Artificial Intelligence In Medicine 96 (2019) 80–92

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Mining heterogeneous network for drug repositioning using phenotypic information extracted from social media and pharmaceutical databases

Christopher C. Yang*, Mengnan Zhao

College of Computing and Informatics, Drexel University, Philadelphia, PA, United States

ARTICLE INFO

Keywords:
Drug repositioning
Heterogeneous network mining
Online health community
Phenotype
Social media

ABSTRACT

Drug repositioning has drawn significant attention for drug development in pharmaceutical research and industry, because of its advantages in cost and time compared with the de novo drug development. The availability of biomedical databases and online health-related information, as well as the high-performance computing, empowers the development of computational drug repositioning methods. In this work, we developed a systematic approach that identifies repositioning drugs based on heterogeneous network mining using both pharmaceutical databases (PharmGKB and SIDER) and online health community (MedHelp). By utilizing adverse drug reactions (ADRs) as the intermediate, we constructed a heterogeneous health network containing drugs, diseases, and ADRs, and developed path-based heterogeneous network mining approaches for drug repositioning. Additionally, we investigated on how the data sources affect the performance on drug repositioning. Experiment results showed that combining both PharmKGB and MedHelp identified 479 repositioning drugs, which are more than the repositioning drugs discovered by other alternatives. In addition, 31% of the 479 of the discovered repositioning drugs were supported by evidence from PubMed.

1. Introduction

Over the past decades, de novo drug development has become costly and time-consuming, with the success rate of less than 10% [1] despite the increasing investment in R&D and the progress in life science and technology [2]. The number of newly developed drugs that can enter preclinical tests and clinical trials has gradually declined [3] and the number of newly approved drugs has not kept up with the consistent increases in pharmaceutical R&D spending. In light of these challenges, drug repositioning receives increasing attention from both academia and pharmaceutical companies, becoming an alternative and promising way for drug development. Drug repositioning has contributed about 30% of the new FDA approved drugs in recent years [4] for instance, repositioning drugs accounted for 20% among 84 new drugs brought to market in 2013 [5].

Drug repositioning is the application of discovering new indications for existing drugs [6] and plays a key role in drug development and healthcare industry. Supported by governments, nonprofit organizations and academic institutions, and under the economic incentives, a number of drug repositioning studies have been conducted and achieved various degrees of success [7]. One notable example of successful drug repositioning is about Gabapentin, which was initially developed for epileptics but was found to be effective for treating anxiety disorders and neuropathic pain [8]; another example is Plerixafor, which was initially developed as an inhibitor of HIV but was repurposed as a stem cell mobilizing drug later [9]. However, most of the successful stories are owing to serendipity. Hence, there is a desire of a systematic approach of drug repositioning.

Compared with the traditional drug development from molecule to product, drug repositioning is more time- and cost-efficient, accelerating drug discovery process. By estimation, it usually takes ten to twelve years to develop a new drug to market, and it costs drug companies an average of $1.2 billion – as high as $5 billion-before a drug available for sale [3]. On the other hand, the repositioned drugs have already been validated by pharmaceutical and toxicological tests. The time and cost in the early-stage development and the risk of failure are reduced significantly. For instance, the time of introducing a repositioned drug to market could be shortened to three years [10]. The other advantages of drug repositioning embody in its potential for cancer, orphan diseases, and personalized medicine. Due to the high demand for anti-cancer drugs and the limited efficacy in current anti-cancer drug development, drug repositioning has been a promising and effective way for searching anti-cancer therapeutics [11]. Due to the limited attention and investment for orphan and rare diseases, drug repositioning has been an alternative and important approach to identify novel therapeutics from known drugs [12]. Beside, advances in

* Corresponding author.
E-mail address: chris.yang@drexel.edu (C.C. Yang).

https://doi.org/10.1016/j.artmed.2019.03.003
Received 1 June 2018; Received in revised form 24 February 2019; Accepted 5 March 2019

medical science and technologies enable personalized genomic studies for each patient and assist to determine the underlying causes of diseases, when possible, repositioned drugs become an efficient option to provide a personalized treatment [13].

Depending on where the discovery derives from, current computational drug-repositioning methods can be classified as either disease-based or drug-based [14]. Disease-based methods usually exploit knowledge of symptomatology, phenotype and pathology while drug-based methods mostly utilized the characteristics of drug compounds such as chemical structures, pharmacological properties and molecular activities [15]. Since different repositioning methods need different aspects of information of drugs and diseases, such as genetic, chemical, pharmacological, clinical, and protein information, multiple data sources, including Drugs@FDA, Gene Expression Omnibus (GEO), Pharmacogenomics Knowledge Base (PharmGKB), Side Effect Resource (SIDER), and DrugBank, have been exploited [2].

In this paper, we adopted the disease-based repositioning strategy by using adverse drug reactions (ADRs) as intermediary to discover novel disease–drug relationships. With the capability of profiling phenotypic expressions of drugs and converting the physiological consequences, ADR is becoming an important intermediary to connect drugs with diseases in drug repositioning and have been exploited to discover new therapeutic uses in some previous studies [16–18]. The rationale for an ADR-based drug repositioning approach is ADR and disease are both behavioral and physiological changes in response to the drug treatment, and if drugs treating a disease share the same ADR, that ADR may serve as a phenotypic "biomarker" for the disease. In other words, ADRs and indications of disease have similar phenotypic expressions because of similar underlying pathways and underlying mechanism of action (MOA) on human subjects [19]. Therefore, ADR are sometimes used as the intermediary between drug and disease or the "biomarker" for diseases in drug repositioning. In addition, considering the fact that ADRs are substantially under-reported in most medical systems and databases, it might lead to the insufficiency of such data sources. In this work, we utilized both social media data and pharmaceutical databases to extract ADRs and their associations with drugs and diseases for drug repositioning. We also investigated how the performance of drug repositioning would be influenced by the choices of data sources and whether social media data would achieve a better performance.

2. Literature review

In general, there are two underlying principles of drug repositioning. Firstly, drugs are confounding by nature. That means a drug can be linked with multiple targets and pathways. Secondly, drugs related to a certain disease may also work on other related diseases due to connections between diseases [6]. Based on these principles, the strategies used in systematic drug repositioning could be categorized into two classes depending on where the discoveries initiate from: (a) drug-based strategy and (b) disease-based strategy [15]. Besides, the computational approaches of drug repositioning usually include data mining, machine learning and network-based analysis [2]. Table 1 provides an overview of current drug repositioning studies.

2.1. Drug-based strategy

Drug-based strategy relies on the pharmacological, chemical, genomic and biomedical data to infer novel drug uses. This approach is preferred when rich information about drug characteristics is available or there is interest or knowledge in understanding how pharmacological properties lead to drug repositioning [15]. A majority of the drug-based strategies hold the assumption that drugs with similar profiles or structures are likely to share common indications. The most frequently used features of drug-based strategies include chemical structure and molecule information [20–24], and genome [25–29].

Chemical structure and molecule information are valuable sources for identify similar drugs for repositioning. Keiser et al. [20] integrated the structural similarity between drug compounds with the knowledge of drug–target relationships to infer novel drug–target relations. With drugs represented by a set of compounds and targets by a set of ligands, the prediction of whether a novel drug–target association is possible was calculated by the sum of similarities between each compound of the drug and each ligand of the target. In result, they predicted thousands of novel associations and tested thirty of them experimentally. Considering the fact that the chemical structures of drugs are so complex that they are not always consistent with the drugs' function, Tan et al. [22] developed a new form of "expression profile" for drugs by integrating chemical structure and gene semantic information to calculate the similarity between drugs. Li and Lu [21] developed a bipartite-graph based approach, with the underlying principle that if two drugs r1 and r2 are defined to be similar, and r1 is indicated for a disease d, then r2 could be seen as a repurposing candidate for treating d. The similarity between drugs combines both the similarity of chemical structures and that of target profiles, and the similarity of sharing target proteins was computed based on a bipartite graph. Zheng et al. [23] developed a new similarity measurement based on "ensemble". A protein is composed of several ligands. The ligands build a set and the set was seen as an ensemble. Instead of comparing two compounds to measure their similarity, this method compared a compound with the whole feature of an ensemble, because the ensemble often covers a small chemical space with structurally similar compounds. Kinnings et al. [24] computed the similarity between drugs by summarizing the transcriptional responses of a drug under multiple treatments, cell lines and dosages. They constructed a drug network, in which, an edge is created when two drugs share similar response profiles. A network partitioning approach was applied to classify the drugs into different communities and repositioning opportunities were discovered within each community.

Genome is another valuable source for drug repositioning in drug-based strategies. Ng et al. [29] introduced the algorithm-ligand Enrichment of Network Topological Similarity (ligENTS) to identify new drug indications by using drug–target interactions on genome scale. Meanwhile, as most of the existing algorithms only focus on finding local neighborhood for drugs, techniques used by ligENTS discover global relationships between chemicals. Rastegar-Mojarad et al. [30] used genes as the intermediary between drugs and diseases. They connected drugs and gene targets with DrugBank as data source and connected diseases and genes by applying large-scale phenome-wide association studies. Through these two connections, they inferred connections between diseases and drugs. Jiang et al. [25] focused on discovering connections between molecules and miRNAs in cancers with data coming from cMap [31]. They constructed a Small Molecule-MiRNA Network (SMirN) to discover new drug candidates. Rukov et al. [27] connected MiRNAs with drug effects via the pathway of MiRNA–gene–drug–drug effect.

2.2. Disease-based strategy

Given the principle that drugs associated to a certain disease or pathways can also be effective in other related diseases or pathways [6], disease-based strategy usually exploits disease-related knowledge such as phenotype (e.g., indication, side effect) and pathology to discover novel relationships between drugs and diseases. This approach is preferred when missing pharmacological knowledge or expertise in drugs, or when repositioning efforts are to be focused on a specific disease or therapeutic category [32,33].

Indication information is utilized in some disease-based approaches. For instance, based on the principle that if two diseases, D1 and D2, share some similar therapies (e.g., drugs), then the drugs that are current used for D1 can be seen as a candidate for the treatment of D2, Chiang et al. [34] applied a "guilt by association" approach to discover

Table 1
An overview of drug-repositioning studies.

		Drug repositioning strategies	
		Drug-based strategy	Disease-based strategy
Computational approaches	Data mining	Li and Lu [39] Okada et al. [28] Zhu et al. [41] Rastegar-Mojarad et al. [30] Zheng et al. [23]	Campillos et al. [36] Andronis et al. [38] Rastegar-Mojarad et al. [59] Nugent et al. [37]
	Machine learning	Lamb et al. [31] Napolitano et al. [43] Leaman et al. [46]	Gottlieb et al. [42] Yang et al. [45] Zhang et al. [44]
	Network-based analysis	Keiser et al. [20] Kinnings et al. [24] Jiang et al. [25] Li and Lu [21] Rukov et al. [27] Ng et al. [29] Tan et al. [22] Iorio et al. [26]	Chiang and Butte [34] Hu and Agarwal [50] Yang and Agarwal [19] Cheng et al. [48] Fukuoka et al. [14] Wu et al. [47] Wang et al. [58] Rakshit et al. [49]

new indications for drugs. Some other disease-based approaches are built on the assumption that a drug can be repositioned from one indication to another because the two indications share some aspects of underlying pathophysiology, which is responsive to the therapeutic effect of a drug [35].

Another approach to connect diseases with drugs is via their side effects (SE). Based on the rationale that side effects and diseases have similar phenotypic expressions because of similar underlying pathways [15]. Campillos et al. [36] used phenotypic side-effect similarities to infer whether two drugs share targets and to identify novel drug–target relationships. Yang and Agarwal [19] constructed a database of disease–SE relationships by using drug-SE data extracted from SIDER and drug–disease relationships from PharmGKB. They generated a confusion matrix for each disease–SE pair, in which, each cell represents the number of drugs listing or not listing a SE when that drug is indicated or not indicated for a disease. The association strength of a pair was measured by Matthews correlation coefficient (MCC), sensitivity and specificity. In a disease–SE matrix, the "false positive" drugs for a disease represents the drugs listing the SE but are not indicated to treat the disease. These "false positive" drugs are the repositioning candidates for the disease identified by Yang and Agarwal's approach [19]. According to the detected disease–ADR associations, they built Naive Bayes models to predict new indications for 145 diseases. The method was extended to predict indications for clinical compounds. Nugent et al. [37] developed a SE-based computational method based on Twitter data and used SE similarity between drugs to construct the drug–drug network, using inverse covariance estimation to find neighboring drugs for each drug.

2.3. Computational approaches

As a large volume of biomedical and pharmaceutical information grows immensely in databases and literature, computational approaches including data mining, machine learning and network analysis are gaining importance in systematical drug repositioning practices [2].

In the studies adopting data mining approach, a majority of them were literature based and adopted text mining techniques such as semantic inference and ontology model [38]. For instance, Li and Lu [39] automatically identified pharmacogenomics (PGx) relationships between genes, drugs and diseases from trial records in ClinicalTrials.gov by developing a dictionary-based text mining method. Rastegar-Mojarad et al. [40] adopted a literature-based method, relying on extracting drug–gene and gene–disease pairs from abstracts in Medline to infer drug–disease pairs with gene as intermediary. The underlying principle is that the association between drug and disease is based on how strong the associations of drug and gene, and gene and disease. Meanwhile, they utilized semantic predications, retrieved from SemMedDB, to infer new connections between drugs and diseases, and ranked the discovered pairs by calculating scores based on the quality of predicates. Zhu et al. [41] developed a meta-ontology model based on the pharmacogenomics data they extracted from PharmGKB, with base classes such as "drug" and "gene" and with relationships such as "associatedwithDrug" and "associatedwithDisease". Based on the model, they exploited semantic inference to identify new drug indications with the principle that a disease D is considered to be associated with a drug R if D is associated with R directly or associated with genes, single nucleotide polymorphisms (SNPs) or pathways that are associated with R.

Machine learning techniques can leverage the data from various data sources to identify medical entities (e.g. gene, compound, protein, drug, and disease), to reveal the underlying associations between these entities, and to explore repositioning opportunities. Gottlieb et al. [42] utilized multiple drug–drug similarity (chemical based, side effect based, sequence based, closeness in a PPI network, GO based) and disease–disease similarity (phenotype based, semantic phenotypic, genetic based) measures as classification features, and used a logistic regression classifier to predict novel drug indications. Napolitano et al. [43] predicted drug therapeutic class by using drug-related features (e.g. drug chemical structure similarity, drug molecular target similarity and drug gene expression similarity). They merged these features into a single drug similarity matrix, which was used as a kernel for SVM classification, and applied collaborative filtering techniques to predict unknown drug–disease associations. Zhang et al. [44] proposed a unified computational framework, DDR (multiple Drug information sources and multiple Disease information sources for Repositioning tasks) for integrating multiple aspects of drug similarity and disease similarity. Based on all this information, the authors formulated the drug–disease network analysis into an optimization problem and solved it using Block Coordinate Descent (BCD) strategy. Yang et al. [45] used a causal inference-probabilistic matrix factorization approach to infer drug–disease associations. Leaman et al. [46] adopted a model combination approach based on two different linear chain conditional random fields (CRF) models to identify chemical entities, where the two CRF models used different tokenizations, feature sets, CRF implementations, CRF parameters, and some variations in post processing, and demonstrated different performances in different aspects.

Network analysis is the most popular and widely used approach in computational drug repositioning, in both drug-based and disease-based strategies. By highlighting the network concept, network-based methods have shown the great capability for deciphering mechanisms,

interactions, and MOAs underlying drugs, diseases and other medical entities [47]. Wu et al. [47] presented various ways to construct the connections between medical entities. Besides a majority of similarity-based methods resort to network knowledge to identify the repositioning candidates, various types of networks have been constructed to extract novel drug targets or drug–disease relationships. In these networks, nodes can be drugs, diseases, targets, side effects, genes and proteins. Links can be the similarity between the same type of nodes or associations between different types of node such as drug–drug similarity and drug–disease association.

"Guilt by association" principle is one common way of utilizing the network to find novel drug–target or drug–disease relationships. For instance, Chiang et al. [34] implemented a network-based method, where the network was constructed with disease as nodes and whether the two corresponding diseases share FDA-approved drugs as links. Novel drug uses were suggested by "guilt by association" approach with weakly suggested drug uses (those with only one suggestion) removed.

Inference based on network topological features is another important approach for repositioning. Cheng et al. [48] utilized bipartite network topology similarity, which was derived from the recommendation algorithms of complex network theory and similar to the collaborative filtering method, to infer new targets for known drugs. Rakshit et al. [49] constructed an indication-drug–target network and used topological measures in the network to find non-Parkinsonian drugs for repositioning.

Classification and clustering algorithms have been adopted for generating novel drug–disease relationships. Hu and Agarwal [50] developed a large-scale drug–disease network and classified the diseases according to the Medical Subject Headings (MeSH). Scores of disease–disease pairs within the same category or between different categories were computed and used to identify new drug indications. Wu et al. [47] built a weighted heterogeneous network and clustered the network to identify modules and then assembled all possible drug–disease pairs from these modules. Tan et al. [22] constructed a drug–drug similarity network and used a popular clustering algorithm-MCODE to find neighbor nodes for drugs.

2.4. Data sources

Since different repositioning methods need different aspects of information of drugs and diseases, such as genetic, chemical, pharmacological, clinical, and protein information, multiple data sources have been used, including the most commonly used databases such as Drugs@FDA, Gene Expression Omnibus (GEO), Pharmacogenomics Knowledge Base (PharmGKB), Side Effect Resource (SIDER), DrugBank, PubMed, and Unified Medical Language System (UMLS) [2]. Table 2 summarizes the mostly used data sources and the approaches that were applied.

2.5. ADR detection

In this work, we used ADR as an intermediate for drug repositioning. As a result, the performance of the repositioning techniques is highly dependent on the quality of ADR extraction, in which to a great extend impacted by the data sources for ADR extraction.

The quality of data source has a great impact on the ADR detection and affects the drug repositioning results of our approach. Currently, data sources for obtaining ADR mainly include: (1) spontaneous reporting system, (2) electronic health records, (3) Administrative Health Databases, (4) Medical Literature, (5) online health communities (OHCs) [51].

FDA (Food and Drug Administration) Adverse Event Reporting System (FAERS) is the most important spontaneous reporting system as well as the primary data source for the study and identification of ADRs in United States. However, there are two major restrictions in FAERS. Users report ADRs spontaneously and voluntarily, which leads to a surprisingly low reporting rate because of the nature of passiveness, with a median of 6%. It usually takes FDA a long time to complete the whole process of collecting reports, investigating cases and releasing alerts. As a result, it is difficult to obtain the timely information about ADRs.

Compared with the spontaneous reporting system, data in the electronic health records is more timely and authoritative since the electronic health records are generated by the health professionals. However, due to the privacy and policy issues, as well as the difficulty in integrating electronic health databases from multiple resources, the ADR detection can only be restricted to a particular electronic health record system with a limited demographic coverage.

Administrative health databases provide high-level information about a patient's illness and medication history, but they lack the record information of treatment outcomes, which is required to assess a drug. Another type of missing data is the information on non-prescription medicine consumed by patients.

Medical literature has been used to generate useful resources for identifying ADRs, for the literature is easily accessible via Internet with or without subscriptions. Although the data quality is high, the information is available after a long delay of scientific research and publication process. The data is not as timely as spontaneous reporting system, electronic health records, and administrative health database.

In the recent years, the development of Web 2.0 not only breeds the various online social media sites like Facebook and Twitter, but also fosters online health communities (OHCs) such as MedHelp, PatientsLikeMe, and DailyStrength. OHCs generate a great deal of health-related contents and are more informative than some administrative databases. OHCs provide a space for patients and their caregivers to learn about an illness, seek and offer support, and connect with others in similar circumstances. OHCs have been growing in popularity across the world and provide a convenient way to exchange health information. It has been claimed that 80% of adults in US and 66% of adults in Europe seek online health advice [47]. 72% of Internet users said they searched online for health information in 2011 [52]. In addition, taking MedHelp for instance, it empowers over 12 million people each month to seek and offer healthcare information on the site. Since huge volumes of information were generated on OHCs, an increasing number of researches have been focused on OHCs, especially their impact on health consumers.

3. Drug repositioning based on heterogeneous network

The proposed drug repositioning system is comprised of four major modules, as shown in Fig. 1: (1) Dataset construction module, (2) Association mining module, (3) Heterogeneous network mining module, and (4) Drug repositioning module. The external data sources of this system include: (1) PharmGKB, a database providing disease–drug association data; (2) SIDER, a database providing drug–ADR association data; (3) CHV Wiki, a lexicon providing user expressions of ADRs; (4) MedHelp, an online health community providing abundant user contributed content, especially the discussion data of diseases, drugs, and ADRs. As a whole, the drug repositioning system takes in disease(s) as input and generates repositioning drug(s) as output, by referring to several external data sources.

(1) Dataset Construction Module:
Dataset Construction Module takes disease names as inputs and constructs a dataset of disease entities, drug entities and ADR signals as outputs, by referring to resources such as social media websites (e.g. MedHelp), pharmaceutical databases (e.g. PharmGKB, SIDER), and medical ontologies (e.g. CHV Wiki). Firstly, sub-module "Social Media Data Crawling" used the incoming disease names as queries to retrieve related drugs from PharmGKB. We used the identified drugs as queries to retrieve related ADRs from SIDER, and then used the obtained diseases, drugs,

Table 2
Data sources and drug-repositioning approaches.

Data sources		Computational approaches		
		Data mining	Machine learning	Network-based analysis
Genome	GEO		Napolitano et al. [43]	Hu and Agarwal [50]
	cMap		Zhang et al. [44]	Jiang et al. [25]
			Napolitano et al. [43]	Iorio et al. [26]
	MsigDB	Lee et al. [60]	Yang et al. [45]	Jiang et al. [25]
	KEGG			Cheng et al. [48]
				Wu et al. [47]
Molecule	PDB	Zheng et al. [23]		Kinnings et al. [24]
	HPRD			Li and Lu [21]
	BindingDB	Zheng et al. [23]		
	PubChem			Tan et al. [22]
	ChEMBL			Ng et al. [29]
Drug/phenome	PharmGKB	Li and Lu [39]		
		Zhu et al. [41]		
	SIDER			Yang and Agarwal [19]
	ClinicalTrials.gov	Li and Lu [39]		
	Drugs@FDA	Zhu et al. [41]		
	DrugBank	Campillos et al. [36]	Yang et al. [45]	Fukuoka et al. [14]
			Gottlieb et al. [42]	Tan et al. [22]
				Li and Lu [21]
				Rukov et al. [27]
				Rastegar-Mojarad et al. [30]
	PubMed	Andronis et al. [38]		
	UMLS	Campillos et al. [36]	Gottlieb et al. [42]	

and ADRs as keywords to crawl relative threads from MedHelp. Secondly, with the incoming MedHelp data, sub-module "ADR Detection" works on detecting the ADR signals by using a lexicon-based approach, and the disease and drug signals with consideration of their alternative names. In result, the outputs of Dataset Construction Module include two chunks: a set of social media threads (each thread is comprised of an original post and all the following comments), and a set of disease, drug, and ADR signals (which are used in the following Association Mining Module and Heterogeneous Network Mining Module).

(2) Association Mining Module:
Association Mining Module computes the associations and the weights of associations among disease, drug, and ADR signals in the dataset coming from Dataset Construction Module. There include three types of associations: disease–drug, drug–ADR, and disease–ADR associations. Using MedHelp threads as corpus, the computation of association weights is based on co-occurrence principle by exploiting association rule mining methods. In the end, Association Mining Module generates three matrices: disease–drug matrix, drug–ADR matrix, and disease–ADR matrix; each cell in the matrices describes the calculated association weights between two signals.

(3) Heterogeneous Network Mining Module:
Heterogeneous Network Mining Module has two sets of inputs: one

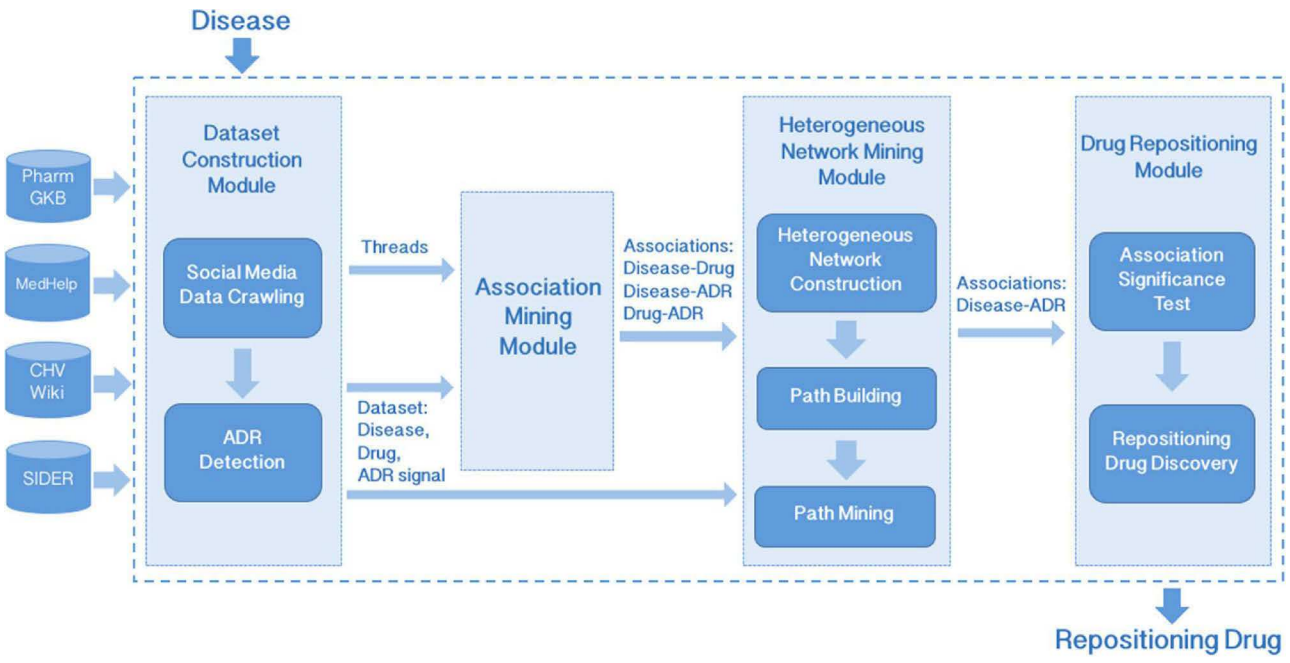

Fig. 1. Architecture of drug repositioning system.

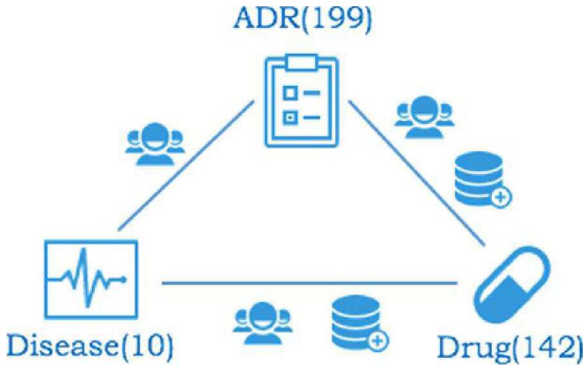

Fig. 2. Medical entities and associations in dataset.

is the dataset of disease, drug, and ADR signals, generated by Dataset Construction Module and serving as nodes in the heterogeneous network; the other one is a set of association matrices, generated by Association Mining Module and serving as links and link weights in the heterogeneous network. Firstly, sub-module "Heterogeneous Network Construction" constructs a heterogeneous healthcare network with the incoming data as nodes and links. Secondly, sub-module "Path Building" defines all the possible paths between each disease–ADR pair. Thirdly, sub-module "Path Mining" infers the final path value of each disease–ADR pair by exploiting different path mining approaches. Finally, Heterogeneous Network Mining Module outputs a list of disease–ADR pairs, along with the corresponding pair values that represent the inferred association strength.

(4) Drug Repositioning Module:
The task of Drug Repositioning Module is to identify repositioning drugs. Firstly, with the set of disease–ADR associations from Heterogeneous Network Mining Module as input, sub-module "Association Significance Test" extracts the significant associations from them, by using statistical analysis. Secondly, sub-module "Repositioning Drug Discovery" identifies the repositioning drugs based on the significant disease–ADR associations, by retrieving the drugs that show the ADR while have not yet been indicated for the disease from SIDER database. In the end, Drug Repositioning Module gives the repositioning drugs for targeted diseases.

4. Dataset construction module

4.1. Medical entities: drug, disease, ADR

The disease-based repositioning strategy enables researchers to focus on specific diseases and identify the drugs for repositioning.

Staring from these diseases of interest, we refer to PharmGKB (www.pharmgkb.org) to identify the corresponding drugs. Pharmacogenomics Knowledge Base (PharmGKB) is a publicly available knowledge resource that collects, curates, integrates and disseminates knowledge about the impact of human genetic variations on drug responses, containing genotypic and phenotypic information, as well as gene & drug & disease relationships. For instance, for each disease, it provides the alternated names, and the related drugs and genes; for each drug, it provides the properties (e.g. chemical structure, absorption, toxicity), pathways, and the related genes and diseases.

We refer to SIDER database (sideeffects.embl.de/) to obtain the corresponding ADRs as potential ADR candidates for these drugs. Side Effect Resource (SIDER) encompasses information about marketed drugs and their recorded ADRs, with data extracted from public documents such as MedSafe and FDA, and package inserts. The database contains 1430 drugs, 5868 ADRs and 139,756 drug–ADR pairs, and 39.9% pairs provide ADR frequency information.

Using the disease and drug name as query terms, we implemented an automatic web crawler to obtain all the related posts and comments from social media websites.

4.2. Medical entities extraction from social media data

When detecting signals of a disease in social media data, we utilized all the alternate names suggested in PharmGKB as well as the abbreviation (e.g. OCD for Obsessive-Compulsive Disorder), for example, the terms used to detect Parkinson included "parkinson" "parkinson disease" and "parkinson's disease". When detecting the signals of a drug, we utilized the terms included in PharmGKB and UMLS.

The expressions of diseases and drugs in social media and pharmaceutical databases are mostly similar and consistent, while the vocabularies of ADRs are quite different, because consumers use diverse and various expressions to describe the concepts and their adverse reactions [53]. Therefore, standard medical lexicons used by professionals like UMLS are not applicable in analyzing health consumer contributed content. To deal with this problem, we resorted to Consumer Health Vocabulary (CHV) Wiki to build up our ADR lexicon. CHV links everyday health-related words to professional terms or jargon, and the goal is to bridge the communication gap between consumers and healthcare professionals [54]. It provides a list of preferred names of ADRs and the corresponding consumer contributed expressions to each of them, for example, "anorexia" is a professional expression of ADR, CHV Wiki extends it to "appetite lost" "appetite loss" "appetite lack" "no appetite" and several other common expressions of health consumers. In our study, we used all the expressions suggested by CHV Wiki to detect ADR signals in user-generated information.

4.3. Associations between medical entities

There were three types of medical entities in the dataset: disease, drug and ADR, and there are three types of associations between these entities: disease–drug, drug–ADR and disease–ADR, which could be obtained from pharmaceutical databases or social media data, as shown in Fig. 2. Specifically, the associations of disease–drug and drug–ADR associations are accessible from PharmGKB and SIDER respectively. These associations can also be extracted from social media, which will be discussed in Section 5. However, the association of disease–ADR is not embodied in any pharmaceutical databases. We resort to social media for mining the disease–ADR association. Section 5 introduces the method of association mining from social media data in detail.

5. Association mining module

Association mining module deals with the extraction of disease–drug, drug–ADR and disease–ADR associations from social media, as well as the measurement of strength of the associations. When extracting these associations from the unstructured social media data, analysis granularity should be chosen empirically according to the research question to be answered. In this study, we use a thread as an analysis unit, which contains a post and the following comments, because the thread is composed of all discussions on a particular issue raised in the original post. A post or comment can also be considered as an analysis unit, but they are usually very short and the user may jump into his/her point without describing the concerned issue. For example, a comment could look like "I have just got similar reactions…" without mentioning the ADRs raised in the post. As the result, a post or comment is too small as analysis unit to extract ADR signals compared with the thread.

In social media data, if two entities are mentioned together frequently, they are deemed to be strongly associated [55], therefore, a number of studies have applied co-occurrence analysis to evaluate the strength of associations. To find the strong associations between the medical entity pairs (drug–disease, drug–ADR, disease–ADR), we adopted association rule mining here, because association rule mining

is one of the most important techniques in data mining to extract interesting correlations and frequent patterns among sets of items. Specifically, we followed the principle of Apriori algorithm in association rule mining by firstly using a breadth-first search strategy to count the *support* of itemsets and then using *lift* to determine the frequent itemsets.

In association rule mining, let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of items and let $T = \{T_1, T_2, \ldots, T_n\}$ be a set of transactions, where each transaction is a subset of items such that $T_i \subseteq I$. An itemset that contains k items is a k-itemset; the occurrence frequency of an itemset is the number of transactions that contain the itemset. The association rule is an implition of form $A \Rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \varnothing$, which is deemed as an itemset. In our case, I denotes the whole set that contains diseases (D), drugs (R), and ADRs; T denotes the dataset of all threads and each thread represents a transaction; there are both 1-itemset (e.g. {D}, {R}, {ADR}) and 2-itemset (e.g. {D, R}, {D, ADR}, {R, ADR}) involved in our calculation. Our goal is to mine and evaluate the associations presented in 2-itemset, in other words, mining the rules in the form of $D \Rightarrow R$, $R \Rightarrow ADR$, $D \Rightarrow ADR$.

Support is a common indicator used in association rule mining, defined as the percentage of transactions that contain 1-itemset or 2-itemset, for instance:

$$support\,(ADR) = \frac{count\,(ADR)}{total\ count}$$

$$support\,(R \Rightarrow ADR) = \frac{count\,(R \cup ADR)}{total\ count}$$

in which, $count\,(ADR)$ is the number of threads that contain target *ADR*; $count\,(R \cup ADR)$ is the number of threads that contain both drug *R* and *ADR*; *total count* is the total number of threads.

Nevertheless, for the 2-itemset, *support* is appropriate only when the co-occurrence frequency of the items is high. However, when consumers mention a drug, they might discuss different aspects of drugs, so that threads that are related to ADR only occupy a small portion in all the threads. To address this problem, another indicator *lift* is often used. *Lift* is a measure based on probability and reflects the division of the actual probability and theoretical probability. For instance, when measuring the strength of rule $R \Rightarrow ADR$, lift not only takes account of $support\,(R \cup ADR)$ but also the correlation between 1-itemset *R* and 1-itemset *ADR*, by calculating the ratio of the proportion of threads containing both *R* and *ADR* above those expected if *R* and *ADR* are independent of each other. The calculation of $lift\,(D \Rightarrow R)$, $lift\,(D \Rightarrow ADR)$, $lift\,(R \Rightarrow ADR)$ are shown in the following formulas:

$$lift\,(D \Rightarrow R) = \frac{support\,(D \cup R)}{support\,(D) \times support\,(R)}$$

$$lift\,(D \Rightarrow ADR) = \frac{support\,(D \cup ADR)}{support\,(D) \times support\,(ADR)}$$

$$lift\,(R \Rightarrow ADR) = \frac{support\,(R \cup ADR)}{support\,(R) \times support\,(ADR)}$$

The outputs of this module are extracted disease–drug, disease–ADR and drug–ADR associations, the strength of which are indicated by *lift*. Besides, the higher of *lift* value represents the stronger the association between the two items.

6. Heterogeneous network mining module

Network analysis is an important approach applied in drug repositioning, with the great capability of revealing the connections as well as the underlying mechanisms and interactions between multiple medical entities. Most of current studies based on network science views networks as homogeneous, where nodes are objects of the same type and relationships among nodes are of the same type, such as only authors involved in citation network and only users in social network. However, most real world networks such as healthcare information networks are heterogeneous, where nodes and relationships are of different types, for example, in a healthcare information network, nodes could be diseases, drugs, ADRs, genes, and proteins, and the links between these nodes are also different. Analysis based on homogeneous network may miss important semantic and schema-level information, while heterogeneous network can present more essential, accurate and complete features of the real-world network, thus unveiling the underlying knowledge and patterns. However, there is no adequate research in applying heterogeneous network techniques in drug repositioning yet. Therefore, we proposed a heterogeneous network based method to represent associations between diseases, drugs and ADRs, and to explore novel connections.

6.1. Heterogeneous healthcare network definition

A heterogeneous network is defined as a graph consisting of nodes connected by links, with at least two types of nodes and at least two types of links [56]. Let $N = \{n_1, n_2, \ldots, n_k\}$ be a set of nodes and $L = \{l_1, l_2, \ldots, l_m\}$ be a set of links, then $G = (N, L)$ denotes the graph. In the graph G, each node $n_i \in N$ belongs to a particular type from γ; each link $l_i \in L$ belongs to a particular type from τ, and $|\gamma| > 1$ or $|\tau| > 1$, and can be directional or non-directional. Then $M_G = (\gamma, \tau)$ denotes the node types γ and link types τ in graph G.

6.2. Construction of heterogeneous healthcare network

In our heterogeneous healthcare network $G = (N, L)$, there are three types of nodes: drug(R), disease(D) and ADR, and three types of links: drug–ADR, disease–ADR and drug–disease, that is, $\gamma = \{R, D, ADR\}$ and $\tau = \{L_{R-ADR}, L_{D-ADR}, L_{D-R}\}$. The relation between *D* and *R* is *treat* $(R \rightarrow D)$ or *be treated* $(D \rightarrow R)$; the relation between *R* and *ADR* is *cause* $(R \rightarrow ADR)$ or *be caused* $(ADR \rightarrow R)$; the relation between *D* and *ADR* is not a directly causal relationship, but one connected by some underlying MOAs.

The proposed network is non-directional weighted heterogeneous network. The weights of links are uniform and the strength of associations between nodes is not considered in a non-weighted heterogeneous healthcare network. However, in our proposed network, the weights are determined by the association strength between two nodes as discussed in Section 5. Our proposed network is not directional because we are not considering the causality of relations between different types of node. There is no explicit causality information provided when we construct the heterogeneous healthcare network from social media data and pharmaceutical databases. It is too complicated to accurately determine the causal relations between nodes even with the assistant of natural language processing (NLP) techniques from the user contributed content [57]. Meanwhile, non-directional relations here are capable of revealing the associations between drugs, diseases and ADRs for repositioning use. Fig. 3 presents the non-directional heterogeneous healthcare network model.

6.3. Path mining

The rationale for ADR-based repositioning strategy is that ADRs and indications both convert the physiological or behavioral consequences to the treatment, and if drugs treating a disease share the same ADR, there might be some underlying MOA linking the disease with the ADR, thus ADR could be seen as a phenotypic "biomarker" of the disease. Therefore, the goal in heterogeneous network mining is to discover the associations between disease and ADR. Under a strong disease–ADR association, the drugs having the ADR but not indicated for the disease could be evaluated as a repositioning candidate.

In view of the heterogeneous network model, there are two pathways to connect *D* and *ADR*: the first path is a direct link of *D–ADR*, or

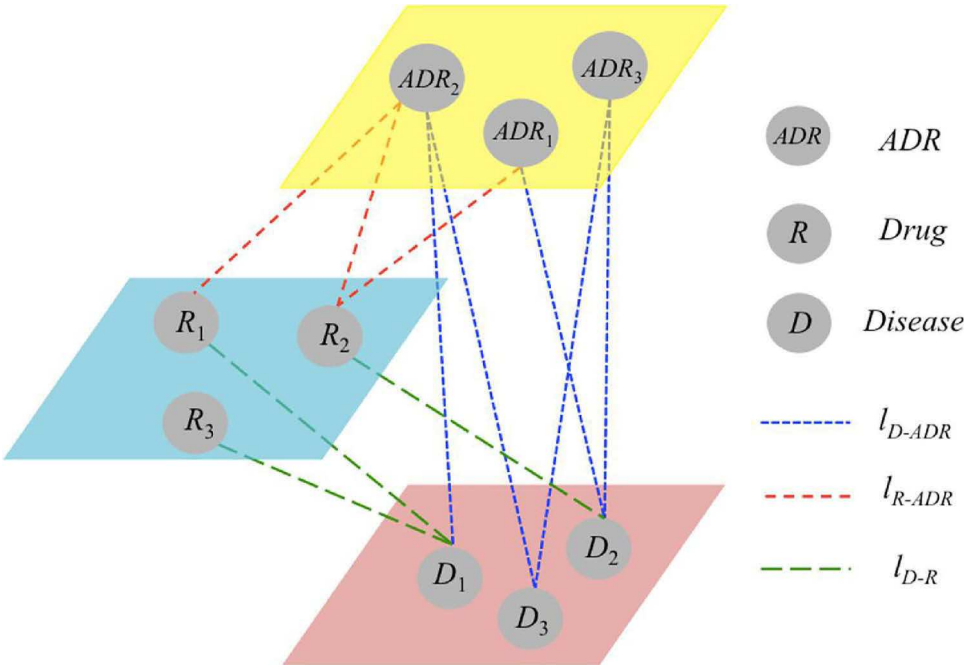

Fig. 3. Heterogeneous healthcare network model.

l_{D-ADR}; the second is a path of *D–R–ADR*, or l_{D-R}–l_{R-ADR}.

(1) *Path(D–ADR)*
Since there are no existing databases revealing disease–ADR associations, we resort to social media to obtain *D–ADR* information. That means such associations are mined from unstructured user-contributed content directly (as shown in Fig. 4) and the strength of D–ADR association, $S_{D-ADR}(l_{D-ADR})$, is measured by *lift* we computed in Section 5:

Fig. 4. *Path(D–ADR)*.

$$S_{D-ADR} = \sum_{\forall P} l_{D-ADR} = \sum_{\forall P} lift(D \Rightarrow ADR)$$

(2) *Path(D–R–ADR)*
ADR represents the harmful and unpleasant reactions of medicine use, and hence, the ADRs that are associated with a disease are highly influenced by the drugs that are treating the disease. In other words, disease indications and ADR are not connected directly with each other, while both of them are direct responses of drugs. Therefore, it is reasonable to use drugs as the bridge between disease and ADR.

Considering the variability of qualities of multiple health data sources, we applied the proposed mining method on different data sources to explore if the results of drug repositioning would be influenced by data source and if social media data would gain a better performance. For instance, since adverse reactions of drugs are substantially under-reported in most medical systems and databases, it might lead to the incompleteness of drug–ADR information, therefore, we obtained R–ADR relationships from both social media and pharmaceutical databases. Fig. 5 demonstrates four different combinations of data sources in applying *Path*(*D–R–ADR*):

The strength of a path, Path(D–R–ADR), is measured by both the weights of l_{D-R} and l_{R-ADR}. When the weight is computed from MedHelp, $lift(D \Rightarrow R)$ and $lift(R \Rightarrow ADR)$ are used. When the weight is computed from PharmGKB or SIDER, the weight is binary depending on if the association is indicated in the databases. The strength of *D–ADR* association via *Path*(*D–R–ADR*) is computed by the following formulation:

$$S_{D-R-ADR} = \sum_{\forall P} l_{D-R} \times l_{R-ADR}$$

where *P* denotes all the possible paths between *D* and *ADR* with *R* as intermediary.

InPharmSIDERandPharmMed,

$$l_{D-R} = \begin{cases} 1 & ifD - Risfoundin\ PharmGKB \\ 0 & ifD - Risnotfoundin\ PharmGKB \end{cases}$$

InMedSIDERandMedMed,

$$l_{D-R} = lift(D \Rightarrow R)$$

InPharmSIDERandMedSIDER,

$$l_{R-ADR} = \begin{cases} 1 & ifR - ADRisfoundinSIDER \\ 0 & ifR - ADRisnotfoundinSIDER \end{cases}$$

In PharmMedandMedMed,

$$l_{R-ADR} = lift(R \Rightarrow ADR)$$

7. Drug repositioning module

A strong association between a disease and an ADR implies an underlying MOA between the disease and the ADR, then the drugs connected with the ADR could be evaluated as a repositioning candidate for this disease. Based on the mined D–ADR associations from Section 6, this module is targeted at: firstly, assessing and extracting the significant D–ADR associations from all associations; secondly, identifying

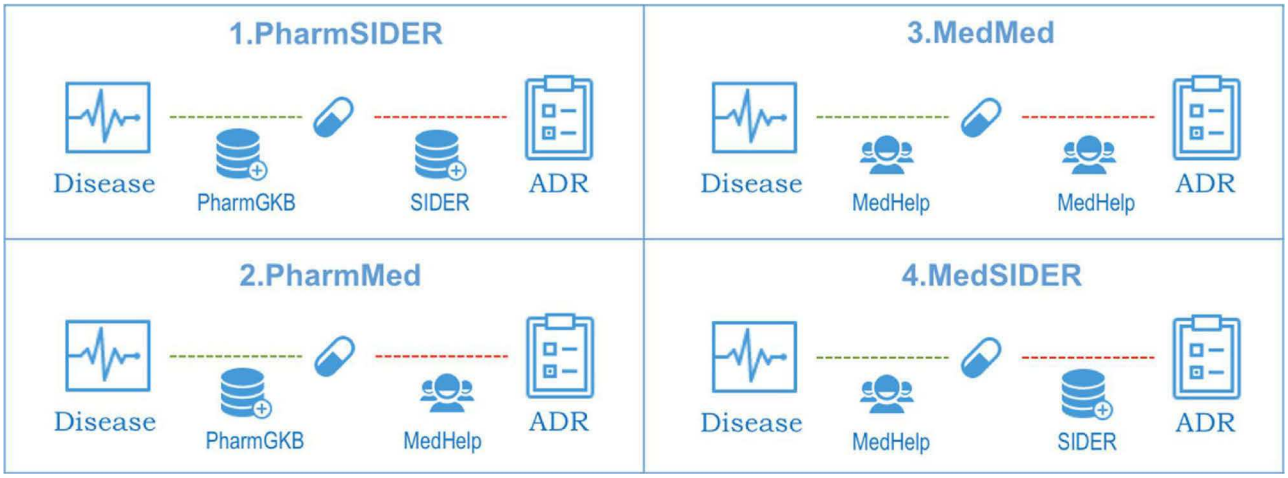

Fig. 5. Different combinations of data sources in *Path*(*D–R–ADR*).

repositioning drugs for each disease.

7.1. Significance test on D–ADR associations

The heterogeneous network mining module generates a large set of D–ADR associations, while not all of the associations are persuasive to be utilized for drug repositioning analysis, especially those with weak association strength. For example, Yang and Agarwal [19] generated a confusion matrix for each D–ADR pair and used Matthews correlation coefficient (MCC), sensitivity and specificity to evaluate the strength of the association. Here we utilized significance test in statistics, specifically, one-sample *T*-test, to differentiate the significant associations with the insignificant ones. For each ADR, we computed the strength of the associations between the ADR and all the diseases. We conducted the one-sample *T*-test, with the null hypothesis that there was no significant difference between the value and the sample mean. For example, we created a sample for ADR "abnormal sensation", which included 10 values of $S_{D-R-ADR}(l_{D-ADR})$ obtained based on MedSIDER. We applied the *T*-test and computed the *t*-value. In hypothesis testing, we compared the calculated *t*-value and the table value, and rejected the null hypothesis for the three associations between abnormal sensation and transplantation, Parkinson, and obsessive-compulsive disorder, as shown in Table 3.

7.2. Identification of repositioning drugs

For each significant disease–ADR association, the goal is identifying the drugs with this ADR while not yet indicated for the corresponding disease. The main procedures include: (1) SIDER provides a list of the drugs for each ADR and the corresponding frequency. We identified the drugs with the ADR frequency labeled with "very common", "common", or the percentage higher than 10%; (2) referring to PharmGKB, we removed the drugs that have already been indicated for the disease. The remaining drugs in the discovery are considered as repositioning drugs for the disease.

Based on each disease–ADR association, we generated a list of repositioning drugs for this disease. That is, if a disease was significantly associated with *k* ADRs, there were *k* lists of repositioning drugs for this disease. If a drug was suggested in more than *k*/2 of the lists, it would be deemed as a repositioning drug, which reinforced the reliability of the underlying MOA between a disease and an ADR.

Table 3
Significance test results on the sample of abnormal-sensation.

ADR	abnormal sensation (AS)	$S_{D\text{-}R\text{-}ADR}$ ($l_{D\text{-}ADR}$)	*t*-value (2.262)
Disease	**transplantation**	1.42E-03	−5.187
	parkinson	1.14E-03	−3.381
	obsessive-compulsive disorder	1.09E-03	−3.052
	hypercholesterolemia	9.44E-04	−2.129
	cystitis	4.51E-04	1.032
	pharyngitis	4.37E-04	1.122
	kidney failure	3.20E-04	1.872
	gastroesophageal reflux	1.77E-04	2.789
	myalgia	8.73E-05	3.364
	glaucoma	5.29E-05	3.584

2.262 is the critical value in t-value table where degree of freedom = 9 and confidence level = 95%.
Bold fonts represent there are significant associations between that disease and the ADR-AS.

8. Experiment

In this study, we used MedHelp (www.medhelp.org), one of the pioneers in online health communities, as the data source for obtaining health consumer-contributed contents and targeted at 10 common diseases: Cystitis, Gastroesophageal Reflux, Glaucoma, Hypercholesterolemia, Kidney Failure, Myalgia, Obsessive-Compulsive Disorder, Parkinson, Pharyngitis, and Transplantation, for which we collected data from both medical databases and social media websites.

We collected more than 41,000 threads from MedHelp by using the 10 diseases and the corresponding 142 drugs suggested by PharmGKB as query terms, where each thread was composed of a post and all the following comments. After the extraction of disease, drug, and ADR entities from user posts, we constructed a heterogeneous network with 351 nodes.

8.1. Experiment results

Based on the constructed heterogeneous network, we applied the path-mining methods and completed drug repositioning. Table 4 shows the number of repositioning drugs we discovered for each disease by applying *Path*(*D–ADR*) approach, and *Path*(*D–R–ADR*) approaches with four combinations of different data sources.

In respect of the number of identified repositioning drugs, *Path*(*D–ADR*) achieved the worst comparing with all of the four approaches in *Path*(*D–R–ADR*). *Path*(*D–ADR*) detected the D–ADR associations by mining such associations directly from unstructured social media data, while *Path*(*D–R–ADR*) inferred these associations based on the features of heterogeneous network, which demonstrated that network-based method was an effective way to uncover the underlying associations between diseases and ADRs especially when their associations were not reflected explicitly in user-contributed content.

Table 4
Number of discovered repositioning drugs.

Disease	Number of repositioning drugs				
	Path(*D–ADR*)	PharmSIDER	PharmMed	MedMed	MedSIDER
Cystitis	0	0	11	18	16
Gastroesophageal reflux	7	15	15	10	10
Glaucoma	0	0	2	14	3
Hypercholesterolemia	0	0	7	16	13
Kidney failure	4	0	74	5	5
Obsessive–compulsive disorder	15	142	147	25	17
Parkinson	25	67	60	0	17
Pharyngitis	0	0	21	11	12
Transplantation	21	118	142	31	14
Total	72	342	479	130	107

Comparing the four approaches in *Path*(*D–R–ADR*), PharmMed, where the association of D–R was obtained from PharmGKB and the association of R–ADR was obtained from MedHelp, achieved the highest number of repositioning drugs (479). On the other hand, MedSIDER, where the association of D–R was obtained from MedHelp and the association of R–ADR was obtained from SIDER, discovered the least number of repositioning drugs (107).

When disease–drug relationships derived from PharmGKB, the repositioning drugs identified by PharmMed were more than PharmSIDER, especially for cystitis, glaucoma, hypercholesterolemia and kidney failure, where PharmSIDER did not discover any repositioning drugs but PharmMed did. The possible reason was that some drug–ADR associations were not collected in SIDER database, but they could be determined from the timely health consumer contributed data. In other words, social media data appeared to be a better data source for obtaining drug–ADR information than SIDER. In addition, when disease–drug relationships coming from MedHelp, MedMed also performed better than MedSIDER.

When drug–ADR relationships derived from SIDER, the repositioning drugs identified by PharmSIDER were substantially more than MedSIDER, which meant PharmGKB included more complete disease–drug information than that was mined from MedHelp. Besides, when drug–ADR relationships coming from MedHelp, PharmMed also achieved a better result than MedMed.

The above observations also explained why PharmMed performed the best while MedSIDER performed the worst, because PharmMed utilized the best data sources for both disease–drug and drug–ADR associations, meanwhile, MedSIDER utilized the worst for both associations.

Most of the drug repositioning results extracted by computing methodologies are not confirmed truth before clinical trials are conducted. Instead, they suggest a possibility for further drug development. As a result, evaluating the performance of these computing methodologies is not a simple task. The common evaluation of repositioning drugs are either computational or experimental [30]. The computational assessments are usually based on the co-occurrence of drug and disease terms in biomedical literature and clinical trials. The experimental assessments are based on in silico or in vitro experiments. In computational assessment, finding evidence in medical articles is one of the evaluation methods that have been adopted by a majority of the previous studies [22,23,39,58]. We resorted to medical literature for result assessment by finding evidence in articles published in PubMed. As shown in Table 5, the results showed that 47% of repositioning-drug → disease associations discovered by *Path*(*D–ADR*) had at least one article in publication type of "Clinical Trial", with the disease as a major subject heading for that article and the drug name mentioned in the title or abstract, followed by 36% of the repositioning-drug → disease associations discovered MedMed and MedSIDER, and 31% by PharmMed.

As shown in Table 5, by considering the number of identified repositioning drugs with evidence out of the number of identified repositioning drugs by each approach, *Path*(*D–ADR*) achieved the highest precision (47%), followed by MedMed (36%) and MedSIDER (36%) approaches in *Path*(*D–R–ADR*). However, with regard to the absolute value rather than the relative value (the number of identified repositioning drugs with evidence), PharmMed discovered 150 drugs, which was 4.41 times more than *Path*(*D–ADR*) (34 drugs) and 3.13 times more than MedMed (48 drugs). Although PharmSIDER achieved relatively lower precision, it discovered the second most of repositioning drugs with evidence (90 drugs). The precision measured one aspect of the experimental results, regarding to the percentage of identified repositioning drugs that were supported by literature. When the precision of an approach was not as high, it did not necessarily mean the approach was not good at identifying repositioning drugs. There was always a trade off between the precision and the number of identified repositioning drugs with evidence. When an approach was capable of identify more repositioning drugs, it might also produce more false positives (identified repositioning drugs without evidence) and therefore lower the precision. An approach might achieve a high precision but might also miss many true positives (identified repositioning drugs with evidence). In this case, PharmMed identified a lot more repositioning drugs with evidence but it achieved a lower precision of 31% compared to the best precision of 47% achieved by *Path*(*D–ADR*). It is also possible that there will be more studies in the future that can provide evidence on the identified repositioning drugs that are not supported yet. That means some of the false positives may indeed have evidences in the future studies.

By taking the drugs identified by all of the five approaches, we had a total of 750 repositioning drugs. Among these 750 drugs, 171 drugs were supported by evidence that we considered as positives. The other 579 drugs were not supported by evidence and therefore considered as negatives. We used this set of drugs to measure the sensitivity (True Positives/Positives) and specificity (True Negatives/Negatives) achieved by the five approaches.

Table 6 presents the results of sensitivity and specificity on each approach. PharmMed achieved the highest sensitivity of 0.877, representing that among all the repositioning drugs with evidence it identified 87.7% of them, whereas the sensitivities of the other methods were much lower. On the other hand, *Path*(*D–ADR*), MedMed and MedSIDER achieved high specificity of 0.934, 0.858, and 0.883 respectively, which means they were good at identifying the true negatives with the cost of missing many true positives.

We also found that PharmSIDER and PharmMed outperformed MedMed and MedSIDER substantially in sensitivity, but MedMed and MedSIDER outperformed PharmSIDER and PharmMed substantially in specificity. That means measuring the associations of disease and drug with PharmGKB in *Path*(*D–R–ADR*) approach can achieve higher sensitivity but measuring the associations of disease and drug with

Table 5
Number of repositioning drugs supported by literature.

Disease	Number of repositioning drugs supported by literature				
	Path(*D–ADR*)	PharmSIDER	PharmMed	MedMed	MedSIDER
Cystitis	0	0	55% (6/11)	33% (6/18)	37% (6/16)
Gastroesophageal reflux	57% (4/7)	47% (7/15)	47% (7/15)	40% (4/10)	30% (3/10)
Glaucoma	0	0	100% (2/2)	29% (4/14)	0 (0/3)
Hypercholesterolemia	0	0	43% (3/7)	31% (5/16)	38% (5/13)
Kidney failure	75% (3/4)	0	47% (35/74)	60% (3/5)	60% (3/5)
Obsessive–compulsive disorder	60% (9/15)	26% (37/142)	27% (39/147)	52% (13/25)	29% (5/17)
Parkinson	40% (10/25)	28% (19/67)	32% (19/60)	0	47% (8/17)
Pharyngitis	0	0	33% (7/21)	27% (3/11)	25% (3/12)
Transplantation	38% (8/21)	23% (27/118)	23% (32/142)	39% (10/31)	43% (6/14)
Total	47% (34/72)	26% (90/342)	31% (150/479)	36% (48/130)	36% (39/107)

Table 6
Sensitivity and specificity of each method.

Method	TP	FN	FP	TN	Sensitivity	Specificity	*F*1
Path(*D–ADR*)	34	137	38	541	0.199	0.934	0.280
PharmSIDER	90	81	252	327	0.526	0.565	0.351
PharmMed	150	21	329	250	0.877	0.432	0.461
MedMed	48	123	82	497	0.281	0.858	0.319
MedSIDER	39	132	68	511	0.228	0.883	0.281

MedHelp in *Path*(*D–R–ADR*) can achieve higher specificity. When we compared PharmSIDER and PharmMed, PharmMed achieved substantially higher sensitivity than PharmSIDER. That means combining the association measurement from PharmGKB and the association measurement from MedHelp produce the highest sensitivity. On the other, MedSIDER achieved slightly higher specificity than MedMed but not substantially. In general, in order to identify the most number of true positives out of the identified repositioning drugs that are supported by evidence, the best performance can be achieved by using the heterogeneous network approach with the extracted associations of diseases and drugs from PharmGKB and the extracted associations of drugs and ADRs from MedHelp. It reflects that social media data is not as reliable as pharmaceutical database in determining the disease–drug associations; however, the social media data is more useful in determining the drug–ADR associations. By selecting the appropriate resources in building the heterogeneous networks for mining, the best performance can be achieved rather than relying on only one type of resource in constructing the heterogeneous network.

In one confusion matrix, the total number of instances refers to all the repositioning drugs that are identified by the five approaches, while the true-positives only refer to drugs that are identified by the current approach, which explains why *F*1 scores of the approaches are not high in Table 6. Within the five approaches, PharmMed achieved the highest *F*1 of 0.461, showing its best performance on the whole.

9. Discussion

Most of the drug repositioning studies have relied on single data sources such as pharmaceutical database and medical record, while few of them integrated multiple medical data sources especially data provided by health consumers. For example, there are works that used data sources such as PharmGKB and SIDER, but there is no existing work that has integrated social media data with the previous data sources. In this work, we utilized both social media data (e.g. MedHelp) and pharmaceutical databases (e.g. PharmGKB and SIDER) to detect repositioning drugs. The integration of multiple data sources allows us to compare the impact of different data sources and different integration ways, and to explore whether using health consumer-contributed data could improve the drug repositioning results. However, since there are no similar studies that integrated social media data for phenotypic information-based drug repositioning, we are unable to compare our results with the other existing approaches. We compare the four integrations as illustrated in Fig. 5. The four integrations use different combinations of data sources, PharmGKB and SIDER (PharmSIDER), PharmGKB and MedHelp (PharmMed), MedHelp and MedHelp (MedMed), and MedHelp and SIDER (MedSIDER).

We analyzed the different performance of different integration approaches in Section 8. The experiment results in Table 4 showed that integrating PharmGKB with MedHelp achieved the best performance for drug repositioning. Furtherly, in order to analyze how each of the data source effects the repositioning results, we did comparisons using controlling variable method and found that PharmGKB contributes the most information on disease–drug relationship and MedHelp contributes the most on drug–ADR relationship. The reason why consumer-contributed data perform better than the databases is that ADRs are substantially under-reported in most medical systems and databases especially in FAERS (FDA Adverse Effect Reporting System). Tables 5 and 6 compare the four integration ways regarding to how the identified repositioning drugs are supported by literature evidence. PharmMed achieved the best performance in terms of the number of repositioning drugs that are supported by literature evidence as well as the sensitivity and *F*1. PharmMed seems to have a lower specificity. However, it is possible that some of the true negatives do not have evidences from the literature because no studies have been done yet.

The way we evaluated our experiment results was based on the consultation with the medical experts. When a medication is recommended for repositioning to the medical experts (e.g. pharmacists and physicians), the medical experts would resort to the medical literature to look for evidence that supports the repositioning (e.g. article, publication, clinical reports or studies). Given this practice traditionally conducted by the medical experts, we conducted our experiment exhaustively which is conventionally done in scientific evaluation of drug repositioning papers. We then presented the results to our medical experts to guarantee the accuracy of evaluation results and the descriptions. The medical experts we have consulted include one pharmacist with over twenty years of experience, one physician with over ten years of experience, and one research scientist working in a pharmaceutical company. They concluded that the repositioning drugs recommended by the proposed algorithms are very helpful for them to narrow down the drugs that can be potentially used for the suggested indication. Some of the recommended repositioning drugs were indeed used as off-label drugs in practice. The experiment also helped them to understand the limitations of the data sources as well as the strength of integrating multiple sources of data in investigation.

This evaluation method has been used in many previous studies as well [19,21–23,39,58,61]. For example, Tan et al. [22] validated their results by searching for current clinical trials. Wang et al. [58] evaluated their predictions by searching for the published literatures. Li and

Lu [21] evaluated their results by searching the disease-repositioning drug pairs in ClinicalTrials.gov and scientific abstracts in PubMed to look for evidences that support the repositioning drug might be indicated for the disease. Li and Lu [61] evaluated their results by searching evidence in both clinical trials and literature and found about 1/3 predictions could be found in PubMed and only a small percentage could be found in ClinicalTrials.gov.

The results of drug repositioning are usually considered as suggestions, predictions, or recommendations, not the results that can be approved for patients immediately. Chiang and Butte [34] even declared that they were unable to examine the results that are only "potential" suggestions rather than FDA-approved drugs, and they saw their results as suggestions for further in vitro and in vivo tests. The main contribution of our work and many repositioning studies is suggesting novel drug uses for pharmaceutical companies and medical associations to conduct further in vitro and in vivo tests, effectiveness and risk evaluation, and clinical trials. Besides, our medical experts also said they would like to see such findings and explore whether there are off-label use opportunities from the repositioning drugs.

10. Conclusion

By suggesting new therapeutic uses for approved drugs, drug repositioning plays a significant role in dealing with the problems of high risks, costly process and long period in drug discovery and development. In this paper, we connected drugs with potential indications for drug repositioning through Adverse Drug Reactions (ADRs). ADR-based repositioning approach were shown to be capable of profiling drug related phenotypic information and can subsequently helped in discovering new therapeutic uses for drugs, and were proved to display better performance than chemical features, biological features (protein features) or their combination. In this work, we developed the path-mining approaches of heterogeneous network mining to explore the associations between drugs, diseases and ADRs. In addition, due to the fact that ADRs were substantially under-reported in most medical systems and databases, which might lead to the insufficiency of such data sources, we utilized both social media data and pharmaceutical databases to extract ADRs, as well as their associations with drugs and diseases, to explore if the results of drug repositioning would be influenced by data source and if social media data would have advantages over pharmaceutical databases in providing information about disease, drug, ADR, and their associations. We obtained data from PharmGKB and SIDER, as well as an online health community – MedHelp, and developed two path-mining methods *Path*(*D–ADR*) *and Path*(*D–R–ADR*) based on the proposed heterogeneous network model. With *Path* (*D–R–ADR*) method, we applied four implementations with data collecting from different data sources, namely PharmMed, PharmSIDER, MedMed, and MedSIDER. To evaluate the performance of different approaches, we resorted to medical literature to identify evidence in articles published by PubMed, and we employed precision, sensitivity and specificity as measurements to evaluate their performance on the identification of repositioning drugs.

Path(*D–ADR*) achieved the highest precision (repositioning drugs with evidence/identified repositioning drugs), while PharmMed discovered the most number of repositioning drugs with evidence. In terms of sensitivity and specificity, PharmMed achieved the highest sensitivity (repositioning drugs with evidence identified by this approach/all repositioning drugs with evidence), showing its capability in identifying true positives, meanwhile, *Path*(*D–ADR*), MedMed and MedSIDER achieved high specificity, showing their advantage in identifying true negatives.

In the experiment, we found that all the *Path*(*D–R–ADR*) approaches identify more repositioning drugs than *Path*(*D–ADR*). *Path*(*D–ADR*) mined the associations between diseases and ADRs directly from the unstructured social media data while *Path*(*D–R–ADR*) utilized the features of heterogeneous network to discover the underlying associations. When we compared the different approaches in *Path*(*D–R–ADR*), we found that social media reveals more ADR related information than pharmaceutical databases, while pharmaceutical database outperformed social media in identifying the disease–drug associations. By incorporating PharmGKB and MedHelp in the heterogeneous network mining approach, it achieved the best sensitivity but relatively worse specificity. In the future, we shall further explore other resources and integration in adverse drug reaction detection and drug repositioning.

Funding

This work was supported in part by the National Science Foundation under the Grants NSF-1741306, NSF-1650531, and NSF-1443019.

Conflict of interest

No potential conflict of interest was reported by the authors.

Acknowledgements

This work was supported in part by the National Science Foundation under the Grant NSF-1741306, IIS-1650531, and DIBBS-1443019. Any opinions, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendix A

The source code of this work is available at https://github.com/Missymeng/Drug_Repositioning. Unfortunately, the data cannot be made publicly available due to the privacy agreement.

References

[1] Gilbert J, Henske P, Singh A. Rebuilding big pharma's business model. In Vivo-New York Then Norwalk 2003;21(10):73–80.
[2] Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. Brief Bioinform 2015;17(1):2–12.
[3] DiMasi JA, Grabowski HG, Hansen RW. The cost of drug development. N Engl J Med 2015;372(20):1972.
[4] Jin G, Wong ST. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discov Today 2014;19(5):637–44.
[5] Graul AI, Cruces E, Stringer M. The year's new drugs & biologics, 2013: Part I. Drugs Today (Barc) 2014;50(1):51–100.
[6] Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 2004;3(8):673–83.
[7] Power A, Berger AC, Ginsburg GS. Genomics-enabled drug repositioning and repurposing: insights from an IOM Roundtable activity. JAMA 2014;311(20):2063–4.
[8] Pollack MH, Matthews J, Scott EL. Gabapentin as a potential treatment for anxiety disorders. Am J Psychiatry 1998;155(7):992–3.
[9] Padhy BM, Gupta YK. Drug repositioning: re-investigating existing drugs for new therapeutic indications. J Postgrad Med 2011;57(2):153.
[10] Elvidge S. Getting the drug repositioning genie out of the bottle. Life Science Leader; 2010.
[11] Jiao M, Liu G, Xue Y, Ding C. Computational drug repositioning for cancer therapeutics. Curr Top Med Chem 2015;15(8):767–75.
[12] Sardana D, Zhu C, Zhang M, Gudivada RC, Yang L, Jegga AG. Drug repositioning for orphan diseases. Brief Bioinform 2011;12(4):346–56.
[13] Li YY, Jones SJ. Drug repositioning for personalized medicine. Genome Med 2012;4(3):27.
[14] Fukuoka Y, Takei D, Ogawa H. A two-step drug repositioning method based on a protein–protein interaction network of genes shared by two diseases and the similarity of drugs. Bioinformation 2013;9(2):89–93.
[15] Dudley JT, Deshpande T, Butte AJ. Exploiting drug–disease relationships for computational drug repositioning. Brief Bioinform 2011;12(4):303–11.
[16] Bisgin H, Liu Z, Fang H, Kelly R, Xu X, Tong W. A phenome-guided drug repositioning through a latent variable model. BMC Bioinform 2014;15(1):267.
[17] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010;6(1):343.
[18] Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning. PLOS ONE 2014;9(2):e87864.
[19] Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. PLoS ONE 2011;6(12):e28025.
[20] Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature 2009;462(7270):175–81.

[21] Li J, Lu Z. A new method for computational drug repositioning using drug pairwise similarity. 2012 IEEE international conference on Bioinformatics and Biomedicine (BIBM). October 2012. p. 1–4.
[22] Tan F, Yang R, Xu X, Chen X, Wang Y, Ma H, et al. Drug repositioning by applying 'expression profiles' generated by integrating chemical structure similarity and gene semantic similarity. Mol BioSyst 2014;10(5):1126–38.
[23] Zheng C, Guo Z, Huang C, Wu Z, Li Y, Chen X, et al. Large-scale direct targeting for drug repositioning and discovery. Sci Rep 2015;5.
[24] Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, Bourne PE. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. PLoS Comput Biol 2009;5(7):e1000423.
[25] Jiang W, Chen X, Liao M, Li W, Lian B, Wang L, et al. Identification of links between small molecules and miRNAs in human cancers based on transcriptional responses. Sci Rep 2012;2.
[26] Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J. Transcriptional data: a new gateway to drug repositioning? Drug Discov Today 2013;18(7):350–7.
[27] Rukov JL, Wilentzik R, Jaffe I, Vinther J, Shomron N. Pharmaco-miR: linking microRNAs and drug effects. Brief Bioinform 2013. bbs082.
[28] Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 2014;506(7488):376–81.
[29] Ng Clara, Hauptman Ruth, Zhang Yinliang, Bourne PE, Xie L. Anti-infectious drug repurposing using an integrated chemical genomics and structural systems biology approach. Pac Symp Biocomput, vol. 19 2014:136–47.
[30] Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. Nat Biotechnol 2015;33(4):342–5.
[31] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006;313(5795):1929–35.
[32] Zhang M, Luo H, Xi Z, Rogaeva E. Drug repositioning for diabetes based on 'omics' data mining. PLOS ONE 2015;10(5):e0126082.
[33] Kim TW. Drug repositioning approaches for the discovery of new therapeutics for Alzheimer's disease. Neurotherapeutics 2015;12(1):132–42.
[34] Chiang AP, Butte AJ. Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther 2009;86(5):507.
[35] Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol 2010;6(2):e1000662.
[36] Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science 2008;321(5886):263–6.
[37] Nugent T, Plachouras V, Leidner JL. Computational drug repositioning based on side-effects mined from social media. PeerJ Comput Sci 2016;2:e46.
[38] Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. Literature mining, ontologies and information visualization for drug repurposing. Brief Bioinform 2011;12(4):357–68.
[39] Li J, Lu Z. Systematic identification of pharmacogenomics information from clinical trials. J Biomed Inform 2012;45(5):870–8.
[40] Rastegar-Mojarad M, Elayavilli RK, Li D, Prasad R, Liu H. A new method for prioritizing drug repositioning candidates extracted by literature-based discovery. 2015 IEEE international conference on Bioinformatics and Biomedicine (BIBM). November 2015. p. 669–74.
[41] Zhu Q, Tao C, Shen F, Chute CG. Exploring the pharmacogenomics knowledge base (PharmGKB) for repositioning breast cancer drugs by leveraging Web ontology language (OWL) and cheminformatics approaches. Pacific symposium on biocomputing. 2014. p. 172.
[42] Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 2011;7(1):496.
[43] Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: a machine-learning approach through data integration. J Cheminform 2013;5(1):30.
[44] Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. AMIA annual symposium proceedings, vol. 2014. 2014. p. 1258.
[45] Yang J, Li Z, Fan X, Cheng Y. Drug–disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization. J Chem Inf Model 2014;54(9):2562–9.
[46] Leaman R, Wei CH, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminform 2015;7(S-1):S3.
[47] Wu C, Gudivada RC, Aronow BJ, Jegga AG. Computational drug repositioning through heterogeneous network clustering. BMC Syst Biol 2013;7(Suppl. 5):S6.
[48] Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, et al. Prediction of drug–target interactions and drug repositioning via network-based inference. PLoS Comput Biol 2012;8(5):e1002503.
[49] Rakshit H, Chatterjee P, Roy D. A bidirectional drug repositioning approach for Parkinson's disease through network-based inference. Biochem Biophys Res Commun 2015;457(3):280–7.
[50] Hu G, Agarwal P. Human disease–drug network based on genomic expression profiles. PLoS ONE 2009;4(8):e6536.
[51] Karimi S, Wang C, Metke-Jimenez A, Gaire R, Paris C. Text and data mining techniques in adverse drug reaction detection. ACM Comput Surv (CSUR) 2015;47(4):56.
[52] White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc 2013;20(3):404–8.
[53] Jiang L, Yang CC. Expanding consumer health vocabularies by learning consumer health expressions from online health social media. International conference on social computing, behavioral-cultural modeling, and prediction. March 2015. p. 314–20.
[54] Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24–9.
[55] Campbell WM, Dagli CK, Weinstein CJ. Social network analysis with content and graphs. Lincoln Lab J 2013;20(1):62–81.
[56] Han J, Sun Y, Yan X, Yu PS. Mining heterogeneous information networks Tutorial at the 2010 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'10). July 2010.
[57] Denecke K. Extracting medical Concepts from medical social media with clinical NLP tools: a qualitative study. Proceedings of the fourth workshop on building and evaluation resources for health and biomedical text processing 2014.
[58] Wang W, Yang S, Zhang X, Li J. Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics 2014;30(20):2923–30.
[59] Rastegar-Mojarad M, Elayavilli RK, Li D, Prasad R, Liu H. A new method for prioritizing drug repositioning candidates extracted by literature-based discovery. 2015 IEEE international conference on Bioinformatics and Biomedicine (BIBM). November 2015. p. 669–74.
[60] Lee H, Kang S, Kim W. Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures. PLOS ONE 2016;11(3):e0150460.
[61] Li J, Lu Z. Pathway-based drug repositioning using causal inference. BMC Bioinform 2013;14(16):S3.