Data Analytics for Pharmaceutical Discoveries
Shobeir Fakhraei, Eberechukwu Onukwugha, Lise Getoor
Healthcare Data Analytics, CRC Press 2015

Chapter 1

Data Analytics for Pharmaceutical Discoveries

Shobeir Fakhraei
Department of Computer Science
University of Maryland
College Park, MD
shobeir@cs.umd.edu

Eberechukwu Onukwugha
Department of Pharmaceutical Health Services Research
University of Maryland
Baltimore, MD
eonukwug@rx.umaryland.edu

Lise Getoor
Department of Computer Science
University of California
Santa Cruz, CA
getoor@soe.ucsc.edu

1.1 Introduction

Interdisciplinary computational approaches that combine statistics, computer science, medicine, chemoinformatics, and biology are becoming highly valuable for drug[1] discovery and development. Data mining and machine learning methods are being more commonly used to properly analyze the emerging high volumes of structured and unstructured biomedical and biological data from several sources including hospitals, laboratories, pharmaceutical companies, and even social media. These data may include sequencing and gene expression, drug molecular structures, protein and drug interaction networks, clinical trial and electronic patient records, patient behavior and self-reporting data in social media, regulatory monitoring data, and biomedical literature.

Data mining methods can be used in several stages of drug discovery and development to achieve different goals. Figure 1.1 summarizes the drug development and FDA[2] approval process diagram. Most new compounds fail during this approval process in clinical trials or cause adverse side effects. The cost of successful novel chemistry-based drug development often reaches millions of dollars, and the time to introduce the drug to market often comes close to a decade [1]. The high failure rate of drugs during this process, make the trial phases known as the "valley of death" [2].

Similar to many other domains, pharmaceutical data mining algorithms aim to limit the search space and provide recommendations to domain experts for hypothesis generation and further analysis and experiments. One way to categorize data mining and machine learning approaches is based on their application to pre-marketing and post-marketing stages. In pre-marketing stage, data mining methods focus on discovery activities, including but not limited to, finding signals that indicate relations between drugs and targets, drugs and drugs, genes and diseases, protein and diseases, and finding bio-markers. In this stage potential interactions that could cause therapeutic or adverse effects are studied. Most of the chemical compounds under study at this stage have not been through clinical trails, and the *in silico* experiments serve as a basis for further explorations for them. In post-marketing stage an important application of data analytics is in finding indications of adverse side effects for approved drugs. These algorithms provide a list of potential drug side effect associations that can be used for further studies.

In this chapter we provide a brief overview of some data analytics applications in this domain, and mainly focus on two major tasks from each stage. We first summarize some of the main methods for drug-target interaction prediction that is highly important during the pre-marketing stage. We then provide an overview of *pharmacovigilance* (or drug safety surveillance) which is an important focus in the post-marketing stage.

1.1.1 Pre-marketing stage

In the pre-marketing stage, data mining algorithms primarily focus on drug discovery and predicting potential adverse effects using characteristics of the compounds (e.g., drug targets, chemical structure) or screening data (e.g., bioassay data) [4]. One of the important challenges where data mining and machine learning methods could be very beneficial is drug-target interaction prediction. This task is also highly important for drug repurposing and drug adverse reactions prediction [5]. *In vitro* identification of drug-target associations is a labor-intensive and costly procedure. Hence, *in silico* prediction methods are promising approaches for focusing *in vitro* investigations [6].

Most drugs affect multiple targets, and *Polypharmacology*, the study of such interactions, is an area of growing interest [7]. These multi-target interactions potentially result in adverse side effects or unintentional therapeutic effects, and is the main cause in the high failure rate of drug

[1] Organic molecules that bind to bio-molecular targets and inhibit or activate their functions.

[2] U.S. Food and Drug Administration.

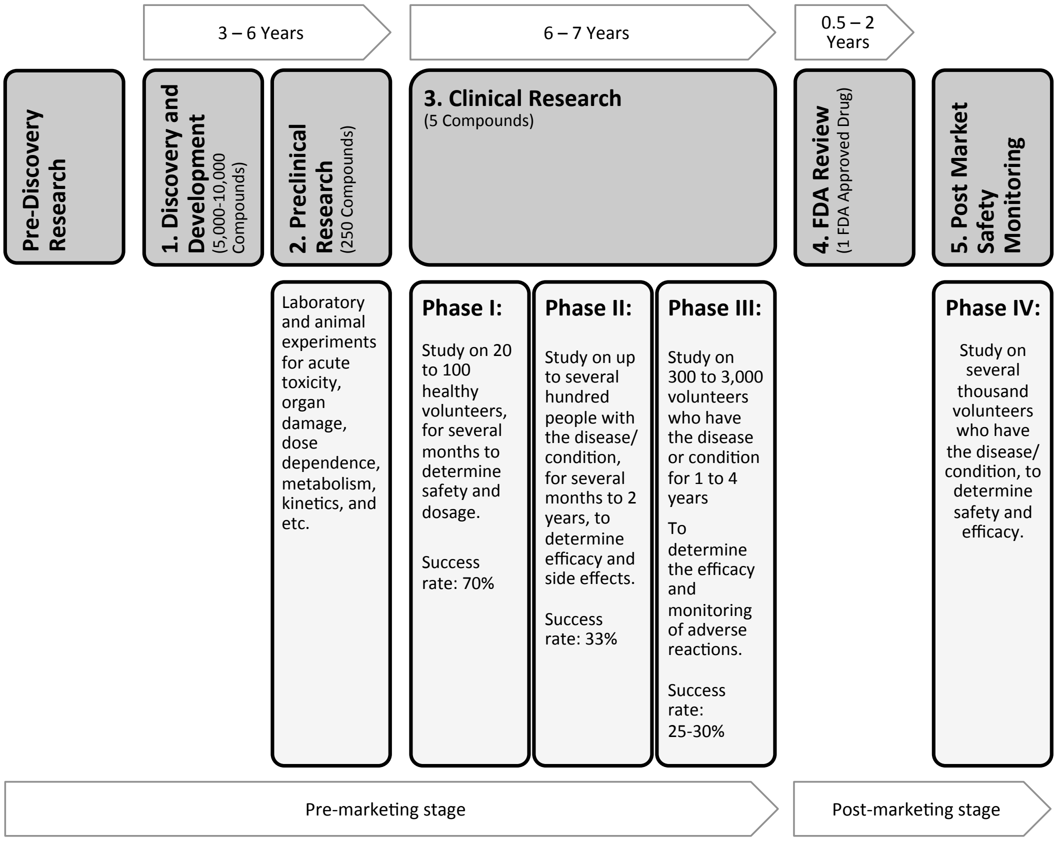
3 – 6 Years
6 – 7 Years
0.5 – 2 Years
Pre-Discovery Research
1. Discovery and Development (5,000-10,000 Compounds)
2. Preclinical Research (250 Compounds)
3. Clinical Research (5 Compounds)
4. FDA Review (1 FDA Approved Drug)
5. Post Market Safety Monitoring
Laboratory and animal experiments for acute toxicity, organ damage, dose dependence, metabolism, kinetics, and etc.
Phase I:
Study on 20 to 100 healthy volunteers, for several months to determine safety and dosage.
Success rate: 70%
Phase II:
Study on up to several hundred people with the disease/ condition, for several months to 2 years, to determine efficacy and side effects.
Success rate: 33%
Phase III:
Study on 300 to 3,000 volunteers who have the disease or condition for 1 to 4 years
To determine the efficacy and monitoring of adverse reactions.
Success rate: 25-30%
Phase IV:
Study on several thousand volunteers who have the disease/ condition, to determine safety and efficacy.
Pre-marketing stage
Post-marketing stage

FIGURE 1.1: The drug development process [3].

candidates in clinical trials. Unacceptable toxicities resulting from these interactions account for approximately 30% of the failures [8]. Predicting these interactions during the drug developmental phase can reduce the high cost of clinical trials and can be crucial for the commercial success of new drugs.

Moreover, due to the high cost and low success rate of novel drug development, pharmaceutical companies are also interested in *drug repositioning* or *repurposing*, which involves finding new therapeutic effects of pre-approved drugs. For example, *Sildenafil*, which was originally developed for pulmonary arterial hypertension treatment, was re-purposed and branded as *Viagra*, based on its side effect of treating erectile dysfunction in men [9].

Another important task where data mining algorithms can also be effective is drug-drug interaction prediction, which may account for up to 30% of unexpected adverse drug events and close to 50% in hospitalized patients [10]. For example, *Tramadol* (a pain reliever) can enhance the effect of *Fluoxetine (Prozac)*, increasing Serotonin levels and potentially leading to seizures [11]. The National Health and Nutrition Examination Survey [12] reports that over 76% of elderly Americans are taking two or more drugs each day. Another study estimated that 29.4% of elderly patients are taking six or more drugs [13]. The drug-drug interactions can be predicted in pre-marketing stage from compounds profiles [14, 15] or identified in post-marketing stage using signals from several sources [16, 17]. For example, Gottlieb et al. [18] infer drug-drug interactions based on several similarities between the drugs and the previously known interactions between them.

1.1.2 Post-marketing stage

In the post-marketing stage an important focus of data mining methods is on finding patterns that indicate potential drug related adverse events [4]. Undiscovered severe adverse events may lead to drug withdrawals which can be financially detrimental for the manufacturers [4]. Several drugs have been withdrawn from the market over the years [19]. For example, *Vioxx*, which was considered a powerful anti-inflammatory drug was withdrawn due to an increased coronary risk [20, 21].

Each year more than two million hospitalizations and injuries, and seven hundred thousand emergency visits in the United States have been estimated to be caused by these effects [22, 4, 23]. They have also been estimated to cost seventy five billion dollars annually [11]. It is also estimated that each year 6–7% of hospitalized patients experience severe adverse drug related events, which can lead to a potential hundred thousand deaths, making it the fourth largest cause of death in the U.S. [23].

Since only a limited number of patient characteristics are studied in clinical trials and for a limited duration, often complex safety issues associated with a new drug cannot be fully studied with clinical trials [11]. Adverse drug effects are often defined as the following [24]:

> “Any unintended and undesirable effects of a drug beyond its anticipated therapeutic effects occurring during clinical use.”

Pharmacovigilance (or drug safety surveillance) is the science that concerns with the detection, assessment, understanding and prevention of adverse drug reactions [4]. Pharmacovigilance is formally defined by World Health Organization (WHO) as [25]:

> “The science and activities related to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems.”

Data analysis algorithms are crucial to narrow the search space and detect the hidden patterns. Harpaz et al. [11] define data mining algorithms for Pharmacovigilance as:

> “Automated high-throughput methods to uncover hidden relationships of potential clinical significance to drug safety.”

They report that volume of publications on data mining methods for Pharmacovigilance index in PubMed[3] has grown from less than 40 in the year 2000 to about 200 a year in 2011. An important focus of data mining algorithms in post-marketing stage is on computing measures of statistical association between pairs of drugs and clinical outcomes recorded in underlying data sources [26].

1.1.3 Data sources and other applications

There are several important data mining applications that we do not address in this chapter. For example, another area of growing interest where data mining algorithms play a significant role is predicting individual drug responses and personalized medicine [27, 28]. Personalized medicine or *Pharmacogenomics*, is using an individual’s genetic profile to make the best therapeutic choice by facilitating predictions about whether that person will benefit from a particular medicine or will suffer serious side effects [29]. For example, *Pharmacogenomics Knowledgebase (PharmGKB)* is a resource that collects, curates, and disseminates information about the impact of human genetic variation on drug responses [30].

Data mining algorithms in different stages of drug development use different data sources.

[3] A search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics.

Chemical and biological data are mainly used in pre-marketing stage for tasks such as hypothesis generation and prediction, while spontaneous reporting systems, electronic health records and administrative claims data are often used in post-marketing data mining tasks mainly to detect signals of association. Biomedical literature and patient-generated data in health-related Internet forums has also received considerable research interest in recent years [11].

In the rest of the chapter, we highlight some of the related data mining tasks and methods based on the data resource they are applied to. First, we summarize some of the methods that use chemical and biological data focusing on approaches that predict drug-target and drug-drug interactions. We then highlight methods that detect patterns of drug related adverse events using spontaneous reports, electronic health records and patient generated data such as web search engine logs. We also mention some of the advances in application of data mining in biomedical literature that can facilitate pharmaceutical discoveries.

There is a plethora of high quality research recently published related to data analytics in pharmaceutical discoveries which we could not cover in this chapter. We did not aim to provide a complete or comprehensive survey; our goal was to provide highlights of some of the important data analytics methods in this domain.

1.2 Chemical and Biological Data

One of the important goals of data mining methods that use chemical and biological data is predicting interactions between chemical compounds (e.g., drugs) and biological targets (e.g., proteins) which could cause therapeutic or adverse effects, or interactions between two or more chemical compounds that could cause potential adverse effects. Openly available databases, including multiple resources available on the Internet that include drug related data and information about their targets are highly used for this task. These databases are used to study properties of drugs for several purposes, including drug-target and drug-drug interaction elucidation. Table 1.1 summarizes some of the more commonly used databases that contain information about drugs, their targets and interactions between them.

There are several methods to model the drug-target interaction prediction task [38]. They can be separated into two categories based on their explicit emphasis on the graph or network representation of drugs and targets interactions. The first category constructs a network structure to predict interactions [39], while others make predictions based on other factors. In this section, we mainly cover network based approaches.

Among the methods that do not use a network representation, the *similarity ensemble approach* (SEA), used ligands[4] to predict interactions between drugs and targets. They used ligands for target representation and chemical similarities between drugs and ligand sets as potential interaction indicators [6]. In CMap, Lamb [40] used RNA expressions to represent diseases, genes, and drugs [35]. They compared up- and down-regulations of the gene-expression profiles from cultured human cells treated with bioactive molecules and provided cross-platform comparisons. They predicted new potential interactions based on opposite-expression profiles of drugs and diseases.

Methods that consider the network structure address two important factors. The first one is how to construct the network and what information to include, and the other is how to predict new interactions. In the following sections, we summarize the main approaches for each task.

[4] A small molecule, that forms a complex with a biomolecule to serve a biological purpose.

Name	URL	Description
Drugbank [31]	www.drugbank.ca	Drug (i.e. chemical, pharmacological and pharmaceutical) data with their targets (i.e. sequence, structure, and pathway) information.
KEGG Drug [32]	www.genome.jp/kegg/drug	Chemical drug structures with their targets.
MATADOR (Manually Annotated Targets and Drugs Online Resource)[33]	matador.embl.de	Drugs and their target interactions.
DCDB (Drug Combination Database) [34]	www.cls.zju.edu.cn/dcdb	Drug combinations and their targets.
DBPedia	www.dbpedia.org	Drugs, diseases and proteins information extracted from Wikipedia.
ChEMBL	www.ebi.ac.uk/chembl	Trial drugs with their targets.
Connectivity Map (CMAP) [35]	www.broadinstitute.org/cmap	Genetic profile information about diseases and drugs.
Pubchem [36]	pubchem.ncbi.nlm.nih.gov	Biological activities of small molecules (i.e. drugs).
Therapeutic Target Database	bidd.nus.edu.sg/group/cjttd	Therapeutic protein and nucleic acid targets, disease, pathway information and the corresponding drugs.
PDTD (Potential Drug Target Database)	www.dddc.ac.cn/pdtd	Drug targets information, focused on the ones with known 3D-structures.
Drug2Gene [37]	www.drug2gene.com	A knowledge base combining the compound/drug-gene/protein information from several publicly available databases.

TABLE 1.1: Databases containing chemical and biological data.

1.2.1 Constructing a network representation

A number of research publications study network structures to predict interactions. Cockell et al. [9] described how to integrate drugs, targets, genes, proteins, and pathways into a network for different tasks. Nodes in this network usually include drugs, proteins and diseases, and edges include their interactions and similarities, where similarities could be extracted from several sources such as chemical structure of the compounds [5]. Figure 1.2 shows an example of a schematic overview of such networks.

Lee et al. [41] described drug repurposing, multi-agent drug development, and estimation of drug effects on target perturbations via network-based solutions. Yildirim et al. [39] explained trends in the drug-discovery industry over time using a network-based analysis and showed that sequencing the genome is changing the traditional trends of drug development. They also discussed different structural aspects of this network including preferential attachment and cluster formation.

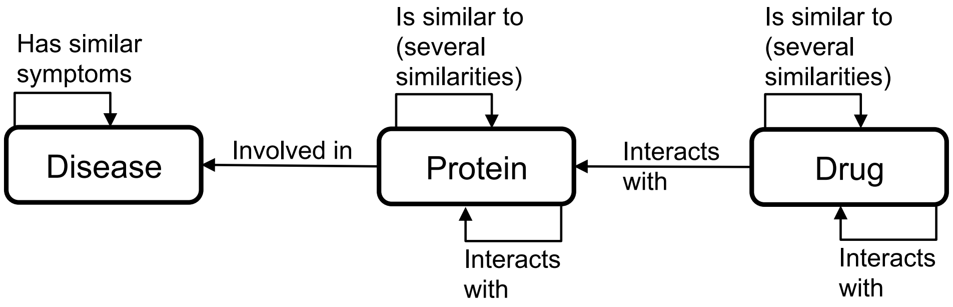

FIGURE 1.2: Network representation example of drugs, targets and diseases.

A common approach to predict new interactions is to construct a bipartite interaction network where nodes represent drugs and targets, and edges denote interactions. Drug-drug and target-target similarities can augment this network on each side. Data from multiple publicly accessible datasets can be integrated to build these networks [41]. The similarities between drugs and between targets can have different semantics. For example, targets can have similarity measures based on their sequences and their ontology annotations [5, 42]. Another example is drug side effects; while potential drug side effects can be predicted via the drug-target interaction predictions [43] they can also be used as similarities between drugs to predict new targets [44]. There are a few databases that contain information about the drugs' known side-effects. Table 1.2 summarizes a few datasets that contain this information.[5]

Name	**URL**	**Description**
SIDER [45]	sideeffects.embl.de	Information on marketed drugs and their recorded adverse drug reactions.
Drugs.com	www.drugs.com/sfx	Information about drugs and their side effects.
MedlinePlus	www.nlm.nih.gov/medlineplus/druginformation.html	National Library of Medicine's website containing drugs and their side effects.

TABLE 1.2: Databases that include drug side effects.

1.2.2 Interaction prediction methods

In the drug-target interaction network, similar targets tend to interact with the same drugs, and similar drugs tend to interact with the same targets [9]. Using variations of this intuition a link prediction method can predict new potential drug-target interactions in a drug-target interaction network [46].

[5] These databases mainly do not focus on chemical and biological data.

1.2.2.1 Single similarity based methods

Network-based approaches integrate drug-drug and target-target similarities extracted via different methods (e.g. SEA and CMap) with the drug-target interactions network [38]. The following methods proposed a single similarity measure for drugs and targets to predict interactions.

Cheng et al. [47] predicted potential interactions using drug-drug and target-target similarities and a bipartite interaction graph. Using SIMCOMP [48], they computed the 2D chemical drug similarities and sequence similarities for targets via the Smith-Waterman score. They used the following three link-prediction methods:

1. Drug-based similarity inference (DBSI) where they only considered similarities between drugs for prediction.
2. Target-based similarity inference (TBSI) where they only considered target similarities for prediction.
3. Network-based inference (NBI) where they combined both drug-drug and target-target similarities.

Alaimo et al. [49] extended this approach by proposing a hybrid drug-target method that integrated prior domain knowledge.

Yamanishi et al. [50] proposed three methods for interaction prediction, including a nearest neighbor approach, a weighted *k*-nearest neighbors approach, and a space integration. In their space integration method, they described a genomic space, using the Smith-Waterman score, and a pharmaceutical space, using the SIMCOMP score. They proposed a method to integrate drugs and targets in a unified latent *pharmacological space*, and they predicted interactions based on the proximity of drugs and targets in that space. Figure 1.3 shows an overview of their method. They separated out four categories of targets, namely enzymes, ion channels, GPCR, and nuclear receptors for their experiments which was adopted by most subsequent drug-target interaction prediction methods [38]. Overall steps in their method include:

1. Embed compounds and proteins on the interaction network into a unified space called “pharmacological space”.
2. Learn a model between the chemical/genomic space and the pharmacological space, and map any compounds/proteins onto the pharmacological space.
3. Predict interacting compound–protein pairs by connecting compounds and proteins which are closer than a threshold in the pharmacological space.

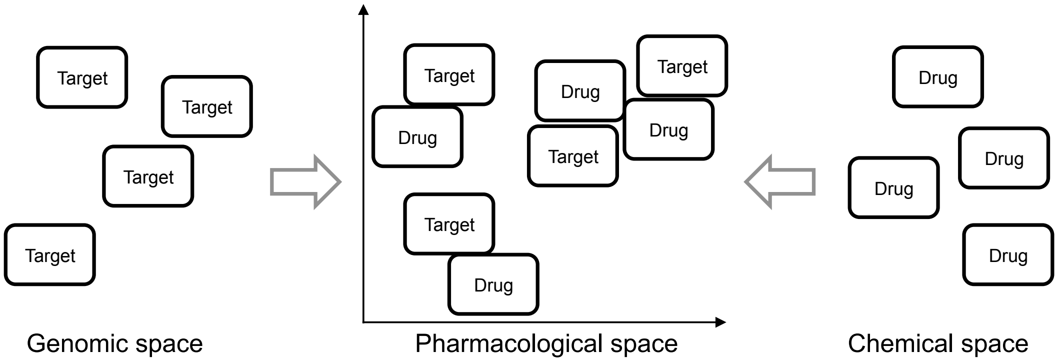

FIGURE 1.3: Overview of Yamanishi et al.’s [50] method.

Bleakley and Yamanishi [51] extended this method by constructing local models for graph inference. They classified each interaction twice and combined the results to provide predictions. First, they built a classifier based on drugs and then based on targets. They used the similarities as the *support vector machine* (SVM) kernels. Further extending this method, Mei et al. [52] proposed an approach to infer training data from neighbors' interaction profiles to make predictions for new drug or target candidates that do not have any interactions in the network. Wang and Zeng [53] proposed a method based on restricted Boltzmann machines for drug-target interaction prediction.

1.2.2.2 Multiple similarity based methods

More complex methods can predict interactions based on multiple heterogeneous similarities. Chen et al. [54] reasoned about the possibility of a drug-target interaction in relation with other linked objects. They used distance, shortest paths, and other topological properties in the network to assess the strength of a relation. Their method assigned scores to paths between drugs and targets and combined path scores for each drug-target pair.

Perlman et al. [42] proposed a supervised learning method along with a feature-engineering approach based on combinations of drug-drug and target-target similarities to predict interactions, called *similarity-based inference of drug targets (SITAR)*. They built their model based on five drug-drug and three target-target similarities. For each potential drug-target interaction, they built a feature based on how similar that potential drug-target interaction is to one of the observed interactions in the network, and computed the interaction similarity based on the weighted combination of the drug-drug and target-target similarities. Overall steps of their method include:

1. They considered chemical-based, ligand-based, expression-based, side-effect-based, and annotation-based similarities between drugs, and computed target similarities using sequence-based, protein-protein interaction network-based, and gene ontology-based information.
2. They built a dataset where each link (i.e., drug-target pair) is a sample (i.e., row) and computed 15 (i.e., 5×3) features for each link based on the similarities. The sample was labeled with class 1 if the drug-target pair was a known interaction, and 0 otherwise.
3. Their model computed the value of each feature based on how similar the potential drug-target interaction is to the closest observed interaction in the network, and computed the similarity of the interaction based on the weighted combination of the drug-drug and target-target similarities.
4. A logistic regression classifier on this dataset was then used to predict new interactions.

Fakhraei et al. [5, 55] proposed a drug-target interaction prediction framework based on *probabilistic soft logic* (PSL), to collectively predict interactions using a structured representation of the network. Their interpretable model captured the multi-relational characteristics of the drug-target interaction network (i.e., nodes and edges with different semantics). They proposed PSL models that reason over rules, based on triad and tetrad structural intuitions, and improved the prediction result of Perlman et al. [42]. Figure 1.4 shows how similarities were used in their method for new drug-target interaction prediction in their triad-based rules, which captures the tendency of similar targets interacting with the same drugs, and similar drugs interacting with the same targets. They used the following steps for their predictions:

1. Similar to SITAR [42], they used five drug-drug and three target-target similarities.
2. They used a blocking threshold to only include the k most similar drugs or targets for each entity in their model.

3. They defined rules based on triad structures with the overall intuition that similar drugs tend to interact with the same target, and a drug tends to interact with similar targets. They introduced a rule for each similarity measure (i.e., eight rules in total).
4. They defined tetrad based rules with the intuition that similar drugs tend to interact with similar targets.
5. They considered a negative prior to capture the sparsity of the network.
6. They studied the effect of collective inference and combination of similarities in improving the performance.

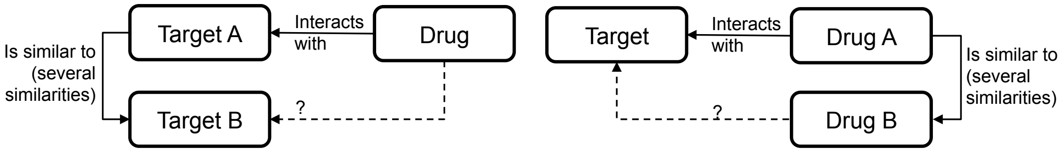
Target A
Interacts with
Drug
Is similar to (several similarities)
Target B
?
Target
Interacts with
Drug A
Is similar to (several similarities)
?
Drug B

FIGURE 1.4: Predicting new drug-target interactions based on drug-drug and target-target similarities.

1.3 Spontaneous Reporting Systems (SRS)

Spontaneous reporting systems are important data sources for post-marketing analysis and data mining, and up until recent years they have been the main resource for pharmacovigilance (or drug safety surveillance) [56, 57].

> “Spontaneous reporting systems are passive systems composed of reports of suspected adverse drug events collected from health-care professionals, consumers, and pharmaceutical companies, and maintained largely by regulatory and health agencies [11].”

They have supported regulatory decisions for a long list of marketed drugs since their inception. There are two main spontaneous reporting systems administrated by *US Food and Drug Administration* (FDA) and *World Health Organization* (WHO) which are described in Table 1.3.

Spontaneous reporting systems (SRS) have a structured format and include information on the drug suspected to cause the adverse reaction. They also contain limited demographic information [11]. There are multiple advantages in using them for pharmacovigilance. They are centralized sources of information focused on drug–adverse event relationships and cover large populations. They are also accessible for analysis and research [58].

While spontaneous reporting systems are the main source of post-market drug adverse effect identification, they have several limitations [59]. Only a fraction of adverse drug events are identified and reported in these systems because reporting to them is on a voluntary basis, except for pharmaceutical companies, which are required to report suspected adverse drug reaction. They may also contain biased reporting rates and missing patient data [60, 61].

There are many reasons causing such limitations in spontaneous reporting systems. Physicians may worry about possible legal issues after disclosure of medical errors, or may not clearly understand the definition of an adverse events. They may consider that the circumstances of a case or its

Name	URL	Description
FDA Adverse Event Reporting System (FAERS)	www.fda.gov/Drugs/ GuidanceCompliance Regulatory Information/Surveillance/ AdverseDrugEffects	Information on adverse event and medication error reports submitted to FDA.
VigiBase	www.umc-products.com/ vigibase_services	World Health Organizations global individual case safety reports database.

TABLE 1.3: Spontaneous reporting systems dataset.

outcome do not warrant reporting, or do not believe that reporting will lead to improvement. They may also not see what added value the growing body of quality and safety guidelines provide in terms of patient outcomes [21]. Amalberti et al. [21] and Strom [20] suggest several improvements to the spontaneous reporting systems to address some of these challenges and also introduce further categories to better organize the reports for analysis.

The main methods that focus on spontaneous reporting systems data are designed to generate measures of statistical association for large sets of drug–outcome pairs. These signals can be used to prioritize and identify risks for further evaluations. Newer approaches have been designed to facilitate identification of higher-order or multi-variate associations that represent more complex safety phenomena such as drug-drug interactions [11]. The main methods for signal identification using the spontaneous reporting systems data are summarized in the following sections.

1.3.1 Disproportionality analysis

Disproportionality analysis plays an important role in most methods applied to spontaneous reporting systems data. Frequency analysis of 2×2 contingency tables (shown in 1.4) is used to estimate measures of statistical association between specific drug–event combinations mentioned in spontaneous reports. Disproportionality analysis methods differ in statistical adjustments for low numbers, and their assumptions. Two main categories of them are frequentist and Bayesian methods [4, 11].

	With target adverse event	Without target adverse event	Total
With target drug	a	b	n = a + b
Without target drug	c	d	c + d
Total	m = a + c	b + d	t = a + b + c + d

TABLE 1.4: Contingency table used in disproportionality analysis of spontaneous reporting systems data.

The *relative reporting ratio* is the most widely discussed measure for disproportionality analysis and is defined as the ratio of the observed incidence rate of a drug–event combination to its *baseline* expected rate under the assumption that the drug and event occur independently. Both the U.S. Food and Drug Administration and the World Health Organization use a Bayesian version of the relative reporting ratio as a basis for monitoring safety signals in their spontaneous reporting systems [11]. Frequentist approaches use one of the measures listed in Table 1.5 to estimate associations and are typically accompanied by hypothesis tests of independence. The hypothesis tests are used as an extra precautionary measure to take into account the sample size used while computing an association.

Measure of association	**Mathematical definition**
Relative reporting ratio (RRR)	$\frac{t \times a}{m \times n}$
Proportional reporting ratio	$\frac{a \times (t-n)}{c \times n}$
Reporting odds ratio	$\frac{a \times d}{c \times b}$
Information component	$\log_2(RRR)$

TABLE 1.5: Mathematical definitions of measures of association.

The uncertainty associated with small counts in Bayesian approaches is addressed by *shrinking* the measure towards no association by a proportional amount [11]. Among the Bayesian approaches *multi-item gamma Poisson shrinker* is a predominant algorithm used in the United States by the FDA, the United Kingdom, and several pharmaceutical companies. An empirical Bayes geometric mean is used in this method, which is a centrality measure of the posterior distribution of the true relative reporting ratio. The World Health Organization uses a similar Bayesian approach, called the *Bayesian Confidence Propagation Neural Network* [11]. However, due to lack of a gold standard to evaluate the performances of these methods, it is accepted that none of them is universally better than any other. Also, their results differences tend to fade with the increase in the number of reports of a specific drug–event combination [11].

1.3.2 Multivariate methods

Traditional disproportionality analysis approaches do not properly support the discovery and analysis of higher-dimensional drug safety phenomena that involve more than one drug or event, and the confounding issue [4, 11]. Hauben and Bate [62] report the importance and difficulties with more complex drug safety phenomena detections. Methods that aim to identify signals of adverse events based on multiple drugs, should be able to detect hidden drug–drug interactions [16].

Confounding is another challenge in these analyses. A confounder is an observed or unobserved variable that mediates an association between other variables. Many related research publications have focused on confounding by drug co-administration. In these cases a drug that is frequently co-prescribed with another drug could be mistakenly associated with an event rather than the correct one [11]. Several multivariate approaches have been proposed to address these issues. We summarized some of them in this section.

- **Disproportionality analysis extensions;** This method has been applied mostly to three-dimensional associations corresponding to drug–drug interactions [63], for which observed-to-expected ratios are calculated in a similar manner but based on three elements (i.e., $drug_1$–$drug_2$–*event*).

- **Multivariate logistic regression;** Stratification is traditionally used to address confounding; however, this method is not effective for studying a large number of potential confounders [64]. Multiple logistic regressions can be a more appropriate approach to deal with confounding. It can estimate the drug–event association by controlling or adjusting for the presence of other potential confounders [64]. Caster et al. [65] applied Bayesian logistic regression [66] –which can carry out regression analyses with millions of covariates– to address confounding in World Health Organization spontaneous reporting system data.

- **Associations rule learning**[6]**;** This method is well established for discovering relations between variables in large databases using specific measures of interestingness. Agrawal et al. [67] introduced association rules for discovering regularities between products in large-scale transaction data in supermarkets. The *Apriori* algorithm is usually used to deal with the huge search space in association rule learning. Rouane-Hacene et al. [68] applied association rule learning to find association of up to three anti-HIV drugs. Harpaz et al. [69] extended this method to capture associations of up to six drugs.

- **Bi-clustering;** Bi-clustering is simultaneous clustering of the data matrix rows and columns to find sub-matrices that exhibit highly correlated activities [70]. Harpaz et al. [71] used a bi-clustering algorithm to identify associations between multiple drugs and adverse effects.

- **Network analysis;** Another approach to adverse event identification from spontaneous reporting systems are based on constructions and analysis of network structures. Ball and Botsis [72] constructed a network for vaccine adverse events where nodes in the network correspond to vaccines and reported events. They observed this network is *scale-free*[7] and proposed using *hubs* in this network to identify patterns of adverse events caused by *HPV4* vaccines. Zhang et al. [73] also constructed bipartite networks of vaccines, diseases and genes to analyze vaccine adverse event data.

In any of the above methods, there are several factors that one should consider in analysis of spontaneous reporting systems data. Amalberti et al. [21] identified examples of incorrect conclusion due to the fact that the studies were performed with simplistic assumptions and only looking for the cause in the researchers own specialty. They also identified that many studies consider a short time-frame and thus miss many adverse events, and proposed three different time frames to study the effect of adverse events. They also highlighted that many studies of adverse events are influenced by emotions and media coverage and are often insurance-driven, while the ones that may have an impact on a larger population are left without much attention.

1.4 Electronic Health Records

As mentioned in the previous section there are multiple limitations with spontaneous reports, that encouraged the use of several other sources to identify drug adverse events signals. One source is electronic health records and administrative claims data. Electronic health records have some advantages when compared to spontaneous reporting systems data. They are captured during the usual course of care and contain more detailed medical data, such as patients clinical history, and the timing of symptom development and medication administration. There is also no need to estimate the reporting frequency as events are captured as part of the standard care [61].

Electronic health records are being increasing used throughout the United States, potentially providing more data [74]. Initiatives like the *Observational Medical Outcomes Partnership* in the USA and the *Exploring and Understanding Adverse Drug Reactions* project in Europe are focusing on building electronic health records based surveillance systems [13]. The efficacy of electronic health records to identify adverse drug events was shown by Ramirez et al. [10]. They used abnormal laboratory signals to identify patients with adverse drug events. Via retrospective studies, Brown et al. [75] have also shown that the adverse events caused by *Vioxx* could have been found sooner based on electronic health records.

[6] Also referred to as association rule mining.

[7] A network with power law degree distribution.

Electronic health records can be categorized into structured coded data and unstructured clinical notes [4]. ICD[8] codes [76], laboratory data and vital measurements [77] are among the structured coded data that have been used to detect association of drugs and adverse events. Wang et al. [78] proposed one of the early methods to use unstructured clinical notes to detect drug adverse event associations. In additional to the challenges with structured coded data, unstructured clinical notes require methods that can extract relevant information from free text clinical narratives. We discuss some details of the natural language processing and text mining methods in section 1.6.

Since electronic health records are mainly captured for diagnoses (usually based on billing codes) and not adverse drug event detection, they often require pre-processing to support analysis. Health-care providers often use different solutions for documentation and encoding the data. There are also legal and privacy concerns in accessing patients data causing logistical issues in sharing, accessing, and storing data.

There is a need for methods that can address confounding in the observational studies. There are methods to apply disproportionality analysis on electronic health records data, and also methods that are based on cohort designs, case-control designs, and self-controlled designs [11]. Cohort designs partition the subject cohorts based on their exposure to the drug, case-control designs divide them based on the event, and self-controlled designs compare the same subjects before and after they were exposed to the drug.

Electronic health records can also be used to detect more complex signals for drug–drug interactions. Iyer et al. [13] proposed adjusted disproportionality ratios to identify significant drug-drug-event associations among 1165 drugs and 14 adverse events. They published the database of population event rates among patients on drug combinations based on the electronic health records corpus from *Stanford Translational Research Integrated Database Environment*. Their method's overall steps include:

1. They first annotated the clinical text, extracting drugs and events of interest, which focused on 14 adverse events.
2. They then constructed 2×2 contingency table based on cohort design, where the exposed group are patients who have taken both drugs, and the comparison group include patients who have taken one or none of the drugs.
3. Then they computed the population event rate for patients who have taken both drugs.

They found that the interaction between *Amiodarone* and *Haloperidol* known to cause *QT prolongation* could have been detected based on signals from *Stanford Translational Research Integrated Database Environment* data as early as 2007. The *FDA Adverse Event Reporting System* started receiving reports for this interaction in 2009 [13]. They also showed that signals from electronic health records can be as useful as signals from spontaneous reporting systems. However, most methods do not indicate causality, instead they only show correlation and are means to provide early warning to focus more extensive investigations.

Harpaz et al. [26] proposed an empirical Bayes model to combine signals extracted from electronic health records and spontaneous reports. They showed an average 40% improvement by combining results from these sources in comparison to using each source independently.

[8] International Classification of Diseases.

1.5 Patient Generated Data on the Internet

Patient generated data such as web search logs, social networks, and health-care related forums are resources that contain medical related information on the Internet. Surveys suggest about 60-70% of adults search for health and medical information online, and about 80% of online health inquiries start at a search engine [79]. Table 1.6 summarizes some of the resources that contain patient generated medical data.

Several systems that make medical predictions based on these types of data have received recent attention from research community that either support or question their findings. For example, Google Flu trends [80] that uses aggregated Google search data to estimate flu activity has been both perceived positively [81] and also criticized [82] by the research community.

Name	**URL**	**Purpose**
Ask a Patient	www.askapatient.com	To share and compare medication experiences.
DailyStrength	www.dailystrength.org	Support group forum to discuss medical conditions.
PatientsLikeMe	www.patientslikeme.com	To find patients with similar condition and share experiences.

TABLE 1.6: Online resources with patient generated medical information.

Although the information provided by patients may be inaccurate or even questionable, such forums can provide valuable supplementary information on drug effectiveness and side effects because they cover large and diverse populations and contain unsolicited, uncensored data directly from patients [11]. However, extracting such information is very challenging and requires statistical and linguistic models to interpret conversational styles, correct spelling and grammatical errors, and separate gossip from real experiences.

Noise, influence of experiences, different bias factors, profession, and online content exposure are some examples that can contaminate the search engine log signals. There could be several reasons for users to search for symptoms and medications; for example, medical professionals could search for these information regularly [79].

Leaman et al. [83] extracted information from DailyStrength posts and found high correlation between user reported drug adverse events and the documented cases. White et al. [84] showed that interaction of *Paroxetine* and *Pravastatin* which can cause hyperglycemia could have been detected based on web search logs prior to its identification. In another example, Freifeld et al. [85] showed the efficacy of the Twitter data for pharmacovigilance.

White et al. [79] combined the signals from search engines logs of 80 million users over 18 months with the FDA's adverse event reporting system, and showed that the detection performance can be improved by 19%. In their analysis they applied the following steps to detect signals form the search queries:

1. They performed entity recognition and resolution to map synonym search terms into a unified representation for drugs, conditions and symptoms.
2. They excluded a portion (approximately 9%) of users from their analysis, based on the frequency of their queries (for internet bots) and the time they first started submitting medical queries (for health-care professionals).

3. For a drug of interest, they considered a surveillance window around the first occurrence of a query (defined as t_0).
4. To exclude exploratory searches and the queries influenced by reading the online articles related to the drug they defined an exclusion window around t_0.
5. They defined and computed a measure for a self-controlled study design called *query rate ratio (QRR)* as the ratio of number of after to before symptom or condition related queries around t_0 to indicate the association of drug-symptom condition.

1.6 Biomedical Literature

Natural language processing and text mining can be used for knowledge representation and hypothesis generation based on biomedical literature. For example, Shetty and Dalal [86] used research articles index on Pubmed[9] that mention specific drugs and adverse events to rank potential drugs adverse event relations. They applied several preprocessing steps and disproportionality analysis for their approach and showed that the association between *Vioxx* and myocardial infarction could have been found sooner. Haerian et al. [61] used the *Medical Language Extraction and Encoding System* (MedLEE) [87], a clinical NLP system developed at Columbia University, to analyze electronic health record and detect drug adverse events. MedLEE has also been used for automated knowledge acquisition from text, extracting adverse events from health records, and quality of care assessment [88, 89, 90].

Text mining could be beneficial to the pharmaceutical industry in several ways; For example, it could facilitate literature reviewing for medical professionals, and identify and extract relevant information. Such information can be extracted from unstructured clinical notes, like those in electronic health records and also biomedical research articles [91, 92]. However, biomedical literature is a very rich resource of information that can be used for discoveries beyond drug adverse events predictions. Mining the biomedical literature has been successfully used to discover new relationships among genes, biological pathways, diseases, or even for drug repurposing [93, 11].

Information extraction from a huge volume of available research literature is a challenging task. For example, Thorn et al. [94] highlight this problem as one of their main challenges in maintaining *pharmacogenomics knowledge base (PharmGKB)*. They have developed a natural language processing framework to streamline the identification of articles of interest and speed up the annotation process [95]. Due to these challenges biomedical literature mining has increasingly become a focus of active research in recent years. *BioCreative*[10] (Critical Assessment of Information Extraction systems in Biology) is an international community-wide effort that evaluates text mining and information extraction systems applied to the biomedical domain. *BioNLP*[11] is organizing events to support application of natural language processing on biomedical literature. There is a considerable amount of research literature and methods addressing the biomedical text mining challenges which we could not summarize in this section. Demner-Fushman et al. [96] and Krallinger et al. [97] provided a survey of biomedical and clinical text mining research, and Hahn et al. [93] summarized the recent advanced in text mining for pharmacogenomics.

In this section we briefly highlight some of the main related tasks and resources to pharmaceutical discoveries and pharmacogenomics. Pharmacogenomics publications span the intersection

[9] www.ncbi.nlm.nih.gov/pubmed

[10] http://www.biocreative.org

[11] http://bionlp.org

of research in genotypes, phenotypes and pharmacology. Some of the applications of pharmacogenomics text mining includes guiding human database curation, discovering interactions and potential cause–effect phenomena such as candidate gene ranking, drug–drug interaction and adverse drug interaction prediction, and drug repurposing [93]. An interesting application of biomedical text mining, pioneered by Don Swanson, is literature-based discovery and hypothesis generation [98], where the goal is to find implicit and novel information relating entities that are not explicitly spelled out in the underlying documents. Some of main tasks in biomedical literature mining includes corpus development, entity recognition and resolutions, relation extraction, and creation and use of ontologies.

There are three main types of entities of interest for recognition and resolutions in biomedical literature for pharmacogenomics; genotypes, phenotypes, and pharmacological entities. The most important genotype entity types are genes and proteins. Phenotype entities mainly include pathological phenomena and diseases in particular, as well as their anatomical sites, conditions and treatment. Pharmacological entities are drugs and other chemicals that are functionally important in treating or causing medically significant phenotypes in the course of treatments and therapies. One of challenges in biomedical literature mining is entity recognition and resolution. Several databases are used as canonical resources for entity resolution in biomedical literature, some of which are highlighted in Table 1.7.

Name	**URL**	**Entities**
EntrezGene	www.ncbi.nlm.nih.gov/gene	genotype
UniProt	www.uniprot.org	genotype
Medical Subject Headings (MeSH)	www.ncbi.nlm.nih.gov/mesh	phenotypes, pharmaceutical
Unified Medical Language System (UMLS)	www.nlm.nih.gov/research/umls	phenotypes, pharmaceutical
International Classification of Diseases (ICD-10)	www.who.int/classifications/icd/en	phenotypes
Systematized Nomenclature of Medicine – Clinical Terms (SNOMED–CT)	www.ihtsdo.org/snomed-ct	phenotypes
Medical Dictionary for Regulatory Activities (MedDRA)	www.meddramsso.com	phenotypes
DrugBank	www.drugbank.ca	pharmaceutical
Chemical Entities of Biological Interest (ChEBI)	www.ebi.ac.uk/chebi	pharmaceutical
KEGG	www.genome.jp/kegg	pharmaceutical, genotype
Human Metabolome Database (HMDB)	www.hmdb.ca	pharmaceutical, genotype
ChemIDplus	chem.sis.nlm.nih.gov/chemidplus	pharmaceutical

TABLE 1.7: Canonical databases for entity resolution.

More complex tasks in biomedical text mining deal with finding relations between entities. Genotype–phenotype relation extraction aims to identify which genotypes can play a role in which diseases [97]. Genotype–pharmaceutical relation extraction focuses on personalized medicine and possibility of tailoring drugs given a genetic context [99]. Phenotype–pharmaceutical relation extraction mostly concentrate on finding drug side effect and associated adverse effects [100]. Genotype–phenotype–pharmaceutical relations are more complex relations that aim to find genetic information and relate them to phenotype–pharmaceutical level. Typically, studies in this area combine text mining with other sources of information, to derive better conclusions [101].

1.7 Summary and Future Challenges

Data mining and data analytics are becoming more widely used in pharmaceutical discoveries. From proposing new hypotheses to detecting adverse event patterns, data mining methods are used to analyze chemical and biological data, spontaneous reports, electronic health records, biomedical literature, and most recently patient generated Internet data. But only in recent years new opportunities and interests have emerged to analyze data that have not been traditionally available and used for pharmaceutical discoveries.

These methods can advance pharmaceutical discoveries by potentially allowing for personalized medicine, drug re-purposing, more effective drug design and developments, and also active and proactive paradigms of surveillance. These new horizons also introduce new interesting challenges that should be addressed by the research community. There is a need for methods that can better address confounding and detect multi-drug interactions and adverse events related to them. Combining different sources of information to provide better predictions and hypotheses is also an interesting area of research.

Most tasks and data in pharmaceutical domain that data mining algorithms can be effective on, lack negative samples. For example, in drug-target interaction prediction, while positive interactions are well documented, a lack of interactions is not properly captured in the commonly used databases. Hence, it is not really known whether the absence of interaction data is due to a lack of real interaction, or it is due to the fact that such interaction has not been properly studied yet. In addition to challenges for standard supervised learning in data mining methods, this limitation introduces problems for standard evaluations.

In addition, further research is needed to understand the relative benefits and limitations of each data source and effectively integrate information from multiple data sources, including biomedical literature, biological/chemical data, user generated data on the internet for pharmaceutical discoveries and pharmacovigilance. The application and development of data mining and machine learning methods to facilitate and help advance pharmaceutical discoveries, has great potentials and need to be further investigated.

Acknowledgments

We are thankful to the editors, anonymous reviewers, and James Foulds for their valuable comments, insightful suggestions, and constructive feedback that greatly helped improving this chapter. We also gratefully acknowledge funding provided by The University of Maryland/Mpowering the State through the Center for Health-related Informatics and Bioimaging, and by the National Science Foundation (NSF) under contract number IIS0746930. Any opinions, findings, 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.

Bibliography

[1] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010.

[2] David J. Adams. The valley of death in anticancer drug development a reassessment. *Trends in Pharmacological Sciences*, 33(4):173 – 180, 2012.

[3] FDA drug development process. http://www.patientnetwork.fda.gov/learn-how-drugs-devices-get-approved/drug-development-process.

[4] Mei Liu, Michael E Matheny, Yong Hu, and Hua Xu. Data mining methodologies for pharmacovigilance. *ACM SIGKDD Explorations Newsletter*, 14(1):35–42, 2012.

[5] Shobeir Fakhraei, Bert Huang, Louiqa Raschid, and Lise Getoor. Network-based drug-target interaction prediction with probabilistic soft logic. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2014.

[6] Michael J. Keiser, Vincent Setola, John J. Irwin, Christian Laggner, Atheir I. Abbas, Sandra J. Hufeisen, Niels H. Jensen, Michael B. Kuijer, Roberto C. Matos, Thuy B. Tran, Ryan Whaley, Richard A. Glennon, Jérôme Hert, Kelan L. H. Thomas, Douglas D. Edwards, Brian K. Shoichet, and Bryan L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462 (7270):175–181, November 2009.

[7] Aislyn D.W. Boran and Ravi Iyengar. Systems approaches to polypharmacology and drug discovery. *Current Opinion in Drug Discovery & Development*, 13(3):297, 2010.

[8] Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4(11):682–690, 2008.

[9] S. J. Cockell, J. Weile, P. Lord, C. Wipat, D. Andriychenko, M. Pocock, D. Wilkinson, M. Young, and A. Wipat. An integrated dataset for in silico drug discovery. *Journal of Integrative Bioinformatics*, 7(3):116, 2010.

[10] E Ramirez, AJ Carcas, AM Borobia, SH Lei, E Piñana, S Fudio, and J Frias. A pharmacovigilance program from laboratory signals for the detection and reporting of serious adverse drug reactions in hospitalized patients. *Clinical Pharmacology & Therapeutics*, 87(1):74–86, 2010.

[11] Rave Harpaz, William DuMouchel, Nigam H Shah, David Madigan, Patrick Ryan, and Carol Friedman. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics*, 91(6):1010–1021, 2012.

[12] National health and nutrition examination survey. http://www.cdc.gov/NCHS/NHANES.htm.

[13] Srinivasan V Iyer, Rave Harpaz, Paea LePendu, Anna Bauer-Mehren, and Nigam H Shah. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association*, 21(2):353–362, 2014.

[14] Sean Ekins and Steven A Wrighton. Application of in silico approaches to predicting drug–drug interactions. *Journal of Pharmacological and Toxicological Methods*, 45(1):65–69, 2001.

[15] Jialiang Huang, Chaoqun Niu, Christopher D Green, Lun Yang, Hongkang Mei, and Jing-Dong J Han. Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS Computational Biology*, 9(3):e1002998, 2013.

[16] Nicholas P Tatonetti, Guy Haskin Fernald, and Russ B Altman. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *Journal of the American Medical Informatics Association*, 19(1):79–85, 2012.

[17] Nicholas P Tatonetti, P Ye Patrick, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125):125ra31–125ra31, 2012.

[18] Assaf Gottlieb, Gideon Y. Stein, Yoram Oron, Eytan Ruppin, and Roded Sharan. INDI: a computational framework for inferring drug interactions and their associated recommendations. *Molecular Systems Biology*, 8(1), July 2012.

[19] Kathleen M Giacomini, Ronald M Krauss, Dan M Roden, Michel Eichelbaum, Michael R Hayden, and Yusuke Nakamura. When good drugs go bad. *Nature*, 446(7139):975–977, 2007.

[20] Brian L Strom. How the US drug safety system should be changed. *Journal of the American Medical Association*, 295(17):2072–2075, 2006.

[21] René Amalberti, Dan Benhamou, Yves Auroy, and Laurent Degos. Adverse events in medicine: Easy to count, complicated to understand, and complex to prevent. *Journal of Biomedical Informatics*, 44(3):390–394, 2011.

[22] Adam L Cohen, Daniel S Budnitz, Kelly N Weidenbach, Daniel B Jernigan, Thomas J Schroeder, Nadine Shehab, and Daniel A Pollock. National surveillance of emergency department visits for outpatient adverse drug events in children and adolescents. *The Journal of Pediatrics*, 152(3):416–421, 2008.

[23] Jason Lazarou, Bruce H Pomeranz, and Paul N Corey. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Journal of the American Medical Association*, 279(15):1200–1205, 1998.

[24] Munir Pirmohamed, Alasdair M Breckenridge, Neil R Kitteringham, and B Kevin Park. Adverse drug reactions. *British Medical Journal*, 316(7140):1295–1298, 1998.

[25] World Health Organization. *The Importance of Pharmacovigilance-Safety Monitoring of Medicinal Products*. World Health Organization, Geneva, 2002.

[26] Rave Harpaz, William DuMouchel, Paea LePendu, and Nigam H Shah. Empirical Bayes model to combine signals of adverse drug reactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1339–1347. ACM, 2013.

[27] Margaret A. Hamburg and Francis S. Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[28] William E. Evans and Howard L. McLeod. Pharmacogenomics drug disposition, drug targets, and side effects. *New England Journal of Medicine*, 348(6):538–549, 2003.

[29] Jill U Adams. Pharmacogenomics and personalized medicine. *Nature Education*, 2008.

[30] M Whirl-Carrillo, EM McDonagh, JM Hebert, L Gong, K Sangkuhl, CF Thorn, RB Altman, and Teri E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012.

[31] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, Alexandra Tang, Geraldine Gabriel, Carol Ly, Sakina Adamjee, Zerihun T. Dame, Beomsoo Han, You Zhou, and David S. Wishart. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 2013.

[32] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(suppl 1):D355–D360, 2010.

[33] Stefan Günther, Michael Kuhn, Mathias Dunkel, Monica Campillos, Christian Senger, Evangelia Petsalaki, Jessica Ahmed, Eduardo Garcia Urdiales, Andreas Gewiess, Lars Juhl Jensen, et al. Supertarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Research*, 36(suppl 1):D919–D922, 2008.

[34] Yanbin Liu, Bin Hu, Chengxin Fu, and Xin Chen. DCDB: Drug combination database. *Bioinformatics*, 26(4):587–588, 2010.

[35] Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A. Clemons, Ru Wei, Steven A. Carr, Eric S. Lander, and Todd R. Golub. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, September 2006.

[36] Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(suppl 2):W623–W633, 2009.

[37] Helge G Roider, Nadia Pavlova, Ivaylo Kirov, Stoyan Slavov, Todor Slavov, Zlatyo Uzunov, and Bertram Weiss. Drug2gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinformatics*, 15(1):68, 2014.

[38] Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in Bioinformatics*, page bbt056, 2013.

[39] Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-Laszlo Barabasi, and Marc Vidal. Drug–target network. *Nature Biotechnology*, 25(10):1119–1126, October 2007.

[40] Justin Lamb. The connectivity map: a new tool for biomedical research. *Nature Reviews Cancer*, 7(1):54–60, January 2007.

[41] Soyoung Lee, Keunwan Park, and Dongsup Kim. Building a drug–target network and its applications. *Expert Opinion on Drug Discovery*, 4(11):1177–1189, November 2009.

[42] Liat Perlman, Assaf Gottlieb, Nir Atias, Eytan Ruppin, and Roded Sharan. Combining drug and gene similarity measures for drug-target elucidation. *Journal of Computational Biology*, 18(2):133–145, February 2011.

[43] Eugen Lounkine, Michael J Keiser, Steven Whitebread, Dmitri Mikhailov, Jacques Hamon, Jeremy L Jenkins, Paul Lavan, Eckhard Weber, Allison K Doak, Serge Côté, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403):361–367, 2012.

[44] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[45] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(1), 2010.

[46] Linyuan Lu and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, March 2011.

[47] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Computational Biology*, 8(5):e1002503, May 2012.

[48] Masahiro Hattori, Yasushi Okuno, Susumu Goto, and Minoru Kanehisa. Heuristics for chemical compound matching. *Genome Informatics Series*, pages 144–153, 2003.

[49] Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. Drug-target interaction prediction through domain-tuned network based inference. *Bioinformatics*, 2013.

[50] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, July 2008.

[51] Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, September 2009.

[52] Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, 29(2): 238–245, 2013.

[53] Yuhao Wang and Jianyang Zeng. Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics*, 29(13):i126–i134, 2013.

[54] Bin Chen, Ying Ding, and David J. Wild. Assessing drug target association using semantic linked data. *PLoS Computational Biology*, 8(7):e1002574, July 2012.

[55] Shobeir Fakhraei, Louiqa Raschid, and Lise Getoor. Drug-target interaction prediction for drug repurposing with probabilistic similarity logic. In *ACM SIGKDD 12th International Workshop on Data Mining in Bioinformatics (BIOKDD)*. ACM, 2013.

[56] MARIE Lindquist and I RALPH Edwards. The who programme for international drug monitoring, its database, and the technical support of the uppsala monitoring center. *The Journal of Rheumatology*, 28(5):1180–1187, 2001.

[57] Marie Lindquist. Vigibase, the WHO global ICSR database system: Basic facts. *Drug Information Journal*, 42(5):409–419, 2008.

[58] Diane K Wysowski and Lynette Swartz. Adverse drug event surveillance and drug withdrawals in the united states, 1969-2002: the importance of reporting suspected reactions. *Archives of Internal Medicine*, 165(12):1363–1369, 2005.

[59] Dianne L Kennedy, Stephen A Goldman, and Ralph B Lillie. Spontaneous reporting in the united states. *Pharmacoepidemiology, Third Edition*, pages 149–174, 2000.

[60] Stephen A Goldman. Limitations and strengths of spontaneous reports data. *Clinical Therapeutics*, 20:C40–C44, 1998.

[61] K Haerian, D Varn, S Vaidya, L Ena, HS Chase, and C Friedman. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical Pharmacology & Therapeutics*, 92(2):228–234, 2012.

[62] M Hauben and A Bate. Decision support methods for the detection of adverse events in post-marketing data. *Drug Discovery Today*, 14(7):343–357, 2009.

[63] June S Almenoff, William DuMouchel, L Allen Kindman, Xionghu Yang, and David Fram. Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. *Pharmacoepidemiology and Drug Safety*, 12 (6):517–521, 2003.

[64] Nicholas P Jewell. *Statistics for Epidemiology*. CRC Press, 2004.

[65] Ola Caster, G Niklas Norén, David Madigan, and Andrew Bate. Large-scale regression-based pattern discovery: The example of screening the who global drug safety database. *Statistical Analysis and Data Mining*, 3(4):197–208, 2010.

[66] Alexander Genkin, David D Lewis, and David Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[67] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[68] Mohamed Rouane-Hacene, Yannick Toussaint, and Petko Valtchev. Mining safety signals in spontaneous reports database using concept analysis. In *Artificial Intelligence in Medicine*, pages 285–294. Springer, 2009.

[69] Rave Harpaz, Herbert S Chase, and Carol Friedman. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics*, 11(Suppl 9):S7, 2010.

[70] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[71] Rave Harpaz, Hector Perez, Herbert S Chase, Raul Rabadan, George Hripcsak, and Carol Friedman. Biclustering of adverse drug events in the fda's spontaneous reporting system. *Clinical Pharmacology & Therapeutics*, 89(2):243–250, 2011.

[72] R Ball and T Botsis. Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS? *Clinical Pharmacology & Therapeutics*, 90(2): 271–278, 2011.

[73] Yuji Zhang, Cui Tao, Yongqun He, Pradip Kanjamala, and Hongfang Liu. Network-based analysis of vaccine-related associations reveals consistent knowledge with the vaccine ontology. *Journal of Biomedical Semantics*, 4(1):33, 2013.

[74] RA Wilke, H Xu, JC Denny, DM Roden, RM Krauss, CA McCarty, RL Davis, T Skaar, J Lamba, and G Savova. The emerging role of electronic medical records in pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 89(3):379–386, 2011.

[75] Jeffrey S Brown, Martin Kulldorff, K Arnold Chan, Robert L Davis, David Graham, Parker T Pettus, Susan E Andrade, Marsha A Raebel, Lisa Herrinton, Douglas Roblin, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiology and Drug Safety*, 16(12):1275–1284, 2007.

[76] Yanqing Ji, Hao Ying, Peter Dews, Ayman Mansour, John Tran, Richard E Miller, and R Michael Massanari. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Transactions on Information Technology in Biomedicine*, 15(3):428–437, 2011.

[77] Jonathan S Schildcrout, Sebastien Haneuse, Josh F Peterson, Joshua C Denny, Michael E Matheny, Lemuel R Waitman, and Randolph A Miller. Analyses of longitudinal, hospital clinical laboratory data with application to blood glucose concentrations. *Statistics in Medicine*, 30 (27):3208–3220, 2011.

[78] Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3): 328–337, 2009.

[79] Ryen W White, Rave Harpaz, Nigam H Shah, William DuMouchel, and Eric Horvitz. Toward enhanced pharmacovigilance using patient-generated data on the internet. *Clinical Pharmacology & Therapeutics*, 2014.

[80] Herman Anthony Carneiro and Eleftherios Mylonakis. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10):1557–1564, 2009.

[81] John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. Digital disease detectionharnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21): 2153–2157, 2009.

[82] David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google flu: Traps in big data analysis. *Science*, 2014.

[83] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the Workshop on Biomedical Natural Language Processing*, pages 117–125. Association for Computational Linguistics, 2010.

[84] Ryen W White, Nicholas P Tatonetti, Nigam H Shah, Russ B Altman, and Eric Horvitz. Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association*, 20(3):404–408, 2013.

[85] Clark C Freifeld, John S Brownstein, Christopher M Menone, Wenjie Bao, Ross Filice, Taha Kass-Hout, and Nabarun Dasgupta. Digital drug safety surveillance: Monitoring pharmaceutical products in Twitter. *Drug Safety*, 37(5):343–350, 2014.

[86] Kanaka D Shetty and Siddhartha R Dalal. Using information mining of the medical literature to improve drug safety. *Journal of the American Medical Informatics Association*, 18(5):668–674, 2011.

[87] George Hripcsak, Carol Friedman, Philip O Alderson, William DuMouchel, Stephen B Johnson, and Paul D Clayton. Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of Internal Medicine*, 122(9):681–688, 1995.

[88] Genevieve B Melton and George Hripcsak. Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association*, 12(4):448–457, 2005.

[89] Jung-Hsien Chiang, Jou-Wei Lin, and Chen-Wei Yang. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using medical language extraction and encoding system (MEDLEE). *Journal of the American Medical Informatics Association*, 17(3):245–252, 2010.

[90] Xiaoyan Wang, Amy Chused, Noémie Elhadad, Carol Friedman, and Marianthi Markatou. Automated knowledge acquisition from clinical narrative reports. In *AMIA Annual Symposium Proceedings*, volume 2008, page 783. American Medical Informatics Association, 2008.

[91] Carol Friedman, George Hripcsak, Lyuda Shagina, and Hongfang Liu. Representing information in patient reports using natural language processing and the extensible markup language. *Journal of the American Medical Informatics Association*, 6(1):76–87, 1999.

[92] Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. MEDEX: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010.

[93] Udo Hahn, K Bretonnel Cohen, Yael Garten, and Nigam H Shah. Mining the pharmacogenomics literaturea survey of the state of the art. *Briefings in Bioinformatics*, 13(4):460–494, 2012.

[94] Caroline F Thorn, Teri E Klein, and Russ B Altman. Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics*, 11(4):501–505, 2010.

[95] Yael Garten and Russ B Altman. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*, 10(Suppl 2):S6, 2009.

[96] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42 (5):760–772, 2009.

[97] Martin Krallinger, Florian Leitner, and Alfonso Valencia. Analysis of biological processes and diseases using text mining approaches. In *Bioinformatics Methods in Clinical Research*, pages 341–382. Springer, 2010.

[98] Tanja Bekhuis. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries*, 3(1):2, 2006.

[99] Jeffrey T Chang and Russ B Altman. Extracting and characterizing gene-drug relationships from the literature. *Pharmacogenetics and Genomics*, 14(9):577–586, 2004.

[100] Pernille Warrer, Ebba Holme Hansen, Lars Juhl-Jensen, and Lise Aagaard. Using text-mining techniques in electronic patient records to identify adrs from medicine use. *British Journal of Clinical Pharmacology*, 73(5):674–684, 2012.

[101] Thomas C Rindflesch, Lorraine Tanabe, John N Weinstein, and Lawrence Hunter. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 517–528. NIH Public Access, 1999.