Data-driven Approach to Detect and Predict Adverse Drug Reactions

Tu-Bao Ho[1,2*], Ly Le[3], Dang Tran Thai[1] and Siriwon Taewijit[1,4]

[1]School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan; [2]John von Neumann Institute, Vietnam National University at Ho Chi Minh City, Vietnam; [3]Department of Biotechnology, International University, Vietnam National University at Ho Chi Minh City, Vietnam, [4]Sirindhorn International Institute of Technology, Thammasat University, Thailand

Tu-Bao Ho

Abstract: ***Background***: Many factors that directly or indirectly cause adverse drug reaction (ADRs) varying from pharmacological, immunological and genetic factors to ethnic, age, gender, social factors as well as drug and disease related ones. On the other hand, advanced methods of statistics, machine learning and data mining allow the users to more effectively analyze the data for descriptive and predictive purposes. The fast changes in this field make it difficult to follow the research progress and context on ADR detection and prediction. ***Methods***: A large amount of articles on ADRs in the last twenty years is collected. These articles are grouped by recent data types used to study ADRs: omics, social media and electronic medical records (EMRs), and reviewed in terms of the problem addressed, the datasets used and methods. ***Results***: Corresponding three tables are established providing brief information on the research for ADRs detection and prediction. ***Conclusion***: The data-driven approach has shown to be powerful in ADRs detection and prediction. The review helps researchers and pharmacists to have a quick overview on the current status of ADRs detection and prediction.

Keywords: Adverse drug reaction, data-driven approach, omics data, social media data, electronic medical records.

Received: March 14, 2016 **Accepted: May 6, 2016**

1. INTRODUCTION

Adverse Drug Reactions (ADRs) can be understood as undesirable effect, reasonably associated with the use of the drug that may occur as a part of the pharmacological action of a drug or may be unpredictable in its occurrence. It is also defined by WHO as “responses to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological function”. ADRs can occur in all settings where healthcare is provided.

It is known that each new drug has gone through a preclinical research and several phases of clinical trials in the pre-market surveillance, but it may still cause ADRs as the drug is clinically tested with at most a few thousands patients and have considerable limitations such as the clinical trials are often done in short time or excluding patients who receive other medications or elderly. As ADRs are major concerns in healthcare, to overcome the shortcomings of pre-market surveillance, the detection and prediction of ADRs in the post-market surveillance play a crucial role in pharmacovigilance (also referred to as drug safety surveillance).

Many factors that directly or indirectly cause ADRs varying from pharmacological, immunological and genetic factors to ethnic, age, gender, social factors as well as drug and disease related ones. It is worth noting that ADR detection and prediction methods intrinsically relate to understanding of the causal factors. Different from ADR detection and prediction methods in pre-market surveillance that are done *in vivo* or *in vitro*, the methods in post-market surveillance are mostly *data-driven* with data collected from the patient drug usage. All data-driven methods depend on two components of data sources and computational methods. In early days the ADR detection methods mainly exploited data from spontaneous adverse event reporting systems or administrative databases with conventional techniques of statistics. On one hand, due to the technology progress, in the last decade several new kinds of data, notably omics data, social media data and electronic medical records (EMRs), have been generating and offering more chances to detect and predict ADRs. On the other hand, advanced methods of statistics, machine learning and data mining allow the users to more effectively analyze the data for descriptive and predictive purposes.

This paper provides a review on the data-driven approach to ADR detection and prediction, and is organized as follows. Section 2 summarizes the causal factors of ADRs, the terminology on ADR study, and common computational methods used in ADR detection and prediction. Sections 3, 4, and 5 review the ADR detection and prediction from omics, social media and EMR data, respectively.

2. ADVERSE DRUG REACTIONS AND DATA-DRIVEN METHODS

2.1. Terminology on Adverse Drug Reactions

Ewards and Aronson [1] analyzed the term “adverse drug reaction” defined by WHO and by some other authors and proposed an adverse drug reaction terminology where adverse drug reaction is defined as “an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medical product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product.” They defined “adverse affect” as “encompasses all unwanted effects; it makes no assumptions about mechanism, evokes no ambiguity, and avoids the risk of misclassification”. Thus the term “adverse reaction” (AR) and “adverse effect” (AE) are interchangeable, except that an adverse effect is seen from the point of view of the drug, whereas an adverse reaction is seen from the point of view of the patient. However, these terms are distinguished from “adverse event” that is “an adverse outcome that occurs while a patient is taking a drug, but is not or not necessarily attributable to it.” Also, the terms “adverse drug reaction” and “drug side effect” are interchangeable while “drug side effect”

*Address correspondence to this author at the School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan;
E-mail: bao@jaist.ac.jp

1381-6128/16 $58.00+.00

(DSE) or “side effect” (SE) is used more commonly among non-health professionals, and it can also cover beneficial unintended reaction [2]. This review adopts the above definitions and concepts.

2.2. Causal Factors of ADRs

There are many factors involving in the pathogenesis of ADRs, typically pharmacological, immunological and genetic factors. Besides, reports suggested that ethnic variation may contribute to the development of ADRs, patient characteristics, drug administration also need to be monitored as agents to cause incidents. A non-chronological, systematic review from 1991 to 2012 [3] subdivided factors affecting the existence of ADRs into four groups.

Patient related factors are the inner, specific characteristics of the patients therefore ADRs are differently developed. Depending on ages, neonates and elderly have the highest possibility of getting undesirable side effects because of their metabolism conditions. Gender is another significant factor because of many biological differences between male and female in metabolism rate, enzymes functions, physical illness. Women on pregnancy have many physiological changes like increasing in blood volume, renal function improve which directly lead to abnormal drug pharmacokinetic rates. Besides, the fetus, which is exposed to any drugs circulating in maternal blood, is very sensitive to drug effects. Creatinine clearance reflects the function of the kidneys that are responsible for the excretion of many drugs, kidney diseases or failures, and therefore have massive impact on ADRs occurrence. Allergy is a genetic factor that happens mostly as type I or type IV reactions because of T-cell-mediated drug hypersensitivity. Lastly, body weight and fat distribution may cause ADRs, especially with fat-soluble drugs on obese and older people who have high proportion of fat in body.

Social factors mostly are behavior of the patient when using drugs. Alcohol drinking affects the metabolism of many drugs and it facilitates the development of ADRs. Taking alcohol with certain drugs can cause many ADRs like nausea, vomiting, headaches, drowsiness, fainting, loss of coordination, hypotension. Chronic alcohol consumption activates enzymes that transform some drugs into toxic chemicals that can damage the liver and other body organs. Ethnic background is a factor controlled by genetics and can determine individual susceptibility to both dose-dependent and dose-independent ADRs. Smoking is one of the risk factors of many diseases like peptic ulcer, cancer and cardiovascular diseases as nicotine can counter the pharmacologic actions of some drugs.

Drug related factors based on the drug itself, mostly from its interactions. ADRs may occur due to drug interaction, synergism, duplication, additive effect, discontinuation of therapy, changing the dose to save money, skipping some medications and physiological antagonism. The causes and significance of drug interactions are multifaceted and include drug dose, serum drug level, route of administration, drug metabolism, duration of therapy, and patient factors, such as age, gender, weight and genetic predisposition. Drug interactions are often classified as either pharmacodynamics or pharmacokinetic interactions. Pharmacodynamics interactions include those that result in additive or antagonistic pharmacological effects while pharmacokinetic interactions involve induction or inhibition of metabolizing enzymes, mostly in the liver.

Disease related factors based on the fact that multiple diseases make patients more vulnerable to ADRs due to the presence of many diseases and the use of many drugs. For example, in patients with renal failure, the effect of drugs on the kidneys is lessened because of the loss of the site of action for these drugs. And drugs that are helpful in one disease may induce the harmful in other diseases.

As a verdict, there are many different factors affect the development of ADRs in different degrees, some of these factors have a direct effect on ADRs but others are insidious. Patients, especially the elderly, should be given serious and intensive attention during the medication to avoid as much risks as possible. Health education, counseling, reconciliation and information technology are tools for pharmacists and health professionals to make any decisions for optimum effectiveness.

2.3. Computational Methods Used for ADR Detection

The relationship between the set D of all drugs and the set A of their known ADR can be represented by as a (bipartite) graph where each drug in D links to a number of ADR and vice versa each ADR is linked to number of drugs. The problem of ADR detection or prediction essentially is the problem of link prediction, i.e. to detect and predict new links between D to A. There are different computational methods that have been used to solve the problem varying from conventional statistics methods to advanced methods in the fields of machine learning and data mining. For the ease of reading articles on ADR detection and prediction by computation for readers who are not familiar with its methods, we briefly describe typical methods and their properties that have been used in ADR detection and prediction.

The traditional statistical methods are basically divided into descriptive ones (with the core is distribution functions) and inferential ones (typically estimation and hypothesis testing). It is worth noting two main features of the conventional statistical methods. One is data that were collected to answer the predefined questions by questionnaire, measurement, observation, etc. and thus their size is usually small, commonly at most several hundreds cases. The other is methods developed to analyze such data, many of them were univariate and created long before we have computers and for small datasets. The multivariate statistics aims to analyze the relations among many random variables, mostly developed around 50 years ago, and divided into two groups: CDA (confirmative data analysis) mostly for hypothesis testing and EDA (explorative data analysis) aiming at generate hypotheses from the data. The EDA methods such as factor analysis, principal component analysis, correspondence analysis, linear discriminant analysis, clustering, regression, and others have been widely employed to solve practical problems. As the cost of data acquisition, storage and processing is dramatically reducing, statistical methods have been greatly changed to adapt to the new situation.

Machine learning and data mining are two closely related fields. Machine learning has its root from the field of artificial intelligence (AI) with the aim to make computers with some learning ability as that of human. This aim is in fact realized by analyzing datasets, typically complex data. Data mining has its root from the practical needs of analyzing large and complex datasets with the aim to discover novel, hidden but precious knowledge in data. Machine learning and data mining are considerably linked to statistics, and share common methods while each field has its own focus and interests. Roughly, machine learning and data mining methods divided into supervised and unsupervised ones.

The supervised methods are mainly for the prediction/classification purpose where prediction models and their parameters are first learned from the training data. The training data contains data of objects that we know each belongs to one or several predefined classes, represented by a class attribute, which is also called labeled data. The learned models are then used to predict the class or classes of unknown objects. Typical supervised learning methods include rule induction, decision trees, neural networks, Naïve Bayesian classification, support vector machines (SVM), regression, etc.

The unsupervised methods are mainly for the description purpose where description models and their parameters are learned from the training data containing objects that we don’t know their classes (unlabeled data). Those learned models allow us to know about properties of the data. Typical supervised learning methods include clustering, association rule mining, trend detection, etc.

As ADR study from social media and EMRs has mostly to deal with textual data, the techniques in natural language processing (NLP) and information retrieval (IR) are widely used.

In the ADR literature, two terms "detection" and "prediction" are often employed. It is necessary to distinguish that the "detection" aims to find previously existed but unknown ADRs from the data (also called "identification" in the literature) while the "prediction" aims to learn new knowledge from data aiming at guessing new adverse reactions when using either a single drug or multiple drugs. The former is mostly carried out by statistics and the latter is mostly by machine learning and data mining.

3. DETECTION AND PREDICTION OF ADVERSE DRUG REACTIONS FROM OMICS DATA

In general, majority of these studies in adverse effects can be classified into two main categories: identification and prediction.

3.1. Adverse Effects Identification

Daly [4] studied the identification by using candidate genes and genome-wide association studies that make contributions of varying extents to each of these forms of reactions are identified to understanding the genetic basis for adverse drug reactions. Many of the associations identified for reactions affecting the liver and skin related human leukocyte antigen (HLA) genes were reported. The other approach that can be complement to omics data is based on information of medical case reports. Gurulingappa *et al.* used MEDLINE (https://www.nlm.nih.gov/bsd/pmresources.html) to extract the adverse effects by employing a Java simple relation extraction system.

The Clinical E-Science Framework (CLEF) initiative shows how to generate semantically annotated medical corpora for information extraction in which yielded robust results [5]. They combined several methods including corpus characteristics, document sampling, annotation guidelines, and annotation methodology and sentence classifier to develop a Benchmark Corpus for adverse effects extraction that was named ADE Corpus [6].

3.2. Adverse Effects Prediction

Kuhn *et al.* introduced SIDER, one public source to predict and investigate adverse effects [7, 8, 9, 10, 11, 12]. The available information included side effect frequency; drug (including placebo) and side effect classifications as well as links to further information are provided in SIDER website, up to now the site has 1430 drugs and 139756 drugs-side effects pairs. Other popular sources included DrugBank for searching the drug information [8, 9, 11, 13, 14] and Pub-Med for pilot studies [8, 9, 11, 12, 13, 14, 15]. Moreover, there are some other sources that are used to predict the adverse effects include Lexicomp [13] and PharmaPendium database from Elsevier [16], both of sources are developed as the drug databases including in vitro, in vivo testing and also clinical trials for drug development strategies. Despite the overlapping of sources, each team has different ways to predict the adverse effects. Their methods range from simple ones analyzing the Pubmed document by Wang *et al.* [15] to have quick results to complexes methods combining many steps by Pouliot *et al.* [14] to exploit analytical process, controlled nomenclature, data sets, normalization of adverse event counts and BioAssay activity and associate adverse events with preclinical assay measurements, screening target specificity, prediction of unrecognized ADRs in marketed drug ingredients, ADR prediction for novel drugs.

Cami *et al.* using the network construction and model prediction to analyze the Pharmacological Network Models [13] while Liu *et al.* use the machine learning methods to predict ADRs on the same data [9]. More complexly, Vilar S. constructed a matrix of drugs and adverse effects to identify and predict unknown effects [11]. La Brute *et al.* use a list of methods including dataset creation, drug-protein target molecular docking calculations using VinaLC, statistical analysis and PubMed text mining to find supportive evidence of ADR-protein associations [8]. Pauwels *et al.* [10] meanwhile use random assignment, nearest neighbors, support vector machines, ordinary canonical correlation analysis. Scheiber *et al.* created the well-established extended connectivity fingerprints comb-ining with Bayesian models and Pearson correlation between the normalized feature probabilities [16]. Izhar *et al.* used molecular docking data to predict the adverse effects [12].

Xiang *et al.* [17] has used UMLS mapping, frequent closed itemsets, uninformative association identification and removal, and statistical validation for efficiently mining multiple drug interactions from Adverse Event Reporting System (AERS) that supports FDA post-marketing safety surveillance program for all approved drugs and therapeutic biologic products. The result showed that their methods returned small *p*-value results and can be improved when cooperated with other external combination.

Integration of different omics data with other data was frequently found in research on side effects. Yamanishi *et al.* use kernel regression, multiple kernel regression (MKR) and canonical correlation analysis (CCA) to analyze the data from SIDER, Pubchem, Drugbank and Matador for predicting side effects [18]. Matador is a drug-protein interaction source where the inclusion of many direct and indirect interactions makes Matador different from other sources. Another drug-protein source was introduced by Mizutani *et al.* using the drug data from DrugBank and SIDER for side effect and protein-drug interaction extraction [19]. They employed ordinary canonical correlation analysis, sparse canonical correlation analysis to predict side effect profiles for new molecules and enrich analyses of targeted proteins to examine the correlation between drug-protein interactions and their side effects on a large scale, without limiting ourselves to proteins of known 3D structures.

Michael *et al.* considered a technique that quantitatively relates proteins based on the chemical similarity of their ligands. They applied a series of methods including molecular descriptor generator, Tanimoto coefficient, similarity ensemble approach (SEA) to construct a random populated pairs of ligand sets and to build an empirical model of background chemical similarity to analyze the data from MDL Drug Data Report [20] The database contains over 132,000 biologically relevant compounds and well-defined deriveactives and have more than 10000 new data added per year. Lee Peters *et al.* developed an approximate matching method for finding the closest drug names within existing RxNorm content for drug name variants found in local drug formularies by using Surescripts test and MEDID test. The drug data was extracted from DrugBank [21]. Lee *et al.* used enrichment scores (ES) calculations, *t*-score calculation and threshold-based filtering to analyze the data from SIDER and Gene Ontology (GO) to find the association of drugs with biological processes [22].

Yamanishi *et al.* focused on drugs targeting four pharmaceutically useful target classes: enzymes, ion channels, GPCRs and nuclear receptors by pharmacological effects prediction from compound chemical structures and Inference of drug-target interactions that are extracted from KEGG, SIMCOMP, JAPIC, SuperTarget and Drugbank [23]. The website description shows that KEGG helps researchers "understand high-level functions and utilities of the biological system from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies". Wienkers *et al.* emphasized on the fundamental factors associated with the early prediction of in vivo DDIs on the basis of kinetic data obtained from in vitro experiments carried out in drug discovery [23]. They employed a series of methods including assessment of CYP inhibition in vitro, generating in vitro inhibition data, biochemical effects on DDI prediction, calculate molecules affecting biology and Genetic variability in DDIs. Unlike Yamanishi *et al.* and Wienkers *et al.*, Bender *et al.* focused on in cell-based screens and suggested a shift from structure-derived chemical descriptors to biological descriptors [24]. GeneGO and Connectivity Map are examples of target prediction in systems pharmacology approach to increase confidence in target and pathway prediction related to phenotypes induce by the compounds of interest. Super-Target is the drug-target interactions database that contains 332828 interactions. SIMCOMP, the website that compares chemical structures in MOL file type by using graph-based methods can be seen as an example.

Xie *et al.* used data from CASTp and ZINC to identify and analyze unknown off-targets for Cholesteryl Ester Transfer Protein (CETP) inhibitors [28]. The analysis consists of 5 steps: Binding site similarity search on a genome scale, Reverse screening of the human structural proteome, Global structure similarity network of off-targets, Volume of the binding pocket, Normalized docking score and Vector distance of the average docking score. Iyer *et al.* worked on Electronic Medical Record (EMR) to identify signals of DDIs from the textual portion of Electronic Health Records [25]. The methods include 2 steps: Preparation of Gold Standard, Annotation of Electronic Health Records and identification of DDI signals. The information about protein target can be analyzed on CAST while the drug data can be collected through ZINC.

Brouwers *et al.* investigated the proportion if side-effect similarities due to targets that are closely to the network compared to shared drug targets by constructing Confidence scores between proteins in the STRING functional protein association database, calculate chemical similarity of drugs, Normalization [26]. White *et al.* introduced a methods using protein biomarker to identify the disease by latent genomic stratification, genetic risk factors for common complex diseases, identification of new disease indications for existing drugs and clinical trial of biomarker [27]. Finally, Shah *et al.* employed phenome-wide association studies (PheWAS), a website that analyzes many phenotypes compared to a single genetic variant (or other attribute) by using the EMR data [29].

NGS platform and relevant technologies have provided an abundant omics data sources for research to gain insights into mechanisms of side and adverse effects at molecular level. The research on side and adverse effects based on omics data not only reduce the unexpected events but also open opportunities for prediction of a drug's unknown molecular activity, thus representing a valuable opportunity in repositioning the drug for a new indication [17].

4. DETECTION AND PREDICTION OF ADVERSE DRUG REACTIONS FROM SOCIAL MEDIA DATA

Not only the OMICS data but also the social media data from medical literature and social networks is a valuable resource for recognizing and extracting the phenotype information of unexpected post-market reactions of drugs. The problem of ADRs detection on the social media data is clarified through several specific tasks such as: Searching for related literature on available medical repositories or collecting patient records to use as evidences for confirming or rejecting knowledge of ADRs or generating and evaluating hypotheses regarding ADRs; Ranking ADRs or side effects (SEs) according to their severity to make a timely warning; Constructing ontologies and lexicon resources to represent medical knowledge; Determining sentences/comments on the social network related to ADRs/SEs which will be useful signals for the ADRs/SEs recognition; Extracting drug-drug interactions and drug-adverse event (AE) relations or patterns indicating ADRs; Mapping terms, phrases extracted from the data to standards lexicon resources.

Section 4.1. shows 8 typical tasks of ADRs detection on the social media data. After that, the researches regarding these tasks are presented in section 4.2.

4.1. Typical Tasks in ADRs Detection on Social Media Data

The social medical data includes medical literature in several public repositories, patient records and comments on forums or social networks that are almost textual data. The medical literature contains the standard medical knowledge, or clinical experiences of pre-market surveillance produced by doctors, pharmacists that is considered the evidences for recognizing and collating the ADRs, ranking the ADRs or building the medical ontologies. Concurrently, it raises several problems in document retrieval. In addition, relying on a group of patient records collected, the hypotheses regarding ADRs can be generated and evaluated. Moreover, the emergence of social network, forums providing a huge textual dataset leads to the problems of Text Mining, Machine Learning (ML), Natural Language Processing (NLP) to automatically exploit such kind of dataset. Eight typical tasks of the ADRs detection on the social media data are as the following:

1. Searching literature concerning ADRs from social medical repositories for confirming or rejecting the know-ledge of ADRs [32, 33, 34, 35, 36, 37, 38].
2. Doing statistical analyses for discovering and evaluating hypotheses related to ADRs [39, 40, 41, 42, 43, 44, 45, 46].
3. Ranking adverse drug events/reactions according to their severity [47, 48, 49, 50].
4. Constructing medical ontologies and medical lexicon resources [51, 52, 53].
5. Comparing adverse events information in different databases/resources [54, 55].
6. Identifying sentences/comments related to potential ADRs [34, 56, 47, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67].
7. Extracting terms, phrases expressing symptoms, syndromes, diseases from a dataset and mapping them to standard lexicon resources [47, 59, 61, 68, 69, 70, 71, 72, 73, 74].
8. Mining drug-drug interactions, drug-AE relations, patterns characterizing ADRs [69, 75, 76, 77, 78, 79, 80, 81, 82, 52, 83, 84, 71, 64, 65, 85, 50, 55, 86, 73].

Although the ADRs detection on the social media data is specified through 8 typical tasks showed above, the boundaries among these tasks are unremarkable. People have not aimed to individually tackle those tasks, the tasks are often combined for a particular purpose.

4.2. Researches in ADRs Detection on Social Media Data

Methods for ADRs detection vary widely, many approaches for pharmacovigilance based on the textual data from medical literature and social networks have been developing. The methods have quickly changed to adapt with the development of the data resources. Several statistical methods are utilized to analyze the medical literature, the reports of patients or physicians or pharmacists gathered and stored in several public repositories such as PubMed, MED-LINE, World Health Organization (WHO). The major characteristics of this data type include fairly homogenous content without lots of noise, and small or medium size that allows a manual analysis. However, the emergence of the social networks with the experiences sharing tendency of patients provided a rich data repository for pharmaceutics researchers to exploit, which makes the classical methods become intractable due to the big size and the noise of the data. Therefore, that demands to develop new computational methods to support detecting and monitoring ADRs on this kind of data. Thus, various data mining, machine learning, Natural Language Processing (NLP) methods have been applying to analyze such kind of data. This section gives a review of methods corresponding to each typical task.

4.2.1. Searching literature concerning ADRs from social medical repositories for confirming or rejecting the knowledge of ADRs

For pharmacovigilance, pharmaceutics researchers or physicians often retrieve the documents related to drugs, their side effects, and symptoms from the social repositories such as PubMed, MEDLINE, Google Scholar.

Table 1. Summary of ADR detection and prediction from omics data.

Reference	Problem	Data	Method
Daly A. K. [4]	Identify adverse effect	Candidate gene and genome wide association studies	• Genes that make contributions of varying extents to each of these forms of reactions are identified.
Gurulingappa H. *et al.* [5]	Identify adverse effect	MEDLINE	• Employ Java Simple Relation Extraction system. • The Clinical E-Science Frame-work (CLEF) initiative investigated how to generate semantically annotated medical corpora for information extraction. • Qualitative evaluation of the system showed robust results
Gurulingappa H. *et al.* [6]	Identify adverse effect	https://sites.google.com/site/adecorpus/	• Characteristic the ADE corpus • Document sampling • Annotation guidelines • Annotation methodology and Modeling a sentence classifier
Kuhn M. *et al.* [7]	Introduce the side effects source	SIDER	• Only include labels from public sources and extend the set of FDA-approved drugs from 746 used in the earlier study to 798. • Drug labels are provided by the FDA and the other sources in two kinds of files: PDF and SPL documents
LaBute M. X. *et al.* [8]	Predict adverse effect	Drugbank, SIDER	
Liu M. *et al.* [9]	Predict adverse effect	SIDER (Side effects data), Pubcheme (drug structure), Drug-bank and KEGG (biological properties)	• Machine-learning-based
Pauwels E. *et al.* [10]	Predict adverse effect	SIDER	• Random assignment, Nearest neighbor, Support vector machine, Ordinary canonical correlation analysis
Vilar S. *et al.* [11]	Predict adverse effect	SIDER, DrugBank	• Matrix construction
Wallach I. *et al.* [12]	Predict adverse effect	SIDER, pubchem, PDB, KEGG	• Docking and Inference
Cami A. *et al.* [13]	Predict adverse effect		• Construct data on drug-ADE associations, drug and ADE taxonomies, and intrinsic drug properties. • Network representation of the drug-ADE associations contained in the 2005 data-base snapshot is constructed
Pouliot Y. *et al.* [14]	Predict adverse effect	Pubchem BioAssay	• Overview of analytical process • Controlled nomenclature • Data sets, Normalization of adverse event counts and BioAssay activity • Associating adverse events with pre-clinical assay measurements • Screening target specificity • Prediction of unrecognized ADRs in marketed drug ingredients, ADR prediction for novel drugs
Wang W. *et al.* [15]	Predict adverse effect	PubMed	• PubMed search • Document classification and Drug-ADE classification
Scheiber J. *et al.* [16]	Predict adverse effect	PharmaPendium data-base from Elsevier	• The well-established extended connectivity fingerprints • Bayesian models built, computing the Pearson correlation between the normalized feature probabilities from the individual Bayesian models

(Table 1) Contd….

Reference	Problem	Data	Method
Xiang Y. *et al.* [17]	To improve the adverse effects mining	Adverse Event Reporting System (AERS)	• UMLS Mapping • Frequent Closed Itemset Mining • Uninformative Association Identification and Removal and Statistical validation
Yamanishi Y. *et al.* [18]	Predict side effects	SIDER, Pubchem, Drug-bank and Matador	• Kernel Regression, Multiple Kernel Regression (MKR) • Canonical Correlation Analysis (CCA)
Mizutani S. *et al.* [19]	Examine the correlation between drug - protein interactions and their side effects on a large scale, without limiting ourselves to proteins of known 3D structures	Drugbank, Matador (drug-protein interaction), SIDER (side effects)	• Ordinary canonical correlation analysis, Sparse canonical correlation analysis • Prediction of side effect profiles for new molecules and Enrichment analyses of targeted proteins
Michael J. Keiser *et al.* [20]	Considers a technique that quantitatively relates proteins based on the chemical similarity of their ligands	MDL Drug Data Report (MDDR)	• Molecular descriptor generator • Tanimoto coefficient • Similarity Ensemble Approach (SEA), construct random populated pairs of ligand sets to build an empirical model of background chemical similarity
Lee Peters *et al.* [21]	Develop an approximate matching method for finding the closest drug names within existing Rx-Norm content for drug name variants found in local drug formularies	RxNorm, Drugbank,	• Surescripts test and MEDID test
Lee S *et al.* [22]	Find the association of drugs with biological processes	SIDER, Gene Ontology (GO)	• Enrichment scores (ES) calculations, t-score calculation • Threshold-based filtering
Wienkers L.C. *et al.* [23]	Focus on the fundamental factors associated with the early prediction of in vivo DDIs on the basis of kinetic data obtained from in vitro experiments carried out in drug discovery and to highlight issues that can confound the success of this undertaking.		• Assessment of CYP inhibition in vitro • Generating in vitro inhibition data • Biochemical effects on DDI prediction • Calculate Molecules affecting biology and Genetic variability in DDIs
Andreas Bender *et al.* [24]	Focuses on two aspects of chemogenomics: In cell-based screens and a shift from structure derived chemical descriptors to biological descriptors		• Summarizes these “performance” descriptors and their applications • Focusing on gene expression profiles and high-content screening data
Iyer S. V. *et al.* [25]	Identifying signals of DDIs from the textual portion of Electronic Health Records	Electronic Health Records (EHRs)	• Preparation of Gold Standard • Annotation of Electronic Health Records • Identification of DDI signals
Brouwer L. *et al.* [26]	Investigate the proportion if side-effect similarities that us due to targets that are close in the network compared to shared drug targets	DrugBank, Matador, PDSP Ki, SIDER and STRING	• Confidence scores between proteins in the STRING functional protein association • Database, calculate chemical similarity of drugs, Normalization
White T.J. *et al.* [27]	Protein biomarker for disease identification		• Latent genomic stratification • Genetic risk factors for common complex diseases • Identification of new disease indications for existing drugs and Clinical trail of biomarker

(Table 1) Contd….

Reference	Problem	Data	Method
Xie L. *et al.* [28]	Identify and analyze a panel of unknown off-targets for Cholesteryl Ester Transfer Protein (CETP) inhibitors	CASTp, ZINC	• Binding site similarity search on a genome scale • Reverse screening of the human structural proteome • Global structure similarity network of off-targets • Volume of the binding pocket • Normalized docking score and Vector distance of the average docking score
Shah N. H. [29]	Summaries about PheWAS	The phenome-wide association study (PheWAS)	
Baker N. C. *et al.* [30]	Explored the potential of using side effect profiles of drugs to predict their bioactivities at the receptor level.	ChemoText, The PDSP Ki data-base (version kidb100-108)	• Compile and curate the modeling datasets • Build and validate statistical models that predict the molecular activity from side effect profiles and perform virtual screening of a large set of chemicals
Yamanishi Y., *et al.* [31]	Focus on drugs targeting four pharmaceutically useful target classes: enzymes, ion channels, GPCRs and nuclear receptors	KEGG (Brite, Genes, Drug and Ligand), SIMCOMP, JAPIC, SuperTarget and Drug-bank	• Prediction of pharmacological effects from compound chemical structures • Inference of drug-target inter actions

To automate the search of literature to verify side effects, Avillach *et al.* [32] constructed the queries for searching the publications in MEDLINE including: the drug, the adverse effects, two subheadings “adverse effects”, “chemical induced” that are mapped to MeSH. After that, they determined a threshold of a number of the extracted publications to confirm whether a given drug/adverse event association has been already known in the literature. Similarly, in [35], confirming or rejecting ADRs are also relied on the documents acquired from MEDLINE.

Lardon *et al.* made their review of the manual ADRs identification and the automatic ADRs extraction from the social media that is considered a new source of knowledge for pharmacovigilance by acquiring the relevant articles from PubMed, Embase, Google Scholar in [36].

The shifts in public health conditions over time can be tracked by considering the fluctuation in the frequent terms sets extracted from Twitter which are used to search the articles from Wikipedia. These articles reflect the health conditions [37].

In order to synthesize the data from published studies and international experiences to identify the evidences of the potential benefits and drawbacks of patient’s reporting for ADRs, Blenkinsopp *et al.* [33] used a structured search on the online repositories such as MEDLINE, CINAHL, and PsyclNFO. In addition, to investigate whether the case reports of statin associated with memory loss, Wagstaff *et al.* utilized available published reports from MedWatch and MEDLINE as the evidences to confirm or against such relation [38].

4.2.2. Making Statistical Analyses for Discovering and Evaluating Hypotheses Related to ADRs

Through the spontaneous patient reports collected from several public repositories, pharmacists or physicians can create and test hypotheses regarding drugs, and side effects. In [43], Cable J. reported the severe side effects of statins by manually gathering and collating the information from the self-reporting data of 351 patients. Additionally, to determine an association between the hospital admissions and the ADRs, after extracting the literature from Cumulative Index to Nursing, Allied Health Literature, EMBASE, MEDLINE, Kongkaew *et al.* [44] determined the ADR prevalence rates of the hospital admission that are calculated as the number of patients admitted to the hospital with at least one ADR divided by the total number of patients admitted to the hospital during the study period. After that, the heterogeneity (the differences) among the ADR prevalence rates was assessed via X^2 and I^2 test.

Cohort study is a form of longitudinal study (a type of observational study) widely used in medicine, social science, business analytics, ecology. A cohort is a group of people who share a common characteristic or experience within a defined period. Several works in ADRs detection used this method. Pal *et al.* [46] proposed a cohort study of adverse events associated with one or more medicines. The cohort study is also designed in [41, 42] to describe the prevalence, the types, the consequences of adverse drug events (ADEs).

Lazarou *et al.* [45] estimated the incidence of serious and fatal adverse drug reactions in hospital patients. Through selecting 39 prospective studies from US hospitals, and analyzing with a random-effects model by two investigators, then computing the overall incidence by combining the incidence of ADRs occurring while in the hospital with the incidence of ADRs causing admission to the hospital, they found that this incidence is extremely high. In other work [39], the frequency of medication errors associated with adverse events is manually evaluated via self-reports of pharmacists, nurse reviews of all patient charts, reviews of all medication sheets. To generate signals for possible unrecognized hazards from spontaneous adverse drug reactions, Evans *et al.* [40] used the Proportion Reporting Ratio (PRR), a statistical aid which involves the comparison of all reactions to a drug for a specific condition of interest with other drugs in the databases.

4.2.3. Ranking Adverse Drug events/Reactions According to the Severity

Ranking ADRs according to their severity plays an important role in drug safety that provides assessments of risks and benefits of drugs, and gives a timely warning for physicians in their treatment. In [48], the order of ADRs based on their serious level was built from the pairwise comparisons of those ADRs which are manually assigned. Moreover, Trifiro *et al.* [49] created a ranked list of high-priority events to deal with the increase of the number of spurious signals that is based on the scientific literature, the medical textbooks, the websites of regulatory agencies. This work was done by two groups of experts who ranked the events according to five criteria, and a consensus score will be obtained in case of a disagreement. In EU-ADR project which aims to exploit the different Europe electronic healthcare record databases for drug safety, creating a ranked list of the events that are deemed to be important in pharmacovigilance is a crucial part [47].

4.2.4. Constructing Medical Ontologies and Medical Lexicon Resources

Medical ontologies, or lexicon resources are important for a medical knowledge representation that is useful for analyzing the data, and looking up the information. There are several works attempting to construct the medical knowledge resources. Liu *et al.* [52] identified associations between a drug class and its side effects by building a hierarchical ontology. This ontology of the side effects is constructed by clustering the extracted side effects (such as "*elevated blood pressure*", "*increase in blood pressure*") into 307 synonym groups which are further grouped into 30 classes (such as "*eyes*", "*blood*", etc). After that, with this ontology, they can quantify the associations between the side effects and the statin drugs via the log-likelihood ratio. In EU-ADR project, eight Electronic Healthcare Records databases in Europe were combined for a large-scale drug safety monitoring [51]. Additionally, the EU-ADR annotated corpus was built up by Mulligen *et al.* [53] by using the Named Entity Recognition system to make the annotations of drugs, drugs disorders, genes, then revising these notations by annotators.

4.2.5. Comparing Adverse Events Information in Different Database Resources

To examine whether the published AE data is different from those in the sponsor's database or not, Scharf *et al.* [54] searched for the literature in National Cancer Institute, and Clinical Data Update System to make a comparison between the AE data in the trial publication with the AE data submitted by the investigators to CDUS. In addition, Xu R *et al.* [55] compared the drug-SE pairs extracted from J Oncology (JCO) tables to those derived from the FDA drug labels.

4.2.6. Identifying Sentences/Comments Related to Potential ADRs

Different from the medical literature, the data from social networks is almost free text which contains lots of irrelevant sentences or comments (noise). Therefore, it causes a problem that how to determine sentences/comments which mention ADRs. That can be treated as a text classification problem. Many methods were proposed to deal with this problem.

In [56], to detect drug users and potential adverse events in the Twitter messages, Bian *et al.* built two binary Support Vector Machine (SVM) classifiers to classify such messages. The first classifier was used to determine whether the users themselves or someone they know has taken the drug or not, the second one was used for identifying which messages containing the side effects on the result obtained by the first one. The features set was built based on two types of features: "textual feature" that constructs a specific meaning in the text such as bag-of-words, the number of words indicating the negation etc.: "ontological/semantic feature" that expresses the existence of semantic properties such as the number of CUIs in each "Semantic Type", etc extracted from Unified Medical Language System (UMLS). In addition, Chee *et al.* [57] treated the problem of unknown ADR identification as a sentiment classification problem and utilized Naïve Bayes (NB) and Support Vector Machines (SVM) with RBF kernel method to classify with unigram, bigram, trigram features. In order to improve the performance of the classification, they used multiple classifiers built on different subsets of total training data, then aggregated those through the bagging (bootstrap aggregation) approach. Due to the purpose of unknown ADRs detection, the evaluation concentrates on the false positive error which occurs when a negative sentence is incorrectly assigned as a positive one. Similarly, the works in [58, 61, 62, 63, 66] performed the classification tasks on the messages from Twitter or DailyStrength to determine whether they have ADR signals or not with the features set built by N-grams and UMLS. After that, the words or phrases were mapped to the formal terminologies to determine expressions indicating the ADR signals.

In [67], Yang *et al.* showed an interesting strategy to enrich the annotated data by detecting a consensus in the communication which provides a greater impact than the individual's adjustment. That made the clusters of messages according to their topics. After that, a similarity evaluation algorithm was used to assign the unlabeled data set to the available informative clusters acquired from previous step. SVM is also applied to separate the messages into positive and negative ones. Not only SVM but also Logistic regression was utilized to classify the documents in PubMed into two groups including the "related-ADR" documents and the "non-related-ADR" documents before making a classification of drug-ADEs to identify which drug causes the ADEs [64]. Moreover, after filtering the messages by SVM, Jiang *et al.* [60] grouped the comments into the reasonable clusters that present the main outcomes of the drugs. This work was done by using semi-supervised Probabilistic Latent Semantic Analysis (PLSA), and the generated topics are guided by the standard outcome descriptions provided by experts for the rationality of those topics. Duda *et al.* [34] evaluated the capability of SVM as a method for locating the articles about drug-drug interaction through making a comparison with the use of queries for the PubMed search.

4.2.7. Extracting Terms, Phrases Expressing Symptoms, Syndromes, Diseases from Dataset and Mapping them to Standard Lexicon Resources

Detecting and mapping words, phrases expressing the symptoms, syndromes, diseases extracted from a data set to the formal terminologies provide a groundwork for solving text mining problems. The words and phrases detection can be based on many available lexicon resources.

To extract the potential drug effects, Jiang *et al.* [59] used MetaMap to recognize the words, phrases indicating the symptoms, the syndromes, and the diseases, then collated with known drug effects. In [70], after selecting a set of generic drugs related to the anesthesiology domain, Levin *et al.* mapped them to their trade names based on the "tradename_of" relationship in RxNorm.

In [61], to extract concepts, Apache Lucene, one of high-performance search engine libraries, was utilized for both indexing and retrieving the ADR lexicon concepts. The Lucene index is generated from the concepts and the associated UMLS IDs.

Yates *et al.* [72] enriched the MedSyn synonym set by an annotated breast cancer drug review dataset. After that, they extracted ADRs by identifying the words and the phrases appearing in their MedSyn synonym set constructed in previous step. In addition, for gauging the experiences of medical devices and drugs by patients with the diabetes mellitus, Akay *et al.* used the Self-organizing maps (SOM) to numerically analyze forum posts to extract a list of words related to positive and negative in [68].

Various machine learning methods and NLP techniques have been applied for the medical words and phrases extraction which is treated as the Named Entity Recognition (NER) problem. In [71], Sampathkumar utilized the lexicon-based NER tool to extract drug names, and terms denoting the side-effects or phrases indicating a relationship between the drugs and the side effects. Conditional Random Field (CRF) is a popular machine learning technique often used for the NER tasks. Yates *et al.* [73] used the dependency parsing to build a dependency graph among terms, each path in the graph presents a candidate of the ADRs. CRF will classify those paths into two groups "FOLLOW" and "DON'T FOLLOW", and the path with the label "FOLLOW" will be selected as a ADR. CRF was also applied for identifying terms indicating the drugs, and the symptoms in [69].

4.2.8. Mining Drug-Drug Interactions, Drug-AE Relations, Patterns Characterizing ADRs

An important work after extracting the words, the phrases indicating the drugs, the symptoms is to discover the relations among such words and phrases which characterize ADRs, or the cause of the ADRs. Several researches are conducted to automatically explore the drug-drug interactions or drug-AE.

The mining drug-ADRs can be treated as a classification problem with the feature set built by NLP methods that aims to determine which drug an ADR or a side effect or a symptom is associated with, or identify a drug-ADR relation is positive or negative. To link the ADRs/symptoms to drugs, there are several works doing classification for the extracted ADRs with the labels are drug names. In [64], after identifying the articles related to ADE, Wang *et al.* used logistic regression to classify these articles with each likely drug in a set of 25 drugs. In addition, given a drug and a side effect, H. Wu [65] formulated the problem of predicting whether the side effect is related to the drug as a binary classification problem. If the side effect is related to the drug, the classifier will assign the label of "relevant". The authors utilized both discriminative classification method and generative modeling method to solve with the training data set based on the information about the drugs from the knowledge base. CRF was also applied to find the most likely label (the drug name) for a ADR that the ADR is caused by this drug [73]. In this work, the features set is constructed from the linguistic information such as part-of-speech, and dependency relations among the terms. In a similar approach, Aramaki *et al.* [69] used two methods pattern-based method and SVM-based method to make the decision which drug caused which symptom.

The targets in [82, 83] are different from the works presented above a little bit. In these works, the authors attempt to classify the relations between the drugs and the adverse drug events into positive and negative by using the SVM with the kernel function. The features set for the relation instances are generated by the dependency parsing to keep the syntactic and semantic information for those instances through identifying the shortest path from the medical events to the treatment entities.

The relations between drugs and AEs can be represented as association rules of entities indicated by words and phrases that allow us to utilize the association rules mining techniques to explore such relations. In [76], the association rules of pairs of terms extracted from the corpus were used to evaluate the capability of co-occurrence of the drug-event pairs. In order to signal the infrequent patterns characterizing ADRs, Jin *et al.* [79] proposed a domain-driven knowledge representation that is Unexpected Temporal Association Rule with two measures for the rules evaluation such as "leverage" that indicates how strength of temporal association, and the interestingness measure "unexlev". They also introduce two mining algorithms MUTARA, HUNT. In the other work [84], since a sentence is a transaction of words belonging to this sentence, the author can find the association of the words by a method including two main steps: Frequent Rule Identification that extracts the combination of words with a mention of ADR; Frequent Pattern Generation that creates the patterns from the extracted rules which can be applied for new sentences to find new mentions of adverse effects. In [86], Yang *et al.* utilized the association rule mining with several typical measures such as "support", "confidence", "leverage" to evaluate the relationship between the drugs and the adverse events. Moreover, Li *et al.* [81] treated the problem of finding risk patterns in the medical data as an optimal rule discovery problem that requires a method to deal with a huge number of rules generated with the low efficiency because of the low support constrain. Relying on the definition of an optimal risk pattern and its anti-monotone property, they proposed the algorithms which post-prune an association rule set to find the optimal ones. In [77], Harpaz *et al.* applied Apriori algorithm to identify the interaction between the drugs and the adverse effects in the spontaneous reporting systems.

Not only classification methods, and association rules mining methods are used to deal with the drug-ADRs mining problem, but also other machine learning techniques and statistical scores for quantifying the relation between drugs and adverse reactions have been utilized. In [78], Jiang *et al.* aimed to develop and evaluate the Semantic Web-based approach for mining the severe drug-drug interaction (DDI) and the induced adverse events (ADE) based on available medical knowledge databases with a statistical work on the EMR data. To do so, after extracting the pairs DDI-ADE with their corresponding AERs outcome codes acquired from the FDA Adverse Event Reporting System - a database of adverse events and medication errors, the signals of the DDI-ADE (the capability of co-occurrence of DDI-ADE) are enhanced through the EMR data with a metric to measure the signal enrichment of the DDI-ADE. In the next step, to identify the serious level of the DDI-ADE pairs, they mapped the AERs outcome codes to the Common Terminology Criteria for the Adverse Event (CTCAE) system to get the CTCAE grades which indicate the level of severity.

Leaman *et al.* [80] extracted the adverse drug reactions from the user comments by utilizing a lexicon-based method that compares a sliding window of tokens from the comments with the lexical terms. The spelling error problem was also handled via a similarity score between two terms. Moreover, the task of extracting the adverse side effects of the drugs can be treated as a sequence labeling problem using Hidden Markov Model in [71]. In this work, the causal relationship between the drug and the side effects or the co-occurrence ability of the drug and the corresponding side effect is formulated as a probabilistic function of states. In [85], the conditional relationship of the drug-gene pairs is extracted by the "drug-gene conditioned" method which classifies sentences in MEDLINE based on the co-occurrence of the known drug-gene pairs before extracting the pair of drug-gene in such sentences. The drug-gene pairs can be extracted from the unclassified sentences via the "unconditioned" algorithm, a simple co-occurrence based method. Xu *et al.* [50] used the syntactic parsing to extract pairs of drugs and side effects from the biomedical literature. The sentences were parsed into the corresponding trees which contain the known drug-SE pairs, the patterns indicating the drug-SE relation can be retrieved by searching on those known drug-SE pairs. In addition, in [55], the authors also performed their work of extracting the drug-SE pairs on the tables downloaded from J Oncology (JCO). After classifying the tables into two groups "SE-related" and "SE-unrelated", the drug-SE pairs are extracted from the "SE-related" tables then compared with those derived from the Food and Drug Administration (FDA) for a quality evaluation. They also measured the correlation between the drug side effects and the drug targets, the metabolism and the indications to investigate the potential value of the anticancer drug-SE pair for the understanding of the drug toxicity prediction, and the drug repurposing.

Bayesian Confidence Propagation Neural Network was applied to quantify the dependency between drugs and ADRs in the WHO databases by Bate *et al.* in [75]. They aimed to compute the strength or the weakness of all drug-ADR dependencies in the data set via a neural network architecture with the Information Component (IC) measures used as the weights of the network. Another measure to quantify an association between drugs and symptoms was showed in [52]. In this research, the log-likelihood ratio was used to estimate the association between the drugs and the symptoms. The log-likelihood ratio is often used to compare the fit of two models one of which called "null model" and the other called "alternative model", and helps to make the decision to reject null model or not. In this case, the author considered the side effect association problem as a coin toss model with the null model which states that a side effect belongs to a drug with the same probability, and the alternative model is in contrast mean. Therefore, the null model will be reject if they found that the side effect indicated by a phrase appears on the reviews of a drug more frequently than on the reviews of the other drugs. That means the alternative model is better to fit with the data set.

Table **2** gives a summary of the methods used in ADRs detection on the social media data.

4.3. Popular Data Sets and Lexicon Resources, NLP Tools for ADRs Detection on Social Media Data

The social media data provides a lot of valuable in-formation to recognize novel ADRs in drug safety.

There are many repositories have been created for collecting the reports from patients, physicians, pharmacists. One of such repositories is MEDLINE/PubMed that is a large source containing lots of journal citations and abstracts for the biomedical literature. Additionally, the SIDER is also useful to provide the information of adverse drug reactions extracted from the public documents. The WHO databases contain over 2.5 million case reports. Moreover, the emergence of the social network and forums such as Twitter, Facebook [89] or DailyStrength, Yahoo!Group about Health & Wellness have been enriching the dataset for ADRs detection.

The medical lexicon resources are an inseparable component in most of methods to explore the medical data. One of popular medical lexicon sources is Unified Medical Language System Metathesaurus that contains one million concepts and five million concept names combined from various databases, and provides the semantic information of medical terms/phrases. Another source is ICD10 that includes the codes of symptoms and diseases.

Since the media data is almost textual data, many NLP tools are used to analyze such data. MetaMap [87, 88] integrates some NLP tools such as the syntactic parsing, the NER, the abbreviation correction to analyze a sentence and mapping the terms or phrases of the sentence to their semantic types. In addition, the Natural Language Toolkit (NLTK), the Stanford Name Entity Recognizer, the Stanford Dependency Parser are widely used in many researches.

5. DETECTION AND PREDICTION OF ADVERSE DRUG REACTIONS FROM ELECTRONIC MEDICAL RECORDS

The digitalized clinical patient's information in EMR is a collection of multi-disciplinary data elements, typically *clinical data* (patient's admission and discharge summaries, doctor daily notes, nurse narratives, medications, etc.) and *para-clinical data* (laboratory test results, X-rays, images, etc.) [90]. Basically, an EMR system is the warehouse to capture all aspects of patient care data in electronic format [91]. This emerging fashion of the data management technology conspicuously benefits to diagnosis support, evidence-based medicine, drug safety surveillance as secondary uses, etc. The advantages include not only the high reliability repository regarding terminology, controlled vocabulary and nomenclature code, facilitated relational structure for effortless data acquisition, longitudinal patient care and outcome, but also flourishing positive and negative patient's risks observations to assess the safety and efficacy of a drug. The favorable merit of EMRs differs from other sources mentioned above in the sense that those are less trustworthiness and lacking in clinical sensibility, and also medical literature, spontaneous report ordinary fall into biased data corresponding to only passive outcome monitoring [92]. These overwhelmed drawbacks make extremely intriguing to recent researches.

Dealing with clinical text, the text representation of EMR in machine-readable form and information extraction are highlight challenge. ADR studies mostly handle textual data with common subtasks in text mining [93, 94]. The main task of NLP [95] is to transform free text to machine-readable data (e.g., lexical analysis, part-of-speech (POS) tagging, NER, word-sense disambiguation (WSD), negation identification, etc.). Then, domain knowledge integration copes with the heterogeneous and massive scale of data through clinical semantic enhancement [96] (e.g., UMLS, ICD, SNOMED CT, Med-DRA, etc.). Hence, its result can augment knowledge representation or extensive comprehension for translational bioinformatics [97]. The subsequently process is relation extraction. The process mainly uncovers an underlying association in the substantial unstructured document using versatile techniques (e.g., co-occurrence analysis, machine learning method, rule-based approach).

In this section, we review the research on using EMRs in detecting and predicting ADRs. The section is organized by starting from the prominent EMR data characteristics, the research questions, materials, and current efficient methods.

5.1. The Prominent EMR Data Characteristics

Due to the vast replaced paper-based by EMR systems, the massive growth of data facilitates to effortless data collection, but it still has limited utilizations, especially unstructured data. The large-scale sources of EMR data for ADR are certainly the doctor daily notes and nurse narrative notes. The enormous values of the invisible ADR information can feasible derive from this underlying data, which are proven by many researches. Nursing document or nursing narrative facilitates a real-time or synchronous patient's status as a recording of a timely log and a summary at the end of shift. The deliverable messages are including the observable patient's health situation, assessment, plan, and recommendation to a next shift. Even though the nursing narrative contains huge redundant data, but there are major advantages for patient monitoring and harmful changed status detection given a certain condition. For example, given the changing a dose of medicine and observe the patient's response.

On the other hand, discharge summary is a primary deliverable document to support communication among health professional teams in the hospital [147]. The content is recorded as a free text that summarizes a patient's hospitalization. Apart from the current admission information, significant finding, procedures and treatment, prescription medication, laboratory test and result, it also conveys family history, illness history, and the follow up instruction. Unlikely nursing document, the discharge summary mostly captures the non-redundant and significant data instead of log data. The utilization is found in many researches [106, 108, 109, 110, 111, 135] by deploying NLP technique to explore the potential ADR from this type of EMR document.

Another dominant note, radiological report, contains a radiology imaging, which derived from an advanced imaging technology, and further free text data consolidation. A diagnostic radiologist, who specializes in the interpretation of these images, can take advantage of radiology imaging for diagnostic and disease treatment. The remaining free text in the report narrates the reason of examination, underlying medical condition including the summarization of radiology examination and interpretation as a final report. This beneficial interpretation of radiology and patient's condition information can contribute towards the ADR signal detection as well.

5.2. Research Questions and Gaps

Many advanced EMR-related researches fall interested in ADR detection. In order to support the automated tool, the effectiveness method is required to tackle the data containing in EMRs. There are remaining of big gaps herein that challenge the researchers. Firstly, the existing methods are inapplicable for EMR due to the distinct of data characteristic of EMRs (e.g. ungrammatical, short phrases, abbreviations, acronyms, etc.) from typical text in articles or literatures. The adaptive NLP methods and their applications are required to deal with the clinical text [98, 99]. Secondly, the exploitation of large corpora of medical terms is crucial in understanding meaning of EMRs terms. The semantics computable is favored as a modern solution to preserve meaning relatedness [99]. Thirdly, the relation extraction to capture the rare events is inefficient. This is because the major characteristic of ADR that the prescription by a physician should be safe. Moreover, the latent confounding factors

Table 2. Summary of ADR detection and prediction from social media data.

Reference	Problem	Data & Tool	Method
P. Avillach et al. [32]	Automate the search of literatures concerning adverse drug reactions to confirm whether the drug/events has been already known in the literature	MEDLINE, UMLS	• Construct the query including drug, adverse effect, two subheading: AE, chemical induced which are mapped to MeSH for literature searching. • Determine a threshold of a number of literatures to confirm the knowledge of drug/event association.
A. Blenkinsopp *et al.* [33]	Synthesize the data from published studies and international experience to identify the evidences of the potential benefits and the drawbacks of direct patient reports of suspected adverse drug reactions	MEDLINE, CINAHL, PsyclNFO	• Structured search of MEDLINE, CINAHL and Psycl-NFO supplemented by internet searches and requests for the information to the key contacts
S. Duda *et al.* [34]	Evaluate the classification capability of SVM as a method for locating articles about drug-drug interactions	MEDLINE	• Make the comparison between the text classification using SVM vs. the queries through the PubMed to identify articles relating to drug-drug interactions.
N. Garcelon *et al.* [35]	To automate the search of publications that correspond to a given Adverse Drug Reaction case	MEDLINE	• Defining a general pattern for the queries used to search MEDLINE. • Determining a threshold of a number of publications to confirm or infirm the adverse drug reaction.
J. Lardon *et al.* [36]	A scoping review was undertaken to explore the breadth of evidence about the use of the social media as a new source of knowledge for pharmacovigilance	PubMed, Embase, Google Scholar for extracting relevant articles	• Determining the scope of review following some research questions. • The relevant articles are extracted from PubMed, Embasem Google Scholar. • Two pairs of reviewers independently screened the selected studies and proposed two themes: manual ADR identification; automatic ADR extraction from the social media.
J. Parker *et al.* [37]	Tracking the public health condition trends via Twitter	Twitter, Wikipedia articles	• Use frequent term sets from tweets as queries to search Wikipedia articles, the retrieval of articles is treated as a health-related condition. • Observing fluctuations in frequent term sets, they can detect the shifts in the public health conditions over time.
L. R. Wagstaff *et al.* [38]	To review case reports of statin-associated memory loss as well as the available published evidence to confirm or against such link	MedWatch, MEDLINE	• Searching the MedWatch drug surveillance system for the reports of statin-associated memory loss. • Reviewing the published literature using the MEDLINE and the prescribing information for these drugs
D. W. Bates *et al.* [39]	To evaluate the frequency of medication errors using a multidisciplinary approach. Determine how often medication errors are associated with adverse drug events.		• From the self-reports of pharmacists, nurse reviews of all patient charts, and reviews of all medication sheets, incidents thought to represent ADEs or potential ADEs are classified by two independent reviewers.
S. J. W. Evans *et al.* [40]	Generateing signals for possible unrecognized hazards from the spontaneous adverse drug reaction reporting data.	ADROIT database	• Using the Proportional Reporting Ratio (PRR), a statistical aid which involves the comparison of the proportion of all reactions to a drug, which are for a particular medical condition of the interest with all other drugs in the database.

(Table 2) Contd….

Reference	Problem	Data & Tool	Method
A. J. Forster *et al.* [41]	To describe the incidence, the severity, the preventability of adverse events affecting patients after discharge from the hospital, and to develop strategies to improve the patient safety during this interval	400 consecutive patients discharged home from the general medical service.	• Prospective cohort study. • Get the information for analysis by making a cases summary for every patient's posthospital course, and making a telephone interview approximately 3 weeks after discharge. • Use this information to create the event summaries through the basic statistical methods.
J. T. Hanlon *et al.* [42]	To describe the prevalence, types, and consequences of adverse drug events in older outpatients with polypharmacy	The data is collected by health services of Durham Veterans Affair Medical Center. The data includes the medical and sociodemographic data, the self-perceived health status, the prescribers, ADE histories.	• Utilize the basic statistical methods.
Cable J. [43]	Reporting the severe side effects of statins through the statistical works on a collection of patient's self-reported information	Text of e-petition which is sent to the World Health Organization	• Gathering and collating manually the information from the self-reported data of 351 patients. • Make the charts to show the obtained results.
C. Kongkaew *et al.* [44]	To determine the prevalence of hospital admissions associated with ADRs and examine the differences in the prevalence rates between the population groups and the methods of ADR detection	Cumulative Index to Nursing and Allied Health Literature, EMBASE, and MEDLINE	• Data extraction. • compute the prevalence measure and do some statistical analyses. • After that they make the data synthesis.
J. Lazarou *et al.* [45]	To estimate the incidence of serious and fatal adverse drug reactions in hospital patients	MEDLINE, Excerpta Medical, International Pharmaceutical Abstracts, Science Citation Index	• Study selection: Selecting 39 prospective studies from the US hospitals. • The data is extracted by 2 investigators and analyzed by the random-effects model. • Combine the incidence of ADRs occurring while in the hospital and the incidence of ADRs causing the admission to hospital to compute the overall incidence
N. S. Pal *et al.* [46]	• To quantify and characterize the risks to individual and communities from the drugs to minimize the harm and improve the use. • To track the problems due to medication errors and poor quality medicines		• Cohort Event Monitoring (CEM) is a prospective, observational, cohort study of adverse events associated with one or more medicines. • Targeted Spontanous Reporting (TSR) that builds on the principles of the spontaneous reporting.
A. Fourrier-Reglat *et al.* [47]	EU-ADR project aims to exploit the different Europe electronic healthcare records databases for the drug safety signal detection		Project framework: • Create a ranked list of the events that are deemed to be important in pharmacovigilance • Detect the drugs that are potentially associated with these events via data mining techniques
A. Gottlieb *et al.* [48]	Rank adverse drug reactions according to their severity	SIDER 2, FDA Legacy AERS, Human Phenotype Ontology (HPO)	They used Internet-based crowsourcing to rank ADRs according to their severity with the following steps: • Assigning the pairwise comparisons of the ADRs. • Using these comparisons to rank the order of the ADRs.

(Table 2) Contd….

Reference	Problem	Data & Tool	Method
G. Trifiro *et al.* [49]	To create a ranked list of high priority events to deal with the increase of the number of spurious signals	EU-ADR project http://www.euadr-project.org	• To review the scientific literature, the medical textbooks, the websites of regulatory agencies to create a preliminary list of the events that are deemed important in pharmacovigilance • Two groups of experts independently rate each event on five criteria. • A consensus score is obtained in case of disagreement.
R. Xu *et al.* [50]	To extract drug-side-effect (drug-SE) pairs from the vast amount of published biomedical literature	MEDLINE, US Food and Drug Administration (FDA)	• Using the syntactic parsing to generate a syntactic tree corresponding to each sentence. • After that, they extracted the syntactic patterns associated with the drug-SE pairs from the trees. • They also developed the patterns-ranking algorithms to prioritize those patterns, then selected a set of patterns with both high precisions and recalls to extract the drug-SE pairs from the text.
P. M. Coloma *et al.* [51]	Combining the electronic healthcare records (EHR) databases in Europe for large-scale drug safety monitoring	8 databases from 4 countries Denmark, Italy, Netherlands, UK	• Aggregate the demographic, and the prescription from 8 databases, then pool them by using a distributed network approach. • Make a comparison of incidence rates of UGIB and NSAID utilization to evaluate the data harmonization and the quality across databases.
J. Liu *et al.* [52]	Identifying associations between a drug class and its side effects from drug reviews on the health-related web sites	drug discussion forums: AskPatient.com, Medications.com, WebMD.com	• Extract a complete set of side effect expressions from the drug reviews. • Use some statistic and heuristic methods to build up a hierarchical ontology of side effects by aggregating the patient-submitted drug reviews. • Use the log-likelihood ratio to quantify the associations between drugs and symptoms.
E. Mulligen *et al.* [53]	Building the corpus	EU-ADR corpus	• Using the named-entity recognition system to make annotations and revise this annotation by annotators. • The corpus has been annotated for drugs, disorders, genes and their inter-relationships.
O. Scharf *et al.* [54]	To examine whether the published AE data differ from those in the sponsor's database and from the data collection requirements stated in study protocols	National Cancer Institute (NCI), Clinical Data Update System (CDUS)	• Comparing the AE data in the trial publication with the AE data submitted by the investigators to CDUS.
R. Xu *et al.* [55]	Aim to extract the drug-SE pairs from a large number of high-profile oncological full-text articles	J Oncology (JCO)	• Classifying the tables downloaded from the JCO into two classes: SE-related, and SE-unrelated. • Extracting the drug-SE pairs from the SE-related tables. • Comparing the drug side effect knowledge extracted from the JCO tables to that derived from the FDA. • Analyzing relationships between anti-cancer drug-associated side effects and drug-associated gene targets, metabolism genes and disease indications.

(Table 2) Contd….

Reference	Problem	Data & Tool	Method
J. Bian et al. [56]	Detect drug users and potential adverse events through the content of twitter messages	Twitter, UMLS Metathesaurus, FDA's Adverse Event Reporting System (AERS), Tool for crawl Tweets: Twitters's user timeline API	• The mining is presented through two binary classification problems: • Building a classifier to identify the drug users. • Building a classifier to identify the side effects caused by the drugs of interest.
B. Chee *et al.* [57]	Exploitation of the post-marketing surveillance information about the drugs to identify the unknown ADRs by aggregating the individual opinions and the review of crowd opinions on the online health forums	Data: Health & Wellness Yahoo!Group, Adverse drug lexicon from MedDRA	• Defining the watchlist of drugs which needs to make surveillance. • Making the sentiment classification with Naive Bayes and SVM, focusing on the false positive error. • To improve the performance, using multiple classifiers with the data permuted before running, then aggregated those classifiers through the bagging approach.
R. Ginn *et al.* [58]	Determining adverse drug reactions in comments of twitter	Twitter, UMLS	• Select a set of drugs to be monitored. • Crawl the data from Twitter, then do the preprocessing: remove advertisements, tweets contain URL. • Manually annotate the data for a binary classification with two classes "has ADR" and "noADR", also identify the span of expressions that convey ADRs then map them to the formal terminologies in UMLS. • Do the binary classification with NB and SVM.
K. Jiang *et al.* [59]	Automatically extract the potential drug effects from the Twitter data	Twitter data, Natural Language Processing Tookits (NLTK), MetaMap, MedLinePlus	• Collecting and preprocessing the drug-related tweets. • Classifying the retrieved tweets with two classes: "personal experience" and "non-personal Experience" by NB, SVM, Maximum Entropy. • Identifying drug effects from the tweet text with the help of MetaMap. • Collating the extracted drug effects with the known drug effects.
Y. Jiang et al. [60]	Clustering patient outcomes by effectively digesting large volumes of the personal health messages	Yahoo! Groups	• Filtering the messages to remove news and advertisements by SVM. • Grouping the sentences in the messages with similar topics into clusters, the Probabilistic Latent Semantic Analysis (PLSA) guided by the expert knowledge.
.O' Connor Karen *et al.* [61]	Aim to present a systematic study of tweets collected for drugs to assess their value as the source of potential signals for ADRs	Corpus of ADRs: http://diego.asu.edu/downloads UMLS	• Data acquisition. • Annotation: with two stages "hasADR", "noADR" for a binary classification, and make the annotations of the symptom, the syndrome, or the disease based on the lexicon resources such as UMLS, etc. • Automatic Concepts Extractions: Use of lexicon-based techniques.
A. Patki *et al.* [62]	Detecting adverse drug reactions and categorizing drugs.	DailyStrength, ComScore, SIDER, IMS Health, WordNet, SentiWords, NLTK toolkit, Weka	• Classifying user comments to determine whether they contain ADRs or not • Combining the classification probabilities (combining the comments for a drug) to classify the drug to the normal or black-box categories.

(Table 2) Contd….

Reference	Problem	Data & Tool	Method
A. Sarker *et al.* [63]	To automatically detect adverse drug reactions mentioned from text	Twitter, DailyStrength	• To Make a binary classifier to determine ADRs by generating a large set of features to represent the semantic properties. • Additionally, they combined the training data from the different corpora to improve the accuracy of classification.
W. Wang *et al.* [64]	Automatically determine whether a specific adverse event is caused by a specific drug	PubMed citations	• Searching documents in the PubMed. • Using logistic regression to classify the documents into two groups ADE-related and non-ADE-related. • Logistic regression is still used to classify the articles acquired from previous step associated with the set of drugs considered.
H. Wu *et al.* [65]	Investigate the feasibility of the discussion exploitation to discover unrecognized drug side effects	DailyMed, Drugs.com, SIDER	• Constructing a Drug Side Effect Knowledge Base. • Detecting Unrecognized Drug Side Effects by making a binary classifier to identify whether the side effect is related to the drug or not. The classifier is built based on both discriminative method and generating modeling method.
C. Yang *et al.* [66]	To detect adverse drug events using Letters to the Editor	Corpus II (1664 LtEs), UMLS, MetaMap	• Make a binary classifier to determine the candidate is ADE or not with the features set generated by using MetaMap, n-grams.
M. Yang *et al.* [67]	To automatically extract and classify the messages posted on social media networks into positive (ADR-related) and negative (non-ADR-related)	Web forums: ProzacAwareness, SSRIsex are forums in Yahoo! Groups about Health & Wellness	• Detecting the consensus through the topics and make the clusters associated with those topics, then use a similarity algorithm to assign the unlabeled set to these informative clusters. • Construct the classifier using SVM.
Akay. A *et al.* [68]	Gauge the experiences of medical devices and drugs by patients	Dataset: forum Diabetes-Daily (http://www.diabetesdaily.com/forum/) Data mining tools: Rapidminer (www.rapidminer.com)	• Using the Self-organizing maps (SOM) to analyze forum posts numerically to extract a list of words related to positive and negative.
E. Aramaki *et al.* [69]	Extracting adverse drug events and effects from the clinical records	The discharge summaries gathered from University of Tokyo Hospital	• The symptom terms, drug terms identification is solved by using CRF. • The relation identification that determines which drugs caused the symptom is solved by both pattern-based method and SVM-based method.
M. Levin *et al.* [70]	Extracting drug names in the free text and normalizing them by mapping to the standardized nomenclatures.	RxNorm, Philips Medical Systems, Andover, MA	• Selecting a set of generic drugs related to the anesthesiology domain, make a cutoff point below which usage of any particular generic drug became uncommon in the electronic records. • Map the generic drugs to their trade names based on the “tradename_of” relationship in the RxNorm.
H. Sampathkumar *et al.* [71]	The objective of this research is to extract reports of adverse drug side-effects from messages in the online healthcare forums and use them as early indicators to assist in post marketing drug surveillance	Dataset: www.medications.com	• They treated the task of extracting adverse side-effects of drug as a sequence labeling problem using Hidden Markov Model (HMM).

(Table 2) Contd….

Reference	Problem	Data & Tool	Method
A. Yates *et al.* [72]	Detecting expected and unexpected adverse drug reactions from the consumer reviews on the social media sites	askapatient.com, drugs.com, drugratingz.com, SIDER, UMLS	• Create an annotated breast cancer drug review dataset. • Generating a comprehensive ADR synonym set focused on breast cancer (MedSyn). • Extracting ADRs by identifying terms and phrases that appeared in the MedSyn synonym set.
A. Yates et al. [73]	Extracting adverse drug reactions (ADRs) from the forum posts and linking the extracted ADRs to the drugs that users claim are responsible for them	Breastcancer.org, MedSyn, Stanford syntactic parsing	• Extracting ADRs: Using the dependency relation and CRF to extract. • Link drugs to ADRs: Using CRF to label the ADR's caused by the drug.
E. Yom-Tow et al. [74]	To monitor adverse drug reactions in a single drug and in the drug combinations.	Yahoo U.S Web search engine, ICD10	• They used the aggregated search data of a large population of Internet users to extract the information related to the drugs and adverse reactions to them, and the correlation of this data over time. • They extended their method for identifying the adverse reactions in the combinations of drugs.
A. Bate *et al.* [75]	Highlighting the dependency between drugs and adverse reactions in the WHO database	WHO database	• Using Bayesian Confidence Propagation Neural Network. • Using the information component (IC) measure as the weight of the neural network.
A. Benton *et al.* [76]	The paper presents two problems: Collecting a corpus of the medical message boards posts, de-identification, extracting the information on ADR; Using the corpus to identify the capability of co-occurrence of drug event pairs	Breastcancer.org, komen.org, Cerner Multum's Drug Lexicon, Stanford NER, CRF, NLTK.	The system is constructed with 4 steps: • Downloading message post pages from sites and remove the unrelated contents. • De-identification. • Developing the controlled vocabulary. • Identifying terms in the corpus. Association rules is used to evaluate the capability of co-occurrence of pair of terms.
R. Harpaz *et al.* [77]	Identifying drug interaction adverse effects (DIAEs) in the spontaneous reporting systems	FDA, MedDRA terminology, MedLEE	• Association rule, using Apriori algorithm.
G. Jiang *et al.* [78]	The objective of this study is to develop and evaluate the Semantic Web-based approach for mining the severe drug-drug interaction (DDI) and the induced adverse drug events (ADE)	FDA Adverse Event Reporting System (AERS), Common Terminology Criteria for Adverse Event (CTCAE), SIDER, World Wide Web consortium (W3C), Mayo Clinic.	They used a normalized AERS dataset with the following steps: • Extracting DDI-ADE pairs with their AERS outcome codes. • Make a filtering pipeline comprising 3 datasets: AERS-DM, SIDER 2, PharmGKB dataset to ensure that the reported ADEs could not be explained by a single drug effect. • Enhance the signals of the DDI-induced ADEs through the EMR data by an enrichment score between DDI and ADEs. • Mapping between the AERs outcome codes and the CTCAE grades, then classifying the filtered DDI-ADEs into the CTCAE (grades corresponding to level of severity).
H. Jin *et al.* [79]	Signaling/highlighting infrequent patterns characterizing ADRs	Pharmaceutical Benefits Scheme (PBS) database, Medicare Benefits Scheme (MBS) database, Queensland Linked Data Set (QLDS)	• The author proposed a domain-driven knowledge representation, Unexpected Temporal Association Rule with the measure unexlev and a mining algorithm MUTARA. • The improved algorithm, HUNT, is used to highlight the infrequent and unexpected patterns.

(Table 2) Contd….

Reference	Problem	Data & Tool	Method
R. Leaman *et al.* [80]	Automatically, extracting and evaluating relationships between drugs and adverse reactions in the user posts to the health-related social network websites	DailyStrength, UMLS, SIDER, Canada Drug Adverse Reaction database, MedEffect.	• Annotate data by drugs names, and concepts. • Do preprocessing data. • Comparing the token in a sliding window with tokens in the lexicon, handling the spelling error though a similarity score between two terms.
J. Li *et al.* [81]	Finding risk patterns in the medical data		• Define the risk patterns by a statistical metric, relative risk. • The problem is treated as an optimal rule discovery problem.
H. Liu *et al.* [82]	Proposing an analytical framework for extracting patient reported adverse drug events from the online patient forums.	U.S. Food and Drug Administration (FDA), FDA Adverse Event Reporting System (FAERS), DailyStrength, PatientsLikeMe, diabetes online community.	• Data collecting, and preprocessing • Medical Entity extracting: lexicon-based approach • Adverse drug event extracting: They proposed a kernel based learning method to extract ADEs in the patient medical forums with two step: relation extraction, relation classification into positive and negative.
X. Liu *et al.* [83]	Identifying patient reported adverse drug events.	American Diabetes Association (ADA) online community, UMLS, FAERS, Consumer Health Vocabulary, MetaMap	• Use the shortest-dependency path kernel function with SVM for classifying the relations instances into positive and negative • Features for the relations instances are generated based on the dependency to keep the syntactic and semantic information for the relations instances by determining the shortest path from the medical events to the treatment entities.
A. Nikfarjam *et al.* [84]	Extracting mentions of adverse drug reactions from user reviews about drugs in the social network websites by mining a set of language patterns	DailyStrength, CO-START vocabulary (a subset of the UMLS Metathesaurus), Canada Drug Adverse Reaction Database, MedEffect.	• Applying the association rule mining on a set of annotated comments to extract the underlying patterns of colloquial expression about the adverse effects.
R. Xu *et al.* [85]	Aim to extract the conditional relationship of drug-gene pairs from the text using the known drug-gene pairs as prior knowledge	MEDLINE, ThinTek	They developed two methods to extract the PGx-specific drug-gene pairs from MEDLINE sentences. • The algorithm “Unconditioned” is a simple co-occurence based method in which the drug-gene pairs are extracted from the unclassified sentences. • “Drug-Gene Conditioned” classifies the sentences (in MEDLINE) based on the occurrence of the known drug-gene pairs before extracting the pair of drug-gene.
C. Yang *et al.* [86]	Mining associations between drugs and adverse reactions from the user constributed content in the social media.	PatientsLikeMe, MedHelp, Facebook, Twitter	• Using the association rule mining with the measures such as support, confidence, leverage to evaluate the relationship between drugs and adverse events.
A. R. Aronson [87]	Describing the algorithm used by MetaMap, and its applications	MetaMap	MetaMap algorithm: • Text Parsing • Variant Generation • Candidate Retrieval • Candidate Evaluation • Mapping Construction

(Table 2) Contd….

Reference	Problem	Data & Tool	Method
A. R. Aronson *et al.* [88]	Reporting on MetaMap's evolution over more than a decade	MetaMap	
X. Marinela *et al.* [89]	To test whether forming a group on Facebook focused on ADRs would lead to the discovery and reporting of ADRs experienced by members of the group	Facebook	

under patient's condition lead to the false positive when considering the causal relation analysis. Lastly, the gold standard corpus for a large-scale assessment has no exist. Therefore, a manual review of top *n* high probability results by physicians is need for performance evaluation endorsement.

5.3. Materials and Methods

5.3.1. Data

According to the increasing of healthcare systems, the digitalized healthcare data becomes accessible with patient privacy procedure such as de-identification [**100**] to prevent a patient's confidential information. Table 3 shows the available sources of electronic clinical and healthcare data for research purpose. There are not only inpatient and outpatient records from patient's history in hospitals, but also simulated data that are available for the study. The Observational Medical Outcomes Partnership (OMOP) [**101**] is the simulated data for studying the effects of medical products. The data expresses over 10 years with 10 million persons with 90 million drug exposures. Another the promising healthcare data, Medical Information Mart for Intensive Care (MIMIC) II **[102]**, was developed by the MIT lab for computational physiology. The data was collected from the ICUs of Beth Israel Deaconess Medical Center from 2001 to 2008. It contains two main data components: (1) Clinical data such as narrative notes, disease diagnosis, medication prescription, laboratory test result, and (2) Physiological waveforms. MIMIC II version 2.6 contains 32,536 patients, hospital admission over than 36,000 cases, and ICU stays over than 40,000 cases. Recently, the PhysioNet has launched MIMIC III which augments the newly data from 2008 - 2012. During our review, the latest version is 1.3. There are over 58,000 hospital admission and 46,520 patients. Furthermore, the encouraging of healthcare improvement is found in some challenges. Informatics for Integrating Biology and the Bedside (I2B2) under an NIH-funded National Center for Biomedical Computing (NCBC) launched the challenge in clinical NLP. In addition, the subtask challenge from SemEval and ShARe/CLEF provides stratified data related to their task including clinical discharge summary corpus, which is derived from MIMIC II. Fortunately, the source of data is available for public access under data sharing policy, and now it becomes popular in the academic medical and computer science departments. On the other hand, some organizations manipulate their own repositories such as STRIDE, NYPH EHR, the Stockholm EPR corpus, VUMC, etc. that contain a huge number of de-identified patient records. To deal with the big data on such rich patient's information repository, the requested knowledge, training, and collaboration for both of clinicians and data scientists have been addressed, for example, see Moskowitz *et al.* [103].

5.3.2. EMR as a Single Source of Data for ADR Detection

A variety of data types in EMRs is a beneficial source for ADR study. The facilitated structured data such as patient's demographics, laboratory test results, medication information, billing codes, or diagnosis codes (ICD), etc. are straightforwardly accessible for data exploration and analysis. In the contrast, more complicated source of data, but remaining reliability including engaging history over the time can be derived from clinician providers such as discharge summaries, nurse notes, medical doctor notes, etc. The mixture utilization among various kinds of data is also challenges. Table 4 shows the research study on EMR for ADR analysis. In the works of Liu *et al.* [119] and Park *et al.* [120], they use only structured data such as medical prescription information, laboratory test results, etc. to analyze adverse drug reaction. This type of data can straightforwardly derive from the database. Alternatively, [109, 110, 115, 122] utilize free text from clinical narratives with NLP techniques to gain more useful information under longitudinal healthcare notes. The hybrid structured and unstructured data from EMR is also appealing for this research area [123, 124].

5.3.3. EMR as a Complementary Repository for ADR Detection Enhancement

In pharmacology study for drug safety, additional integration of heterogeneous across data set is dramatically fashionable [108, 114, 125, 126, 127, 128]. The challenge is the data integration techniques to deal with the variety of data formats. The underlying hypothesis is that the blending of various data characteristics derived from a particular domain (e.g., biological, chemical, genetic, phenotypic, etc.) is more significant benefit to ADR detection than an individual dataset. The identical proof of concept as above is interchangeable to any medical text-based analysis. The utilization of conjugated multidisciplinary data for signal detection was described in several researches [104, 111, 126, 128]. In the work of Li *et al.* [104] and Harpaz *et al.* [111], they employed a framework to consolidate plausible signals discovered from a widely used spontaneous reporting system and complementary observational healthcare data. In [111], the large scale of over 4 millions of spontaneous reports and 1.2 million of narrative notes including structured data from laboratory tests were analyzed to uncover three serious harmful evidences. On the one hand, [104] utilized the similar data characteristics. The clinical data was combined with drug data from extensive repository of insurance claim rather than the laboratory test. Then the integrated data was used in the process of ADR signal prioritized score adjustment. In their work, the four ADRs of interest were straightforwardly recognized from a structured billing code or diagnosis code (ICD-9 /10). In the contrast, Harpaz *et al.* [111] employed NLP techniques on the rich clinical text. In their work, ADRs was recognized from clinical text rather than structured billing code. The authors claimed that the unexpected reaction derived from the healthcare data over the time is more granular than billing codes, and also unbiased data towards reimbursement.

5.3.4. Combinatorial Genome-Wide Association Study (GWAS) and EMR for Genotypic-Phenotypic Relation Comprehension

The expanding of integration across domain is addressed in Jensen PB *et al.* [126] and Wei WQ *et al.* [129]. In their work, the incorporation between clinical observation and genetic information presents the feasibility in practically drug-disease-gene network analysis. In Karnes JH *et al.* work [130], the complementary of EMRs and genetic data was proposed to reveal the AR of a genetic variant that induces the immune thrombosis due to the administra-

Table 3. The electronic clinical data sources have been made available for the adverse reaction research.

Reference	Data source	Description	Data Characteristic
Li Y *et al*. [104]	CCAE[1]	MarketScan Commercial Claims and Encounters (USA) provide information on pharmacoepidemiologic data sources for use in epidemiology, health services research, healthcare economics.	Claim data, symptom and diagnosis data.
Li Y *et al*. [104]	GE EHR[2]	GE Healthcare MQIC (Medical Quality Improvement Consortium) database.	A longitudinal outpatient population, and captures events in structured form that occur in usual care, including patient problem lists, prescriptions of medications, and other clinical observations as experienced in the ambulatory care setting.
Harpaz R *et al.* [105], Doan S, *et al.* [106]	I2B2[3]	Informatics for Integrating Biology and the Bedside (i2b2).	The files contain a random selection of 100,000 records for each of 97 common lab tests, for a total of 9.7 million records.
Suominen H, *et al.* [107]	MIMIC II [102]	Multiparameter Intelligent Monitoring in Intensive Care, Intensive Care Unit (ICU) patients.	Structured data—medical, surgical, coronary care, neonatal, laboratory test, disease diagnosis, etc.; unstructured clinical narratives—medical note, nurse note, discharge summary; waveform data.
Chen ES *et al*. [108], Wang X *et al*. [109], Wang X *et al*. [110] Harpaz R *et al.* [111], Li Y *et al.* [104]	NYPH [112]	New York Presbyterian Hospital at Columbia University Medical Center.	Containing of 1.2 million narrative notes; discharge summaries, operative reports, and reports from numerous ancillary services (e.g., radiology and pathology).
Duan L *et al*. [113], Li Y *et al*. [104]	OMOP[4]	Observational Medical Outcomes Partnership; An observational healthcare databases simulated data for studying the effects of medical products [101].	10 million persons; 90 million drug exposures; 5,000 different drugs; 300 million condition occurrences; 4,500 different conditions; over a span of 10 years; only 1.8% of the 20 million possible drug-condition combinations (population statistic from [113]).
Liu M *et al*. [114], LePendu P *et al.* [115]	STRIDE	The Stanford Clinical Data Warehouse.	Containing of 1.6 million patients; 15 million encounters; 25 million coded ICD-9 diagnoses, and a combination of pathology, radiology, and transcription reports; over 9.5 million unstructured clinical notes over a period of 17 years (population statistic from [115]).
Zhao J *et al*. [117], Henriksson A *et al.* [118]	The Stockholm EPR Corpus [116]	The electronic patients record from Karolinska University Hospital.	Over 512 clinical units; over 2 m patients; structured data-age, gender, ICD-10 diagnosis code, drugs, laboratory result, admission and discharge time; unstructured data—clinical narratives.
Liu M *et al*. [119]	VUMC	The Vanderbilt University Medical Center.	Inpatient and outpatient, clinical information, laboratory values, imaging and pathology reports, billing codes, and clinical narratives; 1.9 million patients with highly detailed longitudinal data for about 1 million.
Liu M *et al*. [119], Yoon D *et al*. [120]	Korean tertiary teaching hospital clinical database	Korean tertiary teaching hospital clinical database.	32,033,710 prescriptions; 115,241,147 laboratory tests; 1,011,055 hospitalizations; 530,829 individual patients (Jan 2000 - Mar 2010).

[1] http://www.bridgetodata.org/node/987
[2] http://www.emr.msu.edu/Documents/mqic_main.htm
[3] https://i2b2.org/
[4] http://omop.org

tion of heparin. The diversity of ICD-9 diagnostic codes and laboratory results from narrative text are employed using NLP techniques to identify the case of interest associated heparin-thrombosis treatment. Hence, DNA information extracted by discarded blood sample from the Vanderbilt DNA databank was linked to the case observation outcome.

To recognize an AR of interest from EMRs, most researches implement one or more from three following approaches: (1) Evidently determines ADRs from laboratory ab-normality results [119, 121] (2) Straightforwardly derives ADRs from a structured billing code or diagnosis code (ICD-9 or 10) [104, 124, 131] (3) Hardly extracts ADRs from clinical notes [108, 111]. Even though, the direct acquisition from the structured data such as ICD code is rather easy, but the unexpected reaction derived from longitudinal healthcare notes is more granular than billing codes, and also unbiased data towards reimbursement [111, 132].

5.4. Methodology

5.4.1. Terminology Normalization

The majority of clinical data generated in EMRs is in the form of unstructured text requiring complex preprocessing methods to support searching, summarization, decision support, or statistical analysis [98]. Even though the clinical narratives in EMR system is recorded by clinicians, they probably encounter problems of domain specific terms, misspelling, un-grammatical, abbreviations, telegraphic phrase, synonyms, acronyms, especially, when written by free text. The automated information extraction is needed to avoid errors and reduce human efforts. Most clinical text mining studies relevant to deal with term annotation or NER as a preprocessing step. BioMedLEE [133]– a NLP system to extract a broad variety of phenotypic information from the biomedical literature– and MedLEE [134]– a growing resource providing a map of NLP systems and research in the medical domain– are widely used in biomedical annotation [108, 109, 135]. Recently, the clinical NLP system cTAKES [136], developed by Mayo Clinic, is become prominent in the community [137, 138]. The system supports multidisciplinary components such as sentence boundary detector, tokenizer, normalizer, POS tagger, chunker, annotater, negation detector, dependency parser, etc. Many clinical NLP systems can result various types of entities with normalization of UMLS [139] concept. The system contains over 12.8 million concept names from 3.2 million concepts. The vocabularies integrate in the UMLS metathesaurus from 153 source families including MedDRA, MeSH, NCBI Taxonomy, SNOMED CT, etc.

5.4.2. ADR Signal Detection and Prediction

There are multidisciplinary methods along the variety of data types and the objectives. Chen ES *et al.* [108] explored two NLP systems, BioMedLEE and MedLEE, and MeSH descriptor to annotate drugs and disease concepts from the biomedical literature and its complementary data source such as discharge summaries from EHRs of New York Presbyterian Hospital. In this work, eight diseases are examined and the relevant Medline articles and discharge summaries are collected to investigate. An evidence-based method is inspected for co-occurrence analysis of the disease and drug relations where the chi-square ($\chi2$) statistic and its p-value are used to test the hypothesis of no association. Finally, the comparative associations from Medline using MeSH and UMLS, discharge summary using UMLS, are discussed to see the overall agreement. According to their experiment, the interesting finding of this combinatorial source of data is that the literature data source contributes to the disease-drug association through the long time therapeutic observation monitoring, and the discharge summary source transmits the current practice corresponding drug prescribing regarding certain conditions. Other two works from Wang and co-workers [109, 110] are rather similar. In [110], the experimental setting mainly separates drugs of interest into (1) long-time drugs in the market but short term side effect, (2) drugs with new adverse event were detected after 2004, and (3) one drug class. The annotation using MedLEE and filtering of confounding factors are deployed. Then the classical co-occurrence analysis is applied to find the strong association. Despite considering only strong drug-disease association, in [109] extensive examines drug-symptom, and disease-symptom derived from clinical narratives to investigate both of direct and indirect associations in which the indirect association can identify by mutual information (MI) and data processing inequality DPI.

Another data type that widely used in pharmacovigilance regarding ADR prevention is the laboratory test results. In the work of Liu *et al.* [119], the utilization of this structured data is considered in the study group and comparison group given drug administration. Then the comparison between patients who is exposed to the abnormal laboratory results and normal one in both groups when administered the specific drug during hospitalization is examined using a two-way contingency table. The several signal measures not only notable $\chi2$ but also PRR, ROR, Yule's Q (YULE), BCPNN, and GPS are used to discover potential drug-adverse reaction pairs. The enhancement by combining of heterogeneous sources of data is found in Harpaz *et al.* [111]. The top k common high ranked drug-adverse reactions from both AERS and EHRs are investigated and evaluated by in-house reference set.

The pattern of patient-specific AE occurrences provides phenotypic information related medication and diagnosis. Recently, Roitmann *et al.* [123] addressed the patient stratification method based on AE profiles derived from clinical narratives. In their work, for each cluster, two methods of AE co-occurrence analysis and weighted edges network analysis were employed to support the comprehensive of phenotypic association from the retrospective observations. The well known term-frequency and inverse document frequency (tf-idf) is applied to fine elucidate cluster characteristics such as common drugs prescription or common diagnosis. Further, the data analytics and visualization are exploited to better understand the patient profiles. This graphical representation can enhance comprehension, support treatment procedure, and benefit diagnosis decision-making.

For temporal related pattern, Liu *et al.* [140] examined comention and drug-first fraction along time dependency to build the features for discriminative model of ADR and drug-indication. The co-occurrence count and its complementary logarithm were considered to fit the SVM model. The data validation was performed with 100-fold cross validation on 1,550 samples of 980 indications and 570 adverse events. From their results, the temporal ordering of the drug-first mentions data representation seems to be promising to achieve the high performance.

Most recently, data mining and machine learning methods can potentially drive ADR analysis. Yildirim *et al.* [142] deployed clustering technique to investigate adverse reactions and allergy (ARA) from antibiotics in children. In their work, the highlight results from data mining techniques can be useful for clinicians to support effective decision-making. The work of Li *et al.* [104] applied LASSO, least absolute shrinkage and selection operator, to investigate confounding factors related to ADR that lead to false positive decision. Similar to the work of Peissig *et al.* [131], the ILP is deployed to generate feature representation using the logic programming. Then comparative machine learning classifiers such as random forest, SMO, PART, J48, and JRIP are built on such representative feature to assess the model performance. Alternatively, Liu et al. [140] produced the representative feature using temporal events with co-occurrence analysis, then applied SVM to discriminate the drug-adverse event pairs from the drug-indication pairs.

For ADR prediction, Karlsson *et al.* [143] trained two tree-based and rule-based classifiers, random forest and JRIP respectively, with the sparse vector of 1,312 different drugs, 9,863 different diagnosis codes, age and gender. They set the experiment into 6 different models of the combination between drug and diagnosis features to investigate the feasible of the machine learning method.

Table 4. The research study in adverse drug reaction detection and prediction using electronic medical records.

Reference	Problem	Materials and Study Population	Method Type	Method
Li Y. *et al.* [104]	ADR detection	• EHR from NYP/CUMC and GE MQIC—including admission notes, discharge summaries, lab tests, structured diagnosis (ICD-9) codes and structured medication lists. • Claim data, CCAE. • Spontaneous reports from FAERS in the 2004-2010. • 4 clinical serious adverse reactions of interest—acute renal failure, acute liver injury, acute myocardial infraction, upper GI bleed • STITCH, MedDRA.	Machine learning	• Deploying LASSO to obtain the confounding adjusted signal score for each drug-ADR pair from NYP/CUMC and FAERS. • Normalizing ADR signal scores using p-value; • Combining calibrated signal scores between couple of FAERS and each EHR source to investigate that the complementary source combination enhances the outcome instead considering of single source, FAERS, whether or not. • Evaluation using reference standard developed by OMOP.
Chen E. S. *et al.* [108]	Identify strong association among drug and disease	• Discharge summary in the 2003-2004 (48,360 reports). • Medline article in the 2006 (81,828 related articles). • To investigate all drugs related to 8 diseases of interest—Acquired immunodeficiency syndrome, Asthma, Breast neoplasms, Congestive heart failure, Diabetes mellitus, Parkinson's disease, Pneumonia, Schizophrenia. • NLP tools—BioMedLee, MedLEE.	Co-occurrence	• For each data source, drug and disease are annotated and mapped to highest-level MeSH descriptor (for Medline) and UMLS concept (for both Medline and discharge summary) using NLP systems. • Co-occurrence and χ^2 statistic are applied to compute and evaluate the association strength among eight diseases of interest and relevant drugs. • The association acquired from Medline using MeSH, Medline using UMLS, and discharge summary using UMLS are compared to see the overall agreement. • The medical expert is used to manual review the top five disease-drug associations.
Wang X. *et al.* [109]	To characterize phenotypic and environmental associations obtained from clinical reports	• Discharge summary from NYPH (25,074 reports) in the 2014. • 1,997 unique drug concepts in scope. • 732 unique symptom concepts in scope. • 947 unique disease concepts in scope • NLP tool—MedLEE. • The evaluation has been done on 2 drugs of interest (Rosiglitazone and Metolazone), and 2 diseases of interest (hypertensive disease and diabetes).	Co-occurrence, MI, and DPI	• Annotating the clinical entities using NLP and map to UMLS concept. • Selecting entity type of interest, disease and symptom as phenotypic entity, and drug as drug entity. • The association tables are constructed for three types of co-occurring—disease-disease, drug-disease/symptom and disease-symptom and perform the hypothesis testing of no association using χ^2. • The indirect drug-adverse reaction derives from the computation of MI and DPI as the association chain of drug-disease → disease-disease → disease-symptom. • Only two drugs and two diseases of interest are evaluated for both of direct and indirect associations corresponding known associations from Micromedex[1] for drug-symptom and WebMD[2] for disease-symptom.

(Table 4) Contd….

Reference	Problem	Materials and Study Population	Method Type	Method
Wang X. *et al.* [110]	Detect associations between drug and adverse reaction	• Discharge summary from NYPH in the 2004. • Drugs of interest divides into: - 3 long-time drugs in the market but short-term side effects—ibuprofen, morphine, warfarin. - 3 drugs with new adverse events were detected after 2004—bupropion, paroxetine, rosiglitazone. - 1 drug class—ACE inhibitors; NLP tool—MedLEE; UMLS concept.	Co-occurrence	• Annotating the narrative text using NLP and map to UMLS concept. • Selecting drugs of interest and possible ADE entities. • Filtering confounding factors such as diseases or symptoms occurring before the drug usage, etc. • Determining drug-adverse event co-occurring pairs and hypothesis testing based on χ^2. • Evaluation on six drugs and one drug class.
Harpaz R. *et al.* [111]	ADR signal detection	• Combining of two data sources: - Clinical narratives—discharge summaries, admission notes, and outpatient office visits, and structured data—laboratory test results from NYPH in the 2004-2010. - Adverse event report system (AERS) of the Food and Drug Administration in the 1968-2101Q3. • 3 serious adverse reactions of interest—rhabdomyolysis, acute pancreatitis, QT prolongation. • NLP tool—MedLEE. • RxNorm, UMLS concept (2011AA), MedDRA(V.13.1).	Pipeline detection, ranking	• Annotation the clinical narratives using NLP system to extract medications, disease, and signs and symptoms by mapping to UMLS concept, and temporal information corresponds to admission, discharge and visit dates, laboratory test data directly obtains from the EHR database. • Signal detection by disproportionality analysis and then considering common high ranking of drug-adverse reaction (top K associations) between outcomes from EHR and AERS. • Evaluation by the in-house reference standard set from a pharmacological expert and reviewed by three physicians, which is divided into 2 classes (i) Established—drugs confirmed to be causally related to the ADR, (ii) Plausible—drugs that have a high likelihood of being causative.
Duan L. *et al.* [113]	Rare ADR detection	• Simulated OMOP dataset. • 1 adverse reaction of interest—catastrophic signal.	Ensemble, ranking	• Deploying the ensemble methods from the three different detection models, two-by-two contingency table, likelihood ratio and a Bayesian network models for potential adverse drug reaction detection. • The weight is applied to the combination of the raw scores from three models then generates final ranked score.
LePendu P. *et al.* [115]	Analyzing patterns of off-label drug usage	• Clinical narratives from STRIDE. • Drug indication data from Medi-Span®—for evaluation • 3 groups of adverse reactions (ICD-9) based on HIPAA requirement—rare diseases, celebrity cases, mental health • NLP Tools—NCBO annotator, NegEx trigger • NCBO BioPortal library, RxNorm, SNOMED CT.	Ranking	• Annotating terms in clinical notes using NCBO Annotator. • Applying NegEx trigger rules to separate negated terms then construct temporally ordered bags of terms vector. • Term normalization. • Creating drug-indication associations using sliding window. • Filter confounding factors. • Scoring the association strength using ROR. • Validating associations against known drug-indication database (Medi-Span®). • Ranking the remaining drug-indication association.

(Table 4) Contd….

Reference	Problem	Materials and Study Population	Method Type	Method
Liu M. *et al.* [119]	Detect drug and adverse reaction association	• Dataset of Yoon *et al.* [120] 470 drug-event pairs (10 drugs and 47 laboratory abnormalities). • Dataset from VUMC's EMR 187,595 patients record with 378 drug-event pairs (9 drugs of interest drugs and 42 laboratory abnormalities).	Co-occurrence	• Separating retrospective observations into two groups, study group and comparison group, based on rule based. • In each group, the laboratory result and drug are examined as abnormal and normal outcome. • A two-way contingency matrix is constructed to compute six signal measures, χ^2, PRR, ROR, Yule's Q (YULE), BCPNN, and GPS to test hypothesis of no association. • The evaluation matrices, precision, recall, and F-score are used to assess the performance of each model corresponding own reference standard set.
Park M. Y. *et al.* [121]	Detecting the signals of ADR focused on laboratory abnormalities after treatment with medication	• Admission, discharge, drug prescription, and laboratory test results of inpatients EMR from Ajou University Hospital, in Korea from Jan 2000 - Mar 2010. • 7 non-oncologic drugs of interest—ciprofloxacin, clopidogrel, ketorolac, levofloxacin, ranitidine, rosuvastatin, valproic acid. • 3 oncologic drugs of interest—etoposide, fluorouracil, metho-trexate. • Laboratory anomaly is determined as adverse reaction. • 56 ADEs of interest from UpToDate® Drug Information Database.	Rule based	• Selecting 10 target drugs. • Retrieve the list of known ADEs related to 10 selected drugs that can be represented by laboratory abnormality from UpToDate database. • Constructing the mapping table to link between laboratory abnormalities detected by CERT algorithm and each known ADEs. • Then derive drug-adverse event pairs.
Skentzos S. *et al.* [122]	Identify ADR to Statins	• Clinical narratives of outpatients from Partners Enterprise Allergy Repository (PEAR) [145] in the 2000-2010. • 1 drug of interest—statin. • 3,175 narrative notes. • UMLS concept, MedDRA code.	Rule based, POS	• Deploying NLP phase structure (parse tree) and manual word class for parts of speech with semantic customization, note-level and sentence-level are examined, statin drug name mapped to UMLS concept and adverse reaction mapped to MedDRA code. • Comparing with in-house reference set of statin-adverse reaction created from 242 randomly narrative notes selected by two trained pharmacy students.
Roitmann E. *et al.* [123]	To exploit fine-grained drug related adverse event information to stratify patients	• Adverse events extracted from clinical narratives from Danish mental health center (6,011 patient records) in the 1998 to 2010. • The structured data for obtaining drug information (ATC), drug dosages, prescription intervals, and diagnoses (ICD-10).	Stratification and co-occurrence	• Identifying 2,347 patient corpus which was prescribed at least one drug and had at least one adverse event. • Constructing patient vectors in a space of 1,190 adverse event dimensions by tf-idf weighted values. • The patients were stratified using the cosine dissimilarity based on their adverse event profiles network. • To compute co-occurrence score and weighted edges to analyze the cluster adverse event. • ATC level 5 based on tf-idf vector is examined to analyze cluster specific drugs. • ICD-10 level 3 based on tf-idf vector is examined to analyze cluster specific diagnosis.

(Table 4) Contd….

Reference	Problem	Materials and Study Population	Method Type	Method
Ji Y. *et al.* [124]	Identify causal relationships between drugs and their associated adverse drug reactions (ADRs)	• Clinical narratives from the Veterans Affairs Medical Center in Detroit, Michigan. • 1 drug of interest-enalapril. • 1,021 patients related to drug of interest 1,290 ICD-9 codes associated with drug of interest.	Rule based	• Detecting the temporal association between taken drug and occurrence of the symptom by considering on re-challenge, re-introduction of the drug and recurrence of the symptom, and de-challenge, withdrawal of the drug and abatement of the symptom from the treatment period. • The experience-based fuzzy RPD model is applied to give the relation strength. • The exclusive causal relation *supp(X→Y)* is a considering of the causal relation between drug and symptom, but ignores causal relation that caused by background noise; temporal association, re-challenge, de-challenge, and the relation strength is used to quantify the degree of association
Peissig P. L. *et al.* [131]	To identify and classify patients at risk	• Healthcare data from CattailsMD EHR-Research Data Warehouse (RDW), Marshfield Clinic in the 1979-2011. • EHR data—diagnoses, procedures, laboratory results, observations, and medications for patients. • 9 diagnosis of interest—acute myocardial infarction, acute liver failure, atrial fibrillation, cataract, congestive heart failure, dementia, type 2 diabetes, diabetic retinopathy and deep vein thrombosis.	Machine learning	• Identifying training set of POS, NEG, and BP (borderline positive) samples, identify training set for congestive heart failure [CHF] and acute liver injury [ALI]. • Deploying Inductive Logic Programming (ILP) for rule learning from given training set. • The rules from ILP are used by comparative ML classifiers—Random forest, SMO, PART, J48, JRIP. • Evaluation based on Receiver Operator Characteristic (ROC) and area under the ROC curve (AUROC).
Liu Y. *et al.* [140]	Discriminating the drug-adverse event pairs from the drug-indication pairs	• Empirical data from STRIDE including narrative notes from inpatient and outpatient records, over 9 million notes, 29,551 SNOMED CT diseases and 2,926 drug ingredients, 86.5 million possible drug-disease pairs. • There are three goal standard datasets—1,550 drug-disease sample from Medi-Span® Adverse Drug Effects Database^{TM}, AERS, and the National Drug File ontology (NDFRT).	Machine learning	• To annotate textual medical records with relevant drug and disease terms by considering the lexicon from UMLS, SNOMED CT, RxNorm and applying NLP technique, NCBO, NegEx. • The drug-disease association is constructed along the patient's timeline, the co-mention and drug-first fractions are used to consider drug and adverse reaction. • SVM classifier is built on three features based on the notion of the drug-first fraction and another three features based on logarithm co-occurrence including four features from z-score of drug-first fraction and co-mention count; the training data consists of 1,550 samples is separated to 980 indications and 570 adverse events. • Evaluation method using 100-fold cross validation and independent validation set.
Iqbal E. *et al.* [141]	Identify instance of adverse drug events (ADEs)	• Clinical narratives from Clinical Record Interactive Search (CRIS) system, the South London and Maudsley NHS Foundation Trust (SLaM) (17,995 patients) in the 2007-2013. • 4 diseases of interest related movement disorders (extrapyramidal side effects [EPSEs], dystonia, akathisia, Parkinsonism and tardive dyskinesia. • NLP tools—GATE [144], Java Annotation Patterns Engine (JAPE).	Rule based	• Constructing EPSE ADE dictionary, including synonym and alternative spelling, then apply GATE to extract ADE related four EPSEs of interest from clinical narratives. • Applying remove rule (defined by JAPE) and retain rule to identify ADE instances. • Iterative to create new and improve rule from misclassification.

(Table 4) Contd….

Reference	Problem	Materials and Study Population	Method Type	Method
Yildirim P. *et al.* [142]	To explore hidden knowledge in the survey data extracted from health records on adverse reactions	• Survey data from the Health Center of Osijek, Eastern Croatia (1,491 children patients).	Clustering	• Deploying *k*-means, *k*-medians, *k*-medoids, and single link clustering (2 <= *k* <= 5). • Cluster evaluation by sum of square error (SSE). • Cluster validity by three measure types (i) external validity indexes—Rand index, Jaccard index, Fowlkes-Mallows index (ii) internal validity indexes—proximity matrix (iii) relative validity indexes.
Karlsson I. *et al.* [143]	ADR prediction	• EMR from the Stockholm EPR Corpus in the 2009-2010. • Diagnosis code of interest—*L27.0* Generalized skin eruption due to drugs and medicaments from ICD-10-SE[3]. • 201 patients related *L27.0* diagnosis as positive samples, and 261 patients from other random diagnosis code as negative samples, the deployment data set related to *L30.9X* 3 (Dermatitis, unspecified) and not include in the negative set. • Structured data was utilized for feature generation, and clinical narratives were used to support the manual evaluation.	Machine learning	• Feature vector generation corresponds to 1,312 drugs, 9,863 diagnosis code, age, and gender, then separate subset of feature into two groups (i) drug only, age, gender (ii) drug and disease, age, gender, then combine considering of the presence drug use (yes, no) and temporal event of drug and diagnosis assignment (before, not), finally there are six combination of feature sets. • Modeling with two machine learning methods: Random forest [146] and JRIP rule learner. • Evaluation using 10-fold cross validation; • Applying the model to deployment set

[1] http://micromedex.com/

[2] http://www.webmd.com/

[3] International Classification of Disease, Version 10, Swedish Modification.

Unfortunately, there are no gold standard dataset for evaluation the drug-adverse reaction association that discovered by the proposed method. Most of studies in ADR detection perform manually review the top ranked results by medical expert such as work of Chen *et al.* [108] (the top five disease-drug association). Conversely, Liu *et al.* [119], Harpaz *et al.* [111] compile in-house reference standard based on the strength evidence for previously known associations or manual investigate by a pharmacological expert. Additional source of known drug-adverse reaction is rather adequate to reference. Wang *et al.* [109] evaluates the outcome with two well-known healthcare communities, MicroMedex and WebMD.

Unfortunately, there are no gold standard dataset for evaluation the drug-adverse reaction association that dis-covered by the proposed method. Most of studies in ADR detection perform manually review the top ranked results by medical expert such as work of Chen *et al.* [108] (the top five disease-drug association). Conversely, Liu *et al.* [119], Harpaz *et al.* [111] compile in-house reference standard based on the strength evidence for previously known associations or manual investigate by a pharmacological expert. Additional source of known drug-adverse reaction is rather adequate to reference. Wang *et al.* [109] evaluates the outcome with two well-known healthcare communities, MicroMedex and WebMD.

CONCLUDING REMARKS

We have reviewed the typical works in the last twenty years on the detection and prediction of adverse drug reactions, a key task in the post-market drug safety surveillance. It is well-known that data-driven approach is essential in solving that problem. In addition to the two traditional data resources of spontaneous adverse events reporting systems and administrative health databases, in the last decade the other three data resources of omics data, social network data and electronic medical records have opened great opportunities for detecting and discovering ADRs. Those new data resources are hard to exploit due to the specific properties. However, the fields of machine learning and data mining have also quickly created various powerful methods that are believed to be successful in analyzing those data resources.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is funded by Vietnam National University at Ho Chi Minh City (VNU-HCMC) under the grant number B2015-42-02 and by JAIST project on Data Science.

REFERENCES

[1] Edwards IR, Aronson JK. Adverse drug reactions: Definitions, diagsis, and management. The Lancet 2000; 356(9237): 1255-9.

[2] Karimi S, Gaire RK, Paris C. Text and data mining techniques in adverse drug reaction detection. ACM Computing Surveys 47(4): Article 56, March 2015.

[3] Alomar MJ. Factors affecting the development of adverse drug reactions. Saudi Pharma J 2013; 22(2): 83-94.

[4] Daly AK. Pharmacogenomics of adverse drug reactions. Genome Med 2013; 5(5).

[5] Gurulingappa H, Rajput AM, Toldo L. Extraction of potential adverse drug events from medical case reports. J Biomed Semantics 2012; 3(15).

[6] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Informatics 2012; 42:885-92.

[7] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Molecular Sys Biol 2010; 6: Article 343.

[8] LaBute MX, Zhang X, Lenderman J, Bennion BJ, Wong SE, Lightstone FC. Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. Plos one 9(9), 2014.

[9] Liu M, Wu Y, Chen Y, *et al*. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc 2012; 19: e28-e35.

[10] Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics 2011; 12:169.

[11] Vilar S, Tatonetti NP, Hripcsak G. 3D. Pharmacophoric similarity improves multi adverse drug event identification in pharmacovigilance. Nature, Scientific Reports 2015.

[12] Wallach I, Jaitly N, Lilien R. A structure-based approach for mapping adverse drug reactions to the perturbation of underlying biological pathway. PLoS ONE 5(8), 2010.

[13] Cami A, Arnold A, Shannon Manzi S, Reis B. Predicting adverse drug events using pharmacological network models. Sci Trans Med 2011; 3(114): 1-10.

[14] Pouliot Y, Chiang AP, Butte AJ. Predicting adverse drug reactions using publicly-available pubchembioassay data. Clin-Pharmacol Ther 2011y; 90(1): 90-9.

[15] Wang W, Haerian K, Salmasian H, Harpaz R, Chase H, Friedman C. A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from pubmed citations. In AMIA Annl Symposium Proc 2011; 1464-70.

[16] Scheiber J, Jenkins JL, Sukuru SCK, *et al.* Mapping adverse drug reactions in chemical space. J Med Chem 2009; 3103-7.

[17] Xiang Y, Albin A, Ren K, Zhang P, Etterm JP, Lin S, Li L. Efficiently mining adverse event reporting system for multiple drug inte-ractions. AMIA Jt Summits Transl Sci Proc 2014; 120-25.

[18] Yamanishi Y, Pauwels E, Kotera M. Drug side-effect prediction based on the integration of chemical and biological spaces. J Chem Inform Model 2012; 52:3284-92.

[19] Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y. Relating drug-protein interaction network with drug side effects. Bioinformatics 2012; 28:522-8.

[20] Keiser MJ, Hert J. Off-Target Networks Derived from Ligand Set Similarity. Chapter 8 in Edgar Jacoby (ed.), Chemo-genomics, Methods Mol Biol 2009; 195-205.

[21] Lee P, Kapusnik-Uner JE, Pharm D, T Nguyen, Bodenreider O. An approximate matching method for clinical drug names. AMIA Annl Symposium 2011; pp. 1117-26.

[22] Lee S, Lee KH, Song M, Lee D. Building the process-drug side effect network to discover the relationship between biological process and side effects. Bioinformatics 2010; 12(Suppl 2): S2.

[23] Wienkers LC, Heath TG. Predicting in vivo drug interactions from in vitro drug discovery data. Nature Reviews 2005; 4: 825-33.

[24] Bender A, Daniel W. Youngb, Jeremy L. Jenkins, *et al.* Chemogenomic data analysis: Prediction of small-molecule targets and the advent of biological fingerprints. Combinatorial Chem High Throughput Screening 2007; (10) 719-31.

[25] Iyer SV, LePendu P, Harpaz R, Bauer-Mehren A, Shah NH. Learning signals of adverse drug-drug interactions from the unstructured text of electronic health records. AMIA Joint Summits Translational Science Proc 2013; 98.

[26] Brouwers L, Iskar M, Zeller G, van Noort V, Bork P. Network neighbors of drug targets contribute to drug side-effect similarity.PLoS One 2011;6(7): e22187.

[27] White TJ, Clark AG, Broder S. Genome-based biomarkers for adverse drug effects, patient enrichment and prediction of drug response, and their incorporation into clinical trial design. Personalized Med 2006; 3(2): 177-85.

[28] Xie L, Li J, Xie L, Bourne PE. Drug discovery using chemical systems biology: Identification of the protein-ligand binding network to explain the side effects of cetp inhibitors. PLoS One 2009; 5(5): e1000387.

[29] Shah NH. Mining the ultimate phenome repository. Nat Biotechnol 2013; 31(12): 1095-7.

[30] Baker NC, Fourches D, Tropsha A. Drug side effect profiles as molecular descriptors for predictive modeling of target bioactivity. Mol Inform 2015; 34: 160-70.

[31] Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharma-cological data in an integrated framework. Bioinformatics 2010; 26: i246-54

[32] Avillach P, Dufour JC, Diallo G, *et al*. Design and validation of an automated method to detect known adverse drug reactions in medline: A contribution from the EU-ADR project. J Am Med Inform Assoc 2013; pp: 446-52.

[33] Blenkinsopp A, Wilkie P, Wang M, Routledge PA. Patient reporting of suspected adverse drug reactions: a review of published literature and international experience. Br J Clin Pharmacol 2007; pp: 148-56.

[34] Duda S, Aliferis C, Miller R, Statnikov A, Johnson K. Ex-tracting drug-drug interaction articles from medline to improve the content of drug databases. Am Med Inform Assoc 216, 2005.

[35] Garcelon N, Mougin F., Bousquet C., Burgun A.. Evidence in pharmacovigilance: Extracting adverse drug reactions articles from medline to link them to case databases. Studies Health Technol Inform 2005; pp: 528-33.

[36] Lardon J, Abdellaoui R, Bousquet C. Adverse drug reaction identification and extraction in social media: A scoping review. J Med Internet Res 2015; 17(7).

[37] Parker J, Wei Y, Yates A, Frieder O, Goharian N. A framework for detecting public health trends with twitter. Int Conference Adv Social Networks Analysis Mining 2013; pp: 556-63.

[38] Wagstaff LR, Mitton MW, Arvik BM, Doraiswamy PM. Statin-associated memory loss: Analysis of 60 case reports and review of the literature. J Humman Pharmacol Drug Ther 2003; pp. 871-80.

[39] Bates DW, Boyle DL, M. B. Vander Vliet, J. Schneider, and L. Leape. Relationship between medication errors and adverse drug events. J General Intern Med 1995; pp. 199-205.

[40] Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRS) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Safety 2001; 10(6): 483-6.

[41] Forster AJ, Murff HJ, Peterson JF, Gandhi TK, Bates DW. The incidence and severity of adverse events affecting patients after discharge from the hospital. Ann Intern Med 2003; pp. 161-7.

[42] Hanlon JT, Schmader KE, Koronkowski MJ, *et al.* Adverse drug events in high risk older outpatients. J Am Geriatrics Soc 1997; pp. 945-8.

[43] Cable J. Adverse events of statins - an informal internet-based study. J Independent Med Res 2009; 7(1).

[44] Kongkaew C, Noyce P, Ashcroft D. Hospital admissions associated with adverse drug reactions: A systematic review of prospective observational studies. Annl Pharmacother 2008; pp. 1017-25.

[45] Lazarou J, Pomeranz BH, ¥ Corey PN. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. JAMA 1998; pp. 1200-5.

[46] Pal NS, Duncombe C, Falzon D, Olsson S. Who strategy for collecting safety data in public health programmes: Complementing spontaneous reporting systems. Drug safety 2013; 36(2) 75-81.

[47] Fourrier-Reglat A, Miriam AM. The EU-ADR project: Preliminary results and perspective. Stud Health Technology Inform 2009; pp. 43-49.

[48] Gottlieb A, Hoehndorf R, Dumontier M, Altman R. Ranking adverse drug reactions with crowdsourcing. J med Internet Res 2015; 17(3).

[49] Trifiro G, Pariente A, Coloma PM, *et al.* Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmaco-Epidemiol Drug Safety 2009; pp. 1176-84.

[50] Xu R, Wang Q. Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. J Biomed Inform 2014; 51: 191-9.

[51] Coloma PM, Schuemie MJ, G. Trifiro G, *et al.* Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: The EU-ADR project. Pharmacoepidemiol Drug Safety 2011; pp. 1-11.

[52] Liu J, Li A, Seneff S. Automatic drug side effect discovery from online patient-submitted reviews: Focus on statin drugs. First International Conference Adv Information Mining Management 2011; pp: 23-9.

[53] E. Mulligen, A Fourrier-Reglat, Gurwitzm D, *et al.* The EU-ADR corpus: Annotated drugs diseases, targets, and their relationships. J Biomed Inform 2012; pp. 879-84.

[54] Scharf O, Colevas AD. Adverse event reporting in publications compared with sponsor database for cancer clinical trials. J Clin Oncol 2006; pp. 3933-8.

[55] Xu R, Wang Q. Combining automatic table classification and relationship extraction in extracting anticancer drug-side effect pair from full-text articles. J Biomed Iinform 53Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events, International workshop on Smart health and wellbeing 2012; pp. 25-32.

[56] Chee B, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. Am Med Informatics Assoc 217, 2011.

[57] Ginn R, Pimpalkhute P, Nikfarjam A, *et al.* Mining twitter for adverse drug reaction mentions: A corpus and classification benchmark. Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, 2014.

[58] Jiang K, Zheng Y. Mining twitter data for potential drug effects. Advanced Data Mining and Applications, Springer Berlin Heidelberg 2013; pp. 434-43.

[59] Jiang Y, Liao QV, Cheng Q, Berlin RB, Schatz BR. Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages. Am Med Informatics Assoc 417, 2012.

[60] O' Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith LK, Gonzalez G. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. Am Med Informatics Assoc 924, 2014.

[61] Patki A, Sarker A, Pimpalkhute P, *et al.* Mining adverse drug reaction signals from social media: Going beyond extraction. BioLink-Sig, 2014.

[62] Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biom Inform 2015; 53: 196-207.

[63] Wang W, Haerian K, Salmasian H, Harpaz R, Chase H, Friedman C. A drug-adverse event extraction algorithm to support pharmacovifilance knowledge mining from Pubmed citations. Am Med Inform Assoc 1464, 2011.

[64] Wu H, Fang H, Stanhope SJ. Exploiting online discussions to discover unrecognized drug side effects. Methods Inf Med 2013; 52(2): 152-9.

[65] Yang C, Srinivasan P, Polgreen MP. Automatic adverse drug events detection using letters to the editor. Am Med Informatics Assoc 2012; p. 1030.

[66] Yang M, Xiaodi W, Kiang M. Identification of consumer adverse drug reaction messages on social media. PACIS, 193, 2013.

[67] Akay, Erlandsson B. A novel data-mining approach leveraging social media to monitor and respond to outcomes of diabetes drugs and treatment. Point-of-Care Healthcare Technol 2013; pp. 264-6.

[68] Aramaki E, Miura Y, Tonoike M, *et al.* Extraction of adverse drug effects from clinical records. Stud Health Technol Inform 2010.

[69] Levin M, Krol M, Doshi MA, Reich LD. Extraction and mapping of drug names from free text to a standardized nomenclature. Am Med Inform Assoc 438, 2007.

[70] Sampathkumar H, Chen X, Luo B. Mining adverse drug reaction from online healthcare forums using hidden Markov model. Med Inform Decision Making 2014.

[71] Yates A, Goharian N. Adrtrace: Detecting expected and unexpected adverse drug reactions from user reviews on social media sites. Advanced in Information Retrieval. Springer Berlin Heidelberg 2013; pp. 816-9.

[72] Yates A, Goharian N, Frieder O. Extracting adverse drug reactions from forums posts and linking them to drugs. ACM SIGIR Workshop on Health Search and Discovery, 2013.

[73] Yom-Tow E, Gabrilovich E. Post-market drug surveillance sans trial costs: Discovery of adverse drug reactions via large-scale analysis of web search queries. J Med Internet Res 2013.

[74] Bate A, Lindquist M, Edwards IR, Orre R. A data mining approach for signal detection and analysis. Drug Safety 25.6, 2002; pp. 293-397.

[75] Benton A, Ungar L, Hill S, *et al.* Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. J Biomed Inform 2011; pp. 989-96.

[76] Harpaz R, Haerian K, Chase HS, Friedman C. Statistical mining of potential drug interaction adverse effects in FDA's spontaneous reporting system. Am Med Inform Assoc 281, 2010.

[77] Jiang G, Liu H, Solbrig H, Chute C. Mining severe drug-drug interaction adverse events using semantic web technologies: A case study. BioData mining 8.1, 2015.

[78] Jin H, Chen J, He, Kelman C, McAullay D, O'Keefe C. Signaling potential adverse drug reactions from administrative health databases. Knowledge and Data Engineering, IEEE Transactions 2010; pp. 839-53.

[79] Leaman R, Wojtulewicz R, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. Workshop on biomedical natural language processing, ACL, 2010.

[80] Li J, Fu AW, He H, Chen J. Mining risk patterns in medical data. ACM SIGKDD 2005; pp. 770-75.

[81] Liu H, Chen H. Azdrugminer: An information extraction system for mining patient-reported adverse drug events in online patient forums. Smart Health, Springer Berlin Heidelberg 2013; pp. 134-50.

[82] Liu X, Chen H. Identifying adverse drug events from patient social media a case study for diabetes. Intelligent System, IEEE, Vol 30, Issue 3.

[83] Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of adverse drug reactions from user comments. Am Med Inform Assoc 1019, 2011.

[84] Xu R, Wang Q. A knowledge-driven conditional approach to extract pharmacogenomics specific drug-gene relationships from free text. J Biomed Iinform 2012; pp. 827-34.

[85] Yang C, Jiang L, Yang H. Detecting signals of adverse drug reactions from health consumer contributed content in social media. SIGKDD Workshop on Health Informatics 2012.

[86] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: The metamap program. Am Med Inform Assoc 17, 2001.

[87] Aronson AR, Lang FM. An overview of metamap: Historical perspective and recent advances. J Am Med Informs Assoc 2010; pp. 229-36.

[88] Marinela X, Kenzevic C, Bivolarevic I, Tanja SP, Slobodan MJ. Using facebook to increase spontaneous reporting of adverse drug reactions. Drug safety 34.4, 2011; pp. 351-2.

[89] Morgan M. W. Identifying and Understanding Clinical Care processes. In: Carter JH, Eds. Electronic Medical Records: A Guide for Clinicians and Administrators. New York: ACP Press, 2001; pp. 159-96.

[90] Hannan TJ. Electronic medical records. In: Hovenga E, Kidd M, Cesnik B, Eds. Health informatics: An Overview 1996; pp. 133-48.

[91] Karimi S, Wang C, Metke-Jimenez A, Gaire R, Paris C. Text and data mining techniques in adverse drug reaction detection. ACM Computing Surveys (CSUR) 2015; 47(4): 56.

[92] Feldman R, Sanger J. II. Text Mining Preprocessing Techniques. In: The text mining handbook: advanced approaches in analyzing unstructured data. New York: Cambridge University Press 2007; pp. 57-63.

[93] Feldman R, Dagan I. Knowledge discovery in textual databases (KDT). In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada. AAAI, Aug 20-21, 1995; 95: 112-7.

[94] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. J Am Med Inform Assoc 2011; 18(5): 544-51.

[95] Damjanovic V, Kurz T, Westenthaler R, Behrendt W, Gruber A, Schaffert S. Semantic enhancement: The key to massive and heterogeneous data pools. In Proceeding of the 20th international IEEE ERK (electrotechnical and computer science) conference 2011; pp. 413-6.

[96] Payne PRO. Chapter 1: biomedical knowledge integration. PLoS computational biology 2012; 8(12): e1002826.

[97] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: A review of recent research. Yearb Med Inform 2008; 35: 128-44.

[98] Henriksson A. Semantic Spaces of Clinical Text: Leveraging Distributional Semantics for Natural Language Processing of Electronic Health Records. PhD dissertation, Stockholm: Stockholm University 2013.

[99] Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010; 10(1): 70.

[100] Murray RE, Ryan PB, Reisinger SJ. Design and validation of a data simulation model for longitudinal healthcare data. AMIA Annu Symp Proc 1176, 2011.

[101] Saeed M, Villarroel M, Reisner AT, *et al.* Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Critic Care Med 2011; 39(5): 952.

[102] Moskowitz A, McSparron J, Stone DJ, Celi LA. Preparing a new generation of clinicians for the era of big data. Harvard Med Student Rev 2015; 2(1): 24.

[103] Li Y, Ryan PB, Wei Y, Friedman, C. A method to combine signals from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions. Drug Safety 2015; 38.10: 895-908.

[104] Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data‐mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther 2012; 91(6): 1010-21.

[105] Doan S, Collier N, Xu H, Duy PH, Phuong TM. Recognition of medication information from discharge summaries using ensembles of classifiers. BMC Med Inform Decision Making 2012; 12(1): 36.

[106] Suominen H, Salanterä S, Velupillai S, *et al.* Overview of the ShARe/CLEF eHealth evaluation lab 2013. In Information Access Evaluation, Multilinguality, Multimodality, and Visualization, Springer Berlin Heidelberg, 2013.

[107] Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: An initial study. J Am Med Inform Assoc 2008; 15(1): 87-98.

[108] Wang X, Hripcsak G, Friedman C. Characterizing environmental and phenotypic associations using information theory and electronic health records. BMC Bioinform 2009; 10(Suppl 9): S13.

[109] Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. J Am Med Inform Asso 2009; 16(3): 328-37.

[110] Harpaz R, Vilar S, DuMouchel W, *et al.* Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2013; 20(3): 413-9.

[111] Johnson SB, Hripcsak G, Chen J, Clayton P. Accessing the Columbia clinical repository. In Proceedings of the Annual Symposium on Computer Application in Medical Care, Am Med Inform Assoc 1994.

[112] Duan L, Khoshneshin M, Street WN, Liu M. Adverse drug effect detection. Biomedical and Health Informatics IEEE J 2013; 17(2): 305-11.

[113] Liu M, Wu Y, Chen Y, *et al.* Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc 2012; 19(e1): e28-e35.

[114] LePendu P, Liu Y, Iyer S, Udell MR, Shah NH. Analyzing patterns of drug use in clinical notes for patient safety. AMIA Summits on Translational Science Proceedings 63, 2012.

[115] Dalianis H, Hassel M, Henriksson A, Skeppstedt M. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. Swedish Language Technol Conference 2012.

[116] Zhao J, Henriksson A, Bostrom H. Detecting Adverse Drug Events Using Concept Hierarchies of Clinical Codes. In Healthcare Informatics (ICHI) 2014 IEEE International Conference 2014.

[117] Henriksson A, Kvist M, Hassel M, Dalianis H. Exploration of adverse drug reactions in semantic vector space models of clinical text. In Proceedings of the 29th international conference on machine learning 2012.

[118] Liu M, Hinz ERM, Matheny ME, *et al.* Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. J Am Med Inform Assoc 2013; 20(3): 420-6.

[119] Yoon D, Park MY, Choi NK, Park BJ, Kim JH, Park RW. Detection of adverse drug reaction signals using an electronic health records satabase: Comparison of the Laboratory Extreme Abnormality Ratio (CLEAR) algorithm. Clin Pharmacol Ther 2012; 91: 467-74.

[120] Park MY, Yoon D, Lee K, *et al.* A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. Pharmacoepidemiology and drug safety 2011; 20(6): 598-607.

[121] Skentzos S, Shubina M, Plutzky J, Turchin A. Structured vs. unstructured: factors affecting adverse drug reaction documentation in an EMR repository. In AMIA Annual Symposium Proceedings, Am Med Inform Assoc 2011.

[122] Roitmann E, Eriksson R, Brunak S. Patient stratification and identification of adverse event correlations in the space of 1190 drug related adverse events. 5, Frontiers in physiology 2014.

[123] Ji Y, Ying H, Dews P, *et al.* An exclusive causal-leverage measure for detecting adverse drug reactions from electronic medical records. In Fuzzy Information Processing Society (NAFIPS), Annual Meeting of the North American, IEEE, 2011.

[124] Liu M, Cai R, Hu Y, *et al.* Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning. J Am Med Inform Assoc 2014; 21(2): 245-51.

[125] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012; 13.6: 395-405.

[126] Cami A, Arnold A, Manzi S, Reis B. Predicting adverse drug events using pharmacological network models. Sci Trans Med 2011; 3(114): 114ra127.

[127] Huang LC, Wu X, Chen JY. Predicting adverse side effects of drugs. Bmc Genom 2011; 12(Suppl 5): S11.

[128] Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine Genome med 2015; 7(1): 41.

[129] Karnes JH, Cronin RM, Rollin J. A genome-wide association study of heparin-induced thrombocytopenia using an electronic medical record. Thromb Haemost 2015; 113(4): 772-81.

[130] Peissig PL, Costa VS, Caldwell MD, *et al.* Relational machine learning for electronic health record-driven phenotyping. J Biomed Iinform 2014; 52: 260-70.

[131] Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. Ann Intern Med 1993; 119(8): 844-50.

[132] Chen L, Friedman C. Extracting phenotypic information from the literature via natural language processing. Medinfo 2004; 11(Pt 2): 758-62.

[133] Friedman C, Alderson PO, Austin J, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. JAMIA 1994; 1(2): 161-74.

[134] Jiang M, Chen Y, Liu M, *et al.* A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011; 18(5): 601-6.

[135] Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. J Am Med Inform Assoc 2010; 17(5): 507-13.

[136] Sohn S, Kocher JPA, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. J Am Med Inform Assoc 2011; 18(Supplement 1): i144-9.

[137] Kovačević A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal

expressions and events from clinical narratives. J Am Med Inform Assoc 2013; 20(5): 859-66.

[138] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004; 32(suppl 1): D267-70.

[139] Liu Y, LePendu P, Iyer S, Shah NH. Using temporal patterns in medical records to discern adverse drug events from indications. AMIA Summits on Translational Science proceedings 47, 2012.

[140] Iqbal E, Mallah R, Jackson RG. Identification of adverse drug events from free text electronic patient records and information in a large mental health case register. Plos one 2015; 10(8): e0134208.

[141] Yildirim P, Majnarić L, Ekmekci OI, Holzinger A. Knowledge discovery of drug data on the example of adverse reaction prediction. BMC bioinformatics 2014; 15(Suppl 6): S7.

[142] Karlsson I, Zhao J, Asker L, Boström H. Predicting adverse drug events by analyzing electronic patient records. In Artificial Intelligence in Medicine, Springer Berlin Heidelberg, 2013.

[143] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: an architecture for development of robust HLT applications. In Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2012.

[144] Kuperman GJ, Gandhi TK, Bates DW. Effective drug-allergy checking: methodological and operational issues. J Biomed Iinform 2003; 36(1): 70-9.

[145] Breiman L. Random forests. Machine Learning 2001; 45(1): 5-32.

[146] Penney TM. Dictate a discharge summary. BMJ 298.6680, 1989; pp. 1084-5.