78–81]. These studies produced many valuable results for future studies to further explore, validate and then make use of them. However, there are no central repositories that systematically curate these predicted results. Therefore, these valuable assets are not widely visible, and their impacts are usually limited to the level of individual studies. Drug knowledge bases that curate predicted results of various studies are good knowledge sources that complement drug knowledge bases of known facts. For example, if many studies predict there is an interaction between a pair of drugs, it is highly possible that the two drugs interact. Such information can be integrated to perform more informed and concentrated analyses.

Drug knowledge bases of negative samples

Existing drug knowledge bases curate positive cases. For example, if two drugs interact, then they are curated into a drug knowledge base. On the other hand, if two drugs are known to not interact with each other, this information is not curated. While it is not trivial to identify negative samples, which might be more difficult than identifying positive samples, many applications are possible with these negative samples. With these negative samples, we can obtain more accurate machine learning models because researchers no longer need to artificially generate negative samples, which might be incorrect. Negative samples can also be used to aid prediction of positive samples. For example, if two drugs do not interact with each other, a third drug that has a similar chemical structure with one of the two drugs has a high possibility not to interact with the other one. This information, combined together, can form a large network, which is a valuable resource to apply community detection methods [124] that can be applied to networks with positive and negative links. With negative samples providing additional information to the network, more meaningful results can be obtainable.

Using social media platforms and medical forums

Not only researchers and practitioners can be contributors of drug knowledge bases. Patients and their family members can also provide useful information that can be properly used to construct drug knowledge bases. For example, patients and their family members may talk about drug side effects they experienced on social media platforms and/or medical forums. Although FDA maintains FDA Adverse Event Reporting System (FAERS) [125] for health-care professionals and consumers to report drug adverse events, social media platforms and medical forums have much broader reach and are easier for patients and their families to use. Because drug-related information collected via crowdsourcing may contain noise (typos, use of incorrect terms, etc.), the curation needs a different approach from the one used to construct traditional drug knowledge bases. Even though the process is not trivial, if successfully implemented, the health-care consumer-based drug knowledge bases would be a valuable resource for research.

Key Points

- Advances in biomedical research have generated a large volume of drug-related data, and many drug knowledge bases have been constructed.
- Drug knowledge bases contain valuable information on entities (i.e. drugs, diseases, targets, genes and side effects) and their relations (i.e. drug–drug, drug–gene, drug–disease, drug–target and drug–side effect).
- Drug knowledge bases have been applied to tasks, such as biomedical text mining, drug repositioning, ADR analysis and pharmacogenomic analysis.
- Integrating drug knowledge bases and implementing a scalable drug knowledge graph enable much more granular analyses by providing more comprehensive views with richer information.
- Improving drug knowledge bases using predicted results and social media platforms and medical forums would provide additional insights.

Funding

The work of Fei Wang is supported by NSF IIS-1650723 and NSF IIS-1716432.

References

1. Chen H, Ding L, Wu Z, *et al*. Semantic web for integrated network analysis in biomedicine. *Brief Bioinform* 2009;**10**(2):177–92.
2. Barabási AL. Network medicine–from obesity to the "diseasome". *N Engl J Med* 2007;**357**(4):404–7.
3. Weinshilboum R. Inheritance and drug response. *N Engl J Med* 2003;**348**(6):529–37.
4. Evans WE, Relling MV. Moving towards individualized medicine with pharmacogenomics. *Nature* 2004;**429**(6990):464–8.
5. Sun X, Vilar S, Tatonetti NP. High-throughput methods for combinatorial drug discovery. *Sci Transl Med* 2013;**5**(205):205rv1.
6. Gilson MK, Liu T, Baitaluk M, *et al*. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 2016;**44**(D1):D1045–53.
7. Bento AP, Gaulton A, Hersey A, *et al*. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 2014;**42**:D1083–90.
8. Nelson SJ, Zeng K, Kilbourne J, *et al*. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* 2011;**18**(4):441–8.
9. U.S. National Library of Medicine. National Drug File – Reference Terminology Source Information. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT (31 May 2017, date last accessed).
10. Klein TE, Chang JT, Cho MK, *et al*. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics research network and knowledge base. *Pharmacogenomics J* 2001;**1**(3):167–70.
11. Hewett M, Oliver DE, Rubin DL, *et al*. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res* 2002;**30**(1):163–5.
12. Whirl-Carrillo M, McDonagh EM, Hebert JM, *et al*. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012;**92**(4):414–7.
13. Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. *Nucleic Acids Res* 2002;**30**(1):412–5.
14. Zhu F, Han B, Kumar P, *et al*. Update of TTD: therapeutic target database. *Nucleic Acids Res* 2010;**38**: D787–91.
15. Zhu F, Shi Z, Qin C, *et al*. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 2012;**40**(D1):D1128–36.
16. Qin C, Zhang C, Zhu F, *et al*. Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res* 2014;**42**: D1118–23.
17. Yang H, Qin C, Li YH, *et al*. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res* 2016;**44**(D1):D1069–74.
18. U.S. National Library of Medicine. DailyMed. https://dailymed.nlm.nih.gov/dailymed (31 May 2017, date last accessed).
19. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**:D354–7.
20. Kanehisa M, Araki M, Goto S, *et al*. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 2008;**36**:D480–4.
21. Kanehisa M, Goto S, Furumichi M, *et al*. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010;**38**:D355–60.
22. Kanehisa M, Goto S, Sato Y, *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;**40**:D109–14.
23. Kanehisa M, Goto S, Sato Y, *et al*. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;**42**:D199–205.
24. Kanehisa M, Sato Y, Kawashima M, *et al*. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**(D1):D457–62.
25. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**(90001):D668–72.
26. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;**36**:D901–6.
27. Knox C, Law V, Jewison T, *et al*. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* 2011;**39**:D1035–41.
28. Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014;**42**:D1091–7.
29. Günther S, Kuhn M, Dunkel M, *et al*. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res* 2008;**36**:D919–22.
30. Hecker N, Ahmed J, von Eichborn J, *et al*. SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res* 2012;**40**:D1113–7.
31. Boyce R, Collins C, Horn J, *et al*. Computing with evidence Part I: a drug-mechanism evidence taxonomy oriented toward confidence assignment. *J Biomed Inform* 2009;**42**(6):979–89.
32. Boyce R, Collins C, Horn J, *et al*. Computing with evidence Part II: an evidential approach to predicting metabolic drug-drug interactions. *J Biomed Inform* 2009;**42**(6):990–1003.
33. Kuhn M, Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;**6**:343.
34. Kuhn M, Letunic I, Jensen LJ, *et al*. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**(D1):D1075–9.

35. Griffith M, Griffith OL, Coffman AC, *et al*. DGIdb: mining the druggable genome. *Nat Methods* 2013;**10**(12):1209–10.
36. Wagner AH, Coffman AC, Ainscough BJ, *et al*. DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res* 2016;**44**(D1):D1036–44.
37. Hanna J, Joseph E, Brochhausen M, *et al*. Building a drug ontology based on RxNorm and other sources. *J Biomed Semantics* 2013;**4**(1):44.
38. Hogan WR, Hanna J, Hicks A, *et al*. Therapeutic indications and other use-case-driven updates in the drug ontology: anti-malarials, anti-hypertensives, opioid analgesics, and a large term request. *J Biomed Semantics* 2017;**8**(1):10.
39. Herrero-Zazo M, Segura-Bedmar I, Hastings J, *et al*. DINTO: using OWL ontologies and SWRL rules to infer drug-drug interactions and their mechanisms. *J Chem Inf Model* 2015; **55**(8):1698–707.
40. Ayvaz S, Horn J, Hassanzadeh O, *et al*. Toward a complete dataset of drug-drug interaction information from publicly available sources. *J Biomed Inform* 2015;**55**:206–17.
41. Sharp ME. Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. *J Biomed Semantics* 2017;**8**(1):2.
42. Sherry ST, Ward MH, Kholodov M, *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;**29**(1): 308–11.
43. Goede A, Dunkel M, Mester N, *et al*. SuperDrug: a conformational drug database. *Bioinformatics* 2005;**21**(9):1751–3.
44. Health Canada. *Canadian Adverse Drug Reaction Monitoring Program*. http://www.hc-sc.gc.ca/ahc-asc/activit/atip-aiprp/priv-prot/pia-efvp-a-eng.php (1 May 2017, date last accessed).
45. Kim S, Thiessen PA, Bolton EE, *et al*. PubChem Substance and Compound databases. *Nucleic Acids Res* 2016;**44**(D1): D1202–13.
46. Wang Y, Bryant SH, Cheng T, *et al*. PubChem BioAssay: 2017 update. *Nucleic Acids Res* 2017;**45**(D1):D955–63.
47. National Center for Biomedical Ontology. *BioPortal*. https://bioportal.bioontology.org (31 May 2017, date last accessed).
48. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**(1):25–9.
49. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res* 2015;**43**: D1049–56.
50. U.S. Food & Drug Administration. *National Drug Code Directory*. https://www.accessdata.fda.gov/scripts/cder/ndc (31 May 2017, date last accessed).
51. Hastings J, de Matos P, Dekker A, *et al*. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 2013;**41**: D456–63.
52. Herrero-Zazo M, Segura-Bedmar I, Martínez P, *et al*. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. *J Biomed Inform* 2013; **46**(5):914–20.
53. He Y, Sarntivijai S, Lin Y, *et al*. OAE: the ontology of adverse events. *J Biomed Semantics* 2014;**5**:29.
54. He L, Yang Z, Lin H, Li Y. Drug name recognition in biomedical texts: a machine-learning-based method. *Drug Discov Today* 2014;**19**(5):610–7.
55. Korkontzelos I, Piliouras D, Dowsey AW, *et al*. Boosting drug named entity recognition using an aggregate classifier. *Artif Intell Med* 2015;**65**(2):145–53.
56. Theobald M, Shah N, Shrager J. Extraction of conditional probabilities of the relationships between drugs, diseases, and genes from pubmed guided by relationships in PharmGKB. *Summit Transl Bioinform* 2009;**2009**:124–8.
57. Clematide S, Rinaldi F. Ranking relations between diseases, drugs and genes for a curation task. *J Biomed Semantics* 2012; **3(Suppl 3)**:S5.
58. Percha B, Altman RB. Learning the structure of biomedical relationships from unstructured text. *PLoS Comput Biol* 2015; **11**(7):e1004216.
59. Li YY, An J, Jones SJ. A computational approach to finding novel targets for existing drugs. *PLoS Comput Biol* 2011;**7**(9): e1002139.
60. Tari L, Vo N, Liang S, *et al*. Identifying novel drug indications through automated reasoning. *PLoS One* 2012;**7**(7):e40946.
61. Cobanoglu MC, Liu C, Hu F, *et al*. Predicting drug-target interactions using probabilistic matrix factorization. *J Chem Inf Model* 2013;**53**(12):3399–409.
62. Yamanishi Y, Kotera M, Moriya Y, *et al*. DINIES: drug-target interaction network inference engine based on supervised analysis. *Nucleic Acids Res* 2014;**42**(W1):W39–45.
63. Tao C, Sun J, Zheng WJ, *et al*. Colorectal cancer drug target prediction using ontology-based inference and network analysis. *Database* 2015;**2015**:bav015. pii: bav015.
64. Zhang M, Luo H, Xi Z, *et al*. Drug repositioning for diabetes based on 'omics' data mining. *PLoS One* 2015;**10**(5):e0126082.
65. Seal A, Ahn YY, Wild DJ. Optimizing drug-target interaction prediction based on random walk on heterogeneous networks. *J Cheminform* 2015;**7**(1):40.
66. Yuan Q, Gao J, Wu D, *et al*. DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 2016;**32**(12): i18–27.
67. Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. *PLoS One* 2011;**6**(12):e28025.
68. Bisgin H, Liu Z, Fang H, *et al*. A phenome-guided drug repositioning through a latent variable model. *BMC Bioinformatics* 2014;**15**:267.
69. Wang J, Li ZX, Qiu CX, *et al*. The relationship between rational drug design and drug side effects. *Brief Bioinform* 2012; **13**(3):377–82.
70. Bresso E, Grisoni R, Marchetti G, *et al*. Integrative relational machine-learning for understanding drug side-effect profiles. *BMC Bioinformatics* 2013;**14**:207.
71. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics* 2011;**12**:169.
72. Jahid MJ, Ruan J. An ensemble approach for drug side effect prediction. In: *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, China, December 18–21, 2013*, 440–5.
73. LaBute MX, Zhang X, Lenderman J, *et al*. Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *PLoS One* 2014;**9**(9):e106298.
74. Eshleman R, Singh R. Leveraging graph topology and semantic context for pharmacovigilance through twitter-streams. *BMC Bioinformatics* 2016;**17**(S13):335.
75. Jamal S, Goyal S, Shanker A, *et al*. Predicting neurological Adverse Drug Reactions based on biological, chemical and phenotypic properties of drugs using machine learning models. *Sci Rep* 2017;**7**(1):872.
76. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**(D1):D158–69.
77. Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res* 2000;**28**(1):235–42.

78. Vilar S, Harpaz R, Uriarte E, *et al*. Drug-drug interaction through molecular structure similarity analysis. *J Am Med Inform Assoc* 2012;**19**(6):1066–74.
79. He L, Yang Z, Zhao Z, *et al*. Extracting drug-drug interaction from the biomedical literature using a stacked generalization-based approach. *PLoS One* 2013;**8**(6):e65814.
80. Cheng F, Zhao Z. Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *J Am Med Inform Assoc* 2014;**21**(e2):e278–86.
81. Hameed PN, Verspoor K, Kusljic S, *et al*. Positive-Unlabeled Learning for inferring drug interactions based on heterogeneous attributes. *BMC Bioinformatics* 2017;**18**(1):140.
82. Rance B, Doughty E, Demner-Fushman D, *et al*. A mutation-centric approach to identifying pharmacogenomic relations in text. *J Biomed Inform* 2012;**45**(5):835–41.
83. Rasmussen HB, Dahmcke CM. Genome-wide identification of structural variants in genes encoding drug targets: possible implications for individualized drug therapy. *Pharmacogenet Genomics* 2012;**22**(7):471–83.
84. Pakhomov S, McInnes BT, Lamba J, *et al*. Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies. *J Biomed Inform* 2012; **45**(5):862–9.
85. Iafrate AJ, Feuk L, Rivera MN, *et al*. Detection of large-scale variation in the human genome. *Nat Genet* 2004;**36**(9):949–51.
86. Re M, Valentini G. Network-based drug ranking and repositioning with respect to DrugBank therapeutic categories. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**(6): 1359–71.
87. Lötsch J, Ultsch A. A machine-learned computational functional genomics-based approach to drug classification. *Eur J Clin Pharmacol* 2016;**72**(12):1449–61.
88. Szklarczyk D, Santos A, von Mering C, *et al*. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* 2016;**44**(D1):D380–4.
89. Udrescu L, Sbârcea L, Topîrceanu A, *et al*. Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing. *Sci Rep* 2016;**6**:32745.
90. Papanikolaou N, Pavlopoulos GA, Theodosiou T, *et al*. DrugQuest - a text mining workflow for drug association discovery. *BMC Bioinformatics* 2016;**17(Suppl 5)**:182.
91. Sun J, Wu Y, Xu H, *et al*. DTome: a web-based tool for drug-target interactome construction. *BMC Bioinformatics* 2012; **13(Suppl 9)**:S7.
92. Sun J, Xu H, Zhao Z. Network-assisted investigation of antipsychotic drugs and their targets. *Chem Biodivers* 2012;**9**(5): 900–10.
93. Sun J, Huang LC, Xu H, *et al*. Network-assisted prediction of potential drugs for addiction. *Biomed Res Int* 2014;**2014**: 258784.
94. Iwata H, Sawada R, Mizutani S, *et al*. Large-scale prediction of beneficial drug combinations using drug efficacy and target profiles. *J Chem Inf Model* 2015;**55**(12):2705–16.
96. Keum J, Yoo S, Lee D, *et al*. Prediction of compound-target interactions of natural products using large-scale drug and protein information. *BMC Bioinformatics* 2016;**17**(S6):219.
95. World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC). *ATC Structure and Principles*. https://www.whocc.no/atc/structure_and_principles (31 May 2017, date last accessed).
97. Zweigenbaum P, Demner-Fushman D, Yu H, *et al*. Frontiers of biomedical text mining: current progress. *Brief Bioinform* 2007;**8**(5):358–75.
98. Talukdar PP, Brants T, Liberman M, *et al*. A context pattern induction method for named entity extraction. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning, New York City, New York, June 08–09, 2006*. Association for Computational Linguistics, 141–8.
99. Li Y, Lin H, Yang Z. Incorporating rich background knowledge for gene named entity classification and recognition. *BMC Bioinformatics* 2009;**10**:223.
100. Li D, Kipper-Schuler K, Savova G. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Columbus, Ohio, June 19, 2008*. Association for Computational Linguistics, 94–5.
101. Davis AP, Grondin CJ, Johnson RJ, *et al*. The comparative toxicogenomics database: update 2017. *Nucleic Acids Res* 2017; **45**(D1):D972–8.
102. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;**3**(8):673–83.
103. Cowley MJ, Pinese M, Kassahn KS, *et al*. PINA v2.0: mining interactome modules. *Nucleic Acids Res* 2012;**40**:D862–5.
104. Tong H, Faloutsos C, Pan JY. Random walk with restart: fast solutions and applications. *Knowl Inf Syst* 2008;**14**(3):327–46.
105. Blei DM, AY NG, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003;**3**:993–1022.
106. Pirmohamed M, Breckenridge AM, Kitteringham NR, *et al*. Adverse drug reactions. *BMJ* 1998;**316**(7140):1295–8.
107. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001;17–21.
108. Breiman L. Random forests. *Machine Learning* 2001;**45**(1): 5–32.
109. Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. In: *Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, California, July 12–16, 1992*, 129–34.
110. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. *Neural Process Lett* 1999;**9**(3):293–300.
111. Alahakoon D, Halgamuge SK, Srinivasan B. Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Trans Neural Netw* 2000;**11**(3):601–14.
112. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;**372**(9):793–5.
113. Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. *Transl Res* 2009;**154**(6): 277–87.
114. Kohonen T. The self-organizing map. *Neurocomputing* 1998; **21**(1):1–6.
115. Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics* 2009;**25**(18):2397–403.
116. Boyce RD, Ryan PB, Norén GN, *et al*. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. *Drug Saf* 2014;**37**(8):557–67.
117. The World Wide Web Consortium (W3C). *Resource Description Framework (RDF)*. https://www.w3.org/RDF (31 May 2017, date last accessed).

118. The World Wide Web Consortium (W3C). *Web Ontology Language (OWL)*. https://www.w3.org/OWL (31 May 2017, date last accessed).
119. Musen MA; Protégé Team. The Protégé project: a look back and a look forward. *AI Matters* 2015;**1**(4):4–12.
120. The World Wide Web Consortium (W3C). *SPARQL Query Language for RDF*. https://www.w3.org/TR/rdf-sparql-query (31 May 2017, date last accessed).
121. Webber J. A programmatic introduction to neo4j. In: *Proceedin gs of the Third Annual Conference on Systems, Programming, and Applications: Software for Humanity, Tucson, Arizona, October 21–25, 2012*. Association for Computational Linguistics, 217–18.
122. Dominguez-Sal D, Urbón-Bayes P, Giménez-Vanó A, *et al*. Survey of graph database performance on the HPC scalable graph analysis benchmark. In: *Proceedings of the Workshops of the 11th International Conference on Web-Age Information Management, Jiuzhaigou, China, July 15–17, 2010*, Springer, 37–48.
123. Cattell R. Scalable SQL and NoSQL data stores. *ACM SIGMOD Record* 2011;**39**(4):12–27.
124. Traag VA, Bruggeman J. Community detection in networks with positive and negative links. *Phys Rev E Stat Nonlin Soft Matter Phys* 2009;**80**(3 Pt 2):036115.
125. U.S. Food & Drug Administration. *Questions and Answers on FDA's Adverse Event Reporting System (FAERS)*. https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm (31 May 2017, date last accessed).