Text and Data Mining Techniques in Adverse Drug Reaction Detection

SARVNAZ KARIMI, CHEN WANG, ALEJANDRO METKE-JIMENEZ, RAJ GAIRE, and CECILE PARIS, CSIRO

We review data mining and related computer science techniques that have been studied in the area of drug safety to identify signals of adverse drug reactions from different data sources, such as spontaneous reporting databases, electronic health records, and medical literature. Development of such techniques has become more crucial for public heath, especially with the growth of data repositories that include either reports of adverse drug reactions, which require fast processing for discovering signals of adverse reactions, or data sources that may contain such signals but require data or text mining techniques to discover them. In order to highlight the importance of contributions made by computer scientists in this area so far, we categorize and review the existing approaches, and most importantly, we identify areas where more research should be undertaken.

Categories and Subject Descriptors: A.1 [**General Literature**]: Introductory and Survey; H.2.8 [**Database Applications**]: Data-Mining; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Drug side effect, adverse drug reaction, drug safety, data mining, text mining, statistical analysis, signal detection

ACM Reference Format:
Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015. Text and data mining techniques in adverse drug reaction detection. ACM Comput. Surv. 47, 4, Article 56 (May 2015), 39 pages.
DOI: http://dx.doi.org/10.1145/2719920

1. INTRODUCTION

We survey the *surveillance of adverse drug events* within the health domain, where statistical methods, text mining algorithms, and data mining algorithms are increasingly being developed and applied to detect adverse events and improve drug safety. With new sources of electronic data—such as electronic health records—becoming more available, the role of data scientists becomes more important in order to mine new knowledge that is hidden in such data.

Any injury caused by a medication is called an *adverse drug event* (ADE)*. This injury can be an unintended effect of the recommended usage of a drug, as per its

*A complete list of abbreviations is provided as appendix.

This work is funded by the Australian e-Health Research Centre and Digital Economy Research Program at the Commonwealth Scientific and Industrial Research Organisation (CSIRO).
Authors' addresses: S. Karimi, C. Wang, and C. Paris, CSIRO, Sydney, Australia; A. Metke-Jimenez, The Australian e-Health Research Centre, CSIRO, Brisbane, Australia; R. Gaire, CSIRO, Canberra, Australia. Mail: CSIRO PO Box 76. Epping NSW 1710. Australia; email: firstname.lastname@csiro.au.

DOI: http://dx.doi.org/10.1145/2719920

label or prescription; the off-label usage of a drug; or a medication error. *Adverse drug reactions* (ADRs) are a subset of adverse drug events, referring to an unexpected harm caused by the normal use of medication at the normal dosage. It therefore does not cover those injuries caused by inappropriate or off-label usages of a medication. For ADRs, a *causal* relationship is a reasonable possibility between a medication and an adverse drug event [ICH Working Group 1996]. We also refer to adverse drug events or reactions as *side effects* or *adverse effects*. The term *side effect* is used more commonly among nonhealth professionals, and it can also cover beneficial unintended reactions to a medication. These side effects vary among individuals, and sometimes they are an artifact of taking another medication at the same time. Among the experts, however, using the term *adverse effect* is preferred [Edwards and Aronson 2000].

ADEs create major concerns in public health. They are responsible for thousands of incidents of death or serious injury, as well as millions of hospitalizations and prolonged hospital stays. Consequently, they cost billions of dollars to the health care systems around the world. For example, in Australia, which has a relatively small population of around 23 million people, it is estimated that about 400,000 visits to general practitioners and 190,000 hospital admissions each year are due to adverse drug events [ACSQHC 2011]. Medicine-related hospitalizations, including adverse drug reactions and medication errors, cost approximately $660 million in 2009 in Australia [Roughead and Semple 2009]. One study in the state of Victoria, Australia, during 2003–2004 estimated that the extra costs in hospital admissions due to adverse events per incident is nearly $7,000, with an average of an extra 10 days per stay per patient in the hospital [Ehsani et al. 2006]. These patients also have a 7% higher chance of death compared to the patients without complications [Ehsani et al. 2006]. Similar estimates are reported for the United States and Europe. In the United States, it is estimated that the cost of each adverse drug reaction in community hospitals is approximately $3,000 per case [Hug et al. 2012; Classen et al. 1997]. In 2008, the European Commission reported that 5% of all hospital admissions are due to adverse drug reactions, and adverse drug reactions are the fifth most common cause of death in hospitals. Given the scale of this problem, even small improvements in the process of detecting adverse drug effects potentially can have a major impact on public health.

Roughead and Semple [2009] estimate that approximately 50% of the medication-related incidents are preventable. To address some of these issues, the Australia National Safety and Quality Health Service Standards [ACSQHC 2011] recommends the *use of technology to support information recording and transfer* as one of the main methods to reduce adverse drug events.

How are the side effects of medications we use identified before they are marketed? What happens after they are marketed? What role has information technology played in this process, and what are the future avenues for a stronger presence of *computer science* in *drug safety*? These are the questions that we answer in this survey, with a focus on the latter to highlight the role of data scientists in the drug safety domain.

We survey the most representative of the existing literature in the area of mining for adverse drug events from structured, semistructured, and unstructured data. The published peer-reviewed literature largely focuses on data from six main sources: adverse events reporting systems, administrative databases, medical literature, electronic health records, search engine logs, and social media.

The science of ensuring drug safety is a multidisciplinary area involving administrative, health care, pharmaceutical, and computer science domains. Even though the area has been studied from different perspectives, it still is in its infancy in some domains. To our knowledge, this area has not been comprehensively surveyed from the viewpoint of computer science in order to draw a big picture of the influence of computer science in drug safety detection. We identify strengths and shortcomings of the existing

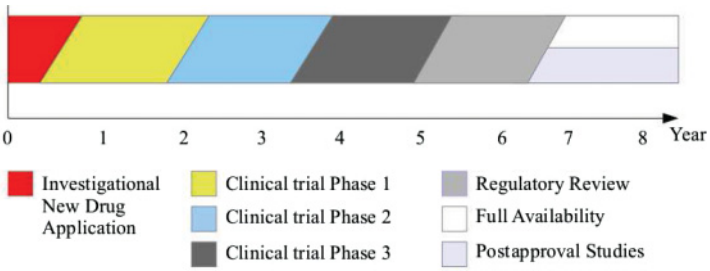

Fig. 1. A typical timeline for a standard drug approval and evaluation. Some of these stages may be repeated or omitted depending on the drug (not shown).

methodologies and algorithms in order to identify pathways for further research that could be conducted by computer and data scientists in this emerging area.

The article is organized to first cover background information in Section 2, including key domain-specific concepts. We introduce the history of applying information technology in the detection of adverse drug events in Section 3. We then depict an overview of the data sources exploited in adverse drug reaction discovery. This view is then used to categorize the related literature based on which data source they employ (Section 4) and what computational methods they propose or apply (Section 5). We outline the open research areas in Section 6.

2. ADVERSE DRUG REACTION DETECTION PROCESS

Pharmacovigilance—the science and activities related to the detection, assessment, understanding, and prevention of adverse drug events or any other drug-related problem (according to the World Health Organization (WHO))—is a complex process, with a number of approaches adopted in different countries. In this section, we briefly explain the process of drug discovery, emphasizing the limitations of premarket trials to identify the adverse drug events. We then move on to the postmarketing stage, or pharmacovigilance, and draw a big picture covering different strategies implemented in the surveyed literature.

2.1. Drug Discovery

The process of drug discovery, formulation, and evaluation is time consuming and expensive. It can easily take 10 to 15 years, or more, from the start of testing of a few thousand compounds that could potentially treat a given disease till one is chosen as the most promising compound for further investigation, tested on animals and later on human volunteers, and finally marketed if everything is successful. To follow, we briefly describe a typical drug evaluation period using a process used in the United States[1] to set the scene for the rest of the article, which is focused on *postmarketing surveillance* and *the role of computer science and information technology* in this process.

Figure 1 shows a simplified timeline for a standard drug evaluation. Once a drug is discovered and preclinical testing is finalized, the pharmaceutical company lodges an application with the corresponding regulator (e.g., Food and Drug Administration (FDA)) for an *investigational drug*. Upon approval of the application, clinical trials start. The first trial is usually run on a maximum of 100 healthy humans. For specific drugs, such as cancer medications, initial trials are run on patients. If the drug passes, a second round starts with the participation of 100 to 500 patients with the condition

[1]More at http://www.fda.gov/drugs/resourcesforyou/consumers/ucm143534.htm.

of interest. If the results of this round are acceptable (i.e., the drug is effective for the given condition, it proves to be safe, and the dosing is right), a third round of testing starts. In this phase, randomized clinical trials are performed on at least 1,000 to 5,000 patients. Some of these patients receive the new drug candidate, and some receive a placebo. One of the goals of this phase is to identify potential adverse drug reactions. After this phase, the results are submitted to the regulatory body for review and approval. Regulators evaluate a number of parameters, including the manufacturing capability, the labeling of the product, and the results from the clinical trials. If the drug is confirmed to be beneficial and its benefits outweigh its adverse reactions, it is released to the market. The adverse reactions identified during the trials are recorded in the product information (PI).

Although the clinical trials are performed carefully, they have some limitations, including the following:

—They are often short in time and involve a limited number of people [Zeng et al. 2002].
—They do not fully represent the target population of the medication as they may exclude patients who receive other medications, an age group (e.g., elderly), and those who have complicated medical conditions.
—Clinical trials may not detect reactions with very low incident rates [Stephens et al. 1985].

To compensate for these shortcomings, postmarketing pharmacosurveillance methods are used, for example, spontaneous reporting, prescription event monitoring, and cohort studies.

2.2. Postmarket Surveillance

Pharmacovigilance evaluates the safety of each drug to enhance the safety profile of the drug over its lifetime in the market [Coloma et al. 2013]. One important step in this process is safety *signal detection*. To follow, we formally introduce relevant terminologies.

2.2.1. Safety Signal Detection. The WHO defines a *safety signal* as reported information on a possible causal relationship between an adverse event and a drug. Recently, the Council for International Organizations of Medical Sciences [CIOMS 2010] extended this definition to cover any new potential causal relationship or a new aspect of a known association between a drug and an event, either adverse or beneficial.

A set of activities that, based on various sources of data, determines whether there is any safety concern regarding an active ingredient or medicinal product is called *signal management*. These activities include signal detection, signal validation, prioritization and assessment, recommendation for action, and exchange of information with authorities [EMA 2012]. *Signal detection* is the process of discriminating between the safety signal and noise [Nelson et al. 2009]. We briefly explain most of the activities related to signal management, but our focus is on signal detection.

2.2.2. Active Versus Passive Surveillance. Two different postmarket surveillance approaches are practiced to ensure the safety of medications: passive and active. Passive approaches rely on individual reports of potential ADEs from different sources such as health professionals or manufacturers. Active methods, however, seek to automatically generate such safety reports from different data sources, such as patient health records and medical and pharmacy claims databases [Furberg et al. 2006; Gordon 2008; Stang et al. 2010]. Ultimately, signals generated from these reports, passive or active, are validated using similar methods, which often require human experts, to establish *causality*.

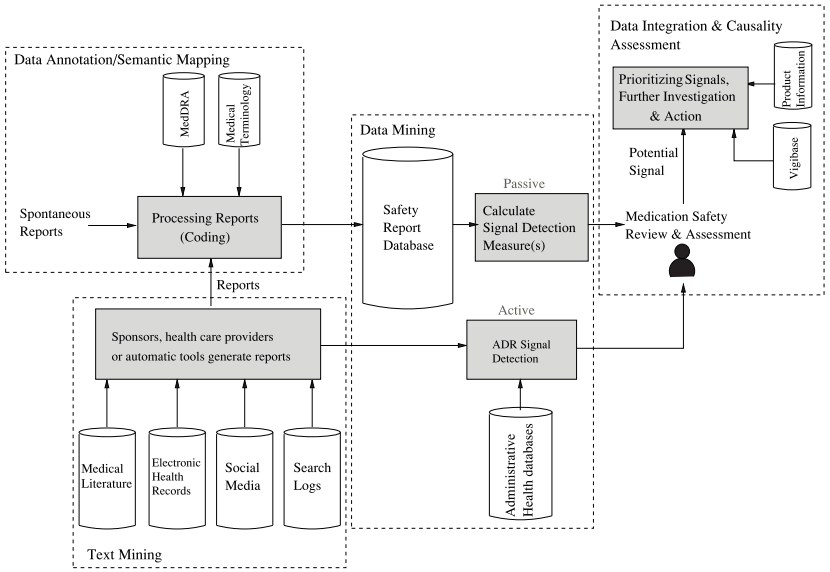

Fig. 2. A generic postmarket surveillance process.

2.2.3. Postmarket Surveillance Procedure. To detect the harmful side effects that go undiscovered during clinical trials, multiple sources of information are consulted. Pharmaceutical companies continue to monitor for potential adverse effects, with regulatory bodies mandating that these companies report suspected adverse effects. We summarize the process of signal detection and validation in Figure 2. It covers possible scenarios for both active and passive surveillance. This process is used in limited scale within pharmaceutical companies for their own products and in larger scale within regulatory bodies for nationwide marketed drugs.

Passive surveillance of adverse drug reactions relies on *spontaneous* reports from sponsors (i.e., pharmaceutical companies), hospitals, physicians, pharmacists, health departments, and consumers. The reports submitted by sponsors could be sourced from consumers, other resources such as medical literature, or even, more recently, social media. Given that a number of sources are used for these reports, there is a high chance of having duplications. One major disadvantage of spontaneous reporting is *underreporting*. It is estimated that, on average, 94% of ADRs are not reported [Hazell and Shakir 2006].

In recent years, active surveillance has received much attention to encourage generating some of the reports automatically from different sources, from more reliable ones such as electronic health records to less reliable ones such as patient forums and search engine logs. Both text and data mining techniques have been developed and applied in different stages of drug safety surveillance. Given that the data types are different, the mining techniques suitable for these tasks can also potentially be different.

Passive and active detection of adverse drug reactions can be used simultaneously in large scale or in a controlled environment such as a specific hospital [Thrmann 2001].

We review the different data sources that have been considered for safety signal detection in Section 4. Once ADR reports are received by regulators, they are processed accordingly and stored in a Safety Report Database. One step is allocating standard terminology, known as the Medical Dictionary for Regulatory Activities (MedDRA), to

the reports if they do not already contain it; this also sometimes involves normalizing the drug names (data annotation or semantic mapping in Figure 2). Given that these reports contain a description of the adverse reaction and may also contain some information on the patient's background, they are potential candidates for natural language processing algorithms that assign standard terminology automatically. We review relevant literature in Section 5.4.1.

Regulators constantly process the incoming reports to detect potential signals (shown in the middle box in Figure 2). These signals are generated using different statistical or data mining techniques. We review the literature on these techniques in Section 5.1. These techniques, however, suffer from inaccurate statistics due to underreporting and the lack of availability of certain statistics, such as off-the-shelf medication consumption or the total number of drug consumers. To compensate for this, some studies suggest complementary solutions such as using data mining techniques on administrative databases. We review these techniques in Section 5.3.

2.2.4. Causality Assessment. One major step after signal detection is establishing *causality* of the signal (Figure 2, the right-most box), that is, to establish if the drug or its active ingredients can cause the given ADR. General causality assessment in the medical domain is normally guided by Bradford-Hill criteria [Hill 1965]. For ADR causality, the following aspects can be examined as exemplified by adapted Bradford-Hill criteria [Anderson and Borlak 2011]: strength of association, consistency of association, dose/response relationship, temporal relationship, coherence and specificity, and plausibility.

The causality assessment is often done manually because domain expertise plays an important role and it is also a requirement for incorporating multiple information sources such as international databases. There are a limited number of studies that attempt to automate this process, which we review in Section 3 and Section 5.3.

3. HISTORY OF USING INFORMATION TECHNOLOGY IN SIGNAL DETECTION

The application of information technology (IT) and its potential benefits in the health domain have been studied for decades. In the domain of detecting adverse reactions, the application of IT started in the 1970s. Most of the early studies focused on using computerized patient records[2] in hospitals to *actively* and *automatically* detect potential signals of adverse reactions. Hulse et al. [1976], for example, implemented a medication monitoring system to alert potential ADRs to the pharmacists in a hospital. Their system positively contributed to changes of patients' therapies to avoid ADRs.

In the late 1970s and early 1980s, researchers argued that clinical judgments were not sufficient for estimating the probability that a drug causes an adverse reaction. Koch-Weser et al. [1977] and Naranjo et al. [1981] pointed out that the lack of a method for establishing causality between an ADE and a drug generates large between-rater and within-rater variability in the safety assessments done by physicians and pharmacists. While Koch-Weser et al. [1977] and researchers of similar studies mostly focused on a clearer definition of ADR, Naranjo et al. [1981] proposed that a systematic method be devised to assess causality. They therefore devised a weighted questionnaire for assessing each reaction. Through a study with participation from both physicians and clinical pharmacologists, they compared traditional methods (expert opinion only) and the use of their devised scoring measures. They then calculated the agreement among the participants and found that their agreement indeed increased substantially with their simple questionnaire-based method. This study became an important basis

[2]These days commonly known as electronic health records or electronic medical records.

for causality assessment, and the current WHO-UMC standardized case causality assessment system was developed based on the Naranjo's method.

Classen et al. [1991, 1992] took it a step further and developed a computerized system that implemented the signal detection method proposed by Naranjo et al. [1981]. They argued that monitoring for ADRs using spontaneous reporting systems (SRSs) detects only a small number of ADRs. Their system, in contrast, actively looks for potential ADRs based on actions recorded in patient records, such as discontinuation of medications, decreases in dosage, or ordering of antidotes. They deployed their system in a hospital in the United States and demonstrated its effect in discovering adverse drug events. Their system identified 731 ADEs over the course of 18 months, as compared to nine ADEs identified using the traditional methods during the same period. They concluded that using a computerized system equipped with a signal detection system together with supervision from the pharmacists on site in a hospital offers a more effective method of detecting ADEs in hospitals.

We can mark this point in history as the start of moving from *passive* postmarketing surveillance done using SRS databases that were mostly managed by regulators to *active* surveillance that does not rely on volunteer reports of adverse events and generates these reports automatically. At this point in time, however, this was only done by researchers in a small number of hospitals, and it was not a systematic movement.

Another branch of work evaluated the effect of using computerized systems in pharmacies that warn for possible ADRs due to drug–drug interactions based on the patients' history of medications and their current prescriptions. Kirking et al. [1986] compared three groups: users of two different computer systems available in pharmacies and also pharmacists who did not possess any computer. The main goal of their study was to see whether or not using computers equipped with pharmacy systems that check for drug–drug interactions would improve detection and follow-up of potential drug interactions. They studied a large number of potential factors, including whether pharmacists in two groups of computer and noncomputer users thought detecting clinically significant drug interactions was important or not, and what the sources of acquiring knowledge on drug interactions were. They found significant differences between the two computerized systems, with one generating significantly more signals and being much more useful to the pharmacists than the other. Both systems were the primary source of knowledge in the known drug interactions for the pharmacists.

Jha et al. [1998] criticized the previous work on computerized systems, such as the work by Classen et al. [1991], which advocated the use of computer systems for ADE detection without comparing them with *chart reviews*. Chart reviews are performed by nurses at the hospital on patients' data. Jha et al. [1998] compared spontaneous reporting, chart reviews, and a rule-based computerized system in a hospital setting for ADE detection. They showed that both chart reviews and computerized monitoring beat spontaneous reporting for the number of ADEs found. They also found that there was little overlap in the ADEs found by these three methods, with the computerized system picking up mostly severe ADEs such as renal failure. They further showed that a computerized system required much fewer human hours for reviewing compared to chart reviews, and that it was cheaper to operate. A similar study by Hwang et al. [2008] provided a similar outcome. Their computerized monitoring system identified almost all the severe ADEs (confirmed by experts) and only missed a small portion of ADEs that would have been picked up by a chart review. However, they also mentioned that their system still needed to be improved for accuracy.

Honigman et al. [2001] took the work by Classen et al. [1991] one step further and implemented four search mechanisms in a hospital setting over patient electronic records to detect ADEs. They also used the assessment methodology of Naranjo et al. [1981] to avoid reviewers' bias in determining the causality. The searching mechanisms

used by their systems were as follows: (1) search in ICD-9 codes;[3] (2) search allergies from a database of drug allergy lists, product names, generic names, and ingredients; (3) event monitoring rules similar to that of Jha et al. [1998]; and (4) text search on visit notes based on a terminology tool called Micromedex M^2D_2. Their text search, though, was far from what we know today as free-text search, as it included a large amount of manual work to break the sentences, find terms, exclude negative terms, and normalize them to uppercase so that they could be filtered through the M^2D_2 structure.

Results from a study by Honigman et al. [2001] confirmed the value of using computerized systems in detecting ADEs and also the value of shifting to electronic health records, which, at the time, were not widely utilized.

Bates et al. [2003] systematically reviewed the use of IT in detecting adverse events in health care, not exclusive to events related to drugs. They emphasized the success of studies on adverse drug events and adverse drug reactions, which, to that date, were based on simple keyword search or natural language processing techniques, to be seen as an example to be followed in other areas of health. Some of the techniques that were highlighted as simple yet successful approaches were lexical matching, with or without mapping to a thesaurus, and syntactic methods to handle negations.

Their review lists all the computerized studies published till 2003 with details including the type(s) of patients considered in the study, the signals used for detection, the level of automation, the rate of false positives or false negatives, the barriers to implementation of the proposed system, and whether the focus was on ADRs or ADEs. They divided the studies' levels of automation into *low-end* and *high-end*, with low-end being a monitoring system that relies on manual entry of specific information to the system for generating an alert, and high-end meaning that the system used multiple resources to generate alerts to be reviewed by an expert. They also divided the studies based on inpatient (hospitals) and outpatient (nonhospital) settings and categorized the barriers in these settings.

Throughout the years, algorithms used to detect ADEs in hospitals did not advance much, with even more recent studies such as that of Tinoco et al. [2011] still relying on the same system described by Classen et al. [1991], which was based on the computerized system established in 1972 at a hospital in the United States to compare computerized surveillance to manual chart review.

Another line of work involving computer science was in the context of regulators, such as the FDA. Woosley [2013] reviewed 100 years of drug regulation in the context of the FDA in the United States and found that the FDA started to analyze the data from its spontaneous reporting system only from 1993. Data mining methods were then developed to identify signals from these reports. Later, in 1998, they made their de-identified data public for researchers to work on. Since then, more advanced data mining techniques have been developed for signal detection. We will discuss them in Section 5.1 in more detail.

Deficiencies of the passive methods, extracting signals from spontaneous reports that rely on volunteer reports, became more recognized as time passed. Meanwhile, electronic patient records became more popular. In 2007, the FDA initiated Sentinel, which allowed for sharing of de-identified patient data among different organizations such as Medicare and hospitals for central processing and analysis while the actual data remained in the original organization. This allows researchers from different fields, including computer scientists, to develop methods for actively monitoring drug safety. We expand on these methods in the following sections.

[3]ICD is the International Classification of Diseases, created to standardize the classification of diseases for international comparability of mortality statistics.

Table I. List of Data Sources and Their Properties That Are Used or Proposed for Adverse Drug Reaction Discovery

Data Source	Type	Reliability	ADR Specific	ADR Monitoring
SRS	Structured	High	Yes	Passive
EHR	Structured	High	No	Active & passive
Administrative DB	Structured	High	No	Active
Medical literature	Unstructured	High	No	Active & passive
Medical forums	Unstructured	Low	No	Active
Social media	Unstructured	Low	No	Active
Search engine logs	Semi-structured	Low	No	Active

4. DATA SOURCES FOR SIGNAL DETECTION

Information technologies available today enable us to easily collect, store, and share patient health care information and quickly communicate about treatment and its adverse reactions. The importance of embracing and exploiting these new communication technologies in health care has been pointed out recently in various studies. A report by Hawn [2009], for example, paints a picture of a high-tech health care ecosystem in the near future, where health care providers and patients are connected both physically through visits to general practitioners and electronically through emails, chats, and video-conferencing. While concerns and potential drawbacks such as privacy and legal issues have been raised, this might still lead to a more patient-centred health care system.

An increasing amount of data in electronic formats is already available to be mined for public health, including to help discover ADRs not revealed by clinical trials. Some of these data sources such as SRS databases are naturally rich in ADE content since they are created to report or share information about ADEs. Other data sources, such as administrative health databases, EHRs, and social media, are created for different purposes and may sparsely contain potential ADEs that might not have been reported elsewhere. We list these data sources in Table I, along with their properties such as their reliability (high or low), whether or not they are specifically made for ADR discovery, and what type of surveillance (passive or active) they support. To follow, we review different types of data sources that have been considered when developing ADR detecting methodologies.

4.1. Spontaneous Reporting Systems

Spontaneous adverse events reporting systems (SAERSs) are information systems that allow the submission of reports of suspected adverse drug events by health professionals, manufacturers, or patients. Some examples of this type of system are the WHO Individual Case Safety Reports (ICSR) Database VigiBase [Lindquist 2008], the U.S. FDA Adverse Event Reporting System (FAERS),[4] the Yellow Card System of Medicines and Healthcare products Regulatory Agency (MHRA) in the United Kingdom (UK) [The Yellow Card Scheme 2013], and the Therapeutic Goods Administration (TGA) Adverse Drug Reaction System (ADRS) in Australia. There are also initiatives such as the Council for International Organizations of Medical Sciences (CIOMS) or the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) that set standards for reporting case safety forms and their reporting means. Here, we use the FAERS as an example to introduce the information available in such sources. The FDA started collecting adverse

[4]http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm.

drug event information through a voluntary, SRS from 1952 [Woosley 2013]. The system has been evolving in two major phases, namely, the Adverse Event Reporting System (AERS) and the current FAERS.

FAERS is designed for postmarketing safety surveillance for drug and therapeutic biological products. In FAERS, the information collected and made available to the public consists of the following elements:

—Patient demographic and administrative information including age, sex, weight, and report date
—Drug information including drug name, expiration date, dose, and indications for use
—Adverse reaction information including the description of event using *preferred terms* defined in MedDRA and patient outcome

The number of reports submitted by consumers and health care professionals is increasing each year. The percentage of consumer-submitted reports has also been increasing. In 2011, 48.4% of reports were submitted by consumers (statistics calculated based on percentage of consumer reports in 2011 over the total number of reports in the United States [FDA 2014]).

These systems have the advantage of being very large, being relatively cheap, and containing information about the whole population [Wilson et al. 2004]. Their analysis can lead to formulation of hypotheses. These hypotheses in turn can lead to further investigations and ultimately to regulatory warnings, changes of product information leaflets, and even withdrawals of marketing authorizations [Pal et al. 2013].

These systems, however, suffer from several disadvantages. Because the initiative of reporting is left to the reporters, the data collected by this method are incomplete in terms of both quality and quantity [Hazell and Shakir 2006]. Underreporting is widespread and significant. Therefore, these systems cannot provide the estimation of prevalence of an ADE with certainty, and the risk factors of ADEs cannot be established with confidence [Pal et al. 2013]. The passive nature of these systems results in reporting bias. A propensity of reporting only unusual reactions, serious reactions, and reactions for new drugs has been found among physicians and pharmacists in different countries [Biriell and Edwards 1997]. Furthermore, the absence of a pharmacovigilance mandate and clear protocols are also barriers to ADE reporting [Olsson et al. 2010].

4.2. Administrative Health Databases

Administrative health databases, also known as longitudinal observational databases (LODs), are systems that routinely record health events such as hospital admissions, medical services, and drug prescriptions for administration purposes [Jin et al. 2006, 2010]. These databases can provide high-level information about a patient's illness and medication history.

In countries where health care funding is provided by the public sector, the relevant patient data are submitted to government agencies for funding requirements. In Australia, for example, the patient-level hospital separation data, Medicare Benefit Scheme (MBS)[5] data, and Pharmaceutical Benefits Scheme (PBS)[6] data are submitted for each patient. Because of the complex multijurisdictional health care funding system, observational health care data are also prepared in linkable databases [Kelman et al. 2007]. The Queensland Linked Dataset (QLDS) [Williams et al. 2002] is

[5]http://www.mbsonline.gov.au/.
[6]http://www.pbs.gov.au/.

an example of such databases. Similarly, The Health Improvement Network (THIN)[7] in the United Kingdom collects data from general practices and includes temporal information detailing the patient's medical events and prescription history [Reps et al. 2012]. Patient data are also collected in the form of health insurance claims in countries where health care service is managed by the private sector.

The main advantage of these systems is that they cover substantial populations at individual levels [Burgess et al. 2005]. When linked together, they provide valuable insights into actual patient care with potentially short time lags [Kelman et al. 2007]. This could be used to not only identify an ADE but also establish causality of a drug and an ADE [Gordon 2008].

One disadvantage of such data, however, is its lack of record information about positive outcomes, which is required to assess the safety and efficacy of a drug [Gordon 2008; Suling and Pigeot 2012]. Another type of missing data is information on nonprescription medicine consumed by patients, such as supplementary vitamins, that are not reported to a database such as PBS or often even patient health records. An additional limitation of this type of data is that it could contain incorrect timestamps of medical events for some patients due to a phenomenon called *registration event dropping*. Registration event dropping occurs when a patient moves from one general practice to another, and the patient's previous events and conditions are recorded with current dates in the new general practice [Reps et al. 2012]. Furthermore, while some individual databases are already linked and coded, reliably linking different databases together can be time consuming and expensive, especially when a database uses a proprietary coding system.

4.3. Electronic Health Records

According to the International Organization for Standardization (ISO) [2005], an EHR is defined as a "repository of information regarding the health status of a subject of care, in computer processable form." Unlike administrative health databases, which are created for administration purposes, EHRs are designed to capture low-level health information of individuals from a health care perspective, which can also be shared and processed electronically.

Some benefits of an EHR include quick access to all available clinical information on a patient at the point of care, increased efficiency, cost savings, and improved safety and quality of care [Cusack 2008]. EHRs can contain both coded and free text narratives, such as *clinical notes* written in a clinical setting. Coding the narratives with the Unified Medical Language System (UMLS) [Bodenreider 2004] helps standardize the data for applying statistical methods to remove confounding associations [Friedman 2009]. Competitions, such as the i2b2[8] challenge, also provide manually annotated clinical notes that can be used to train supervised machine-learning algorithms [Doan et al. 2012]. EHRs when linked to computerized clinical databases are useful for detection of ADEs resulting from specific drugs [Hannan 1999; Patel and Kaelber 2013; Ramirez et al. 2009].

Although EHRs are carefully prepared, they still contain inaccurate and incomplete information that may lead to serious medication errors [Wuerdeman et al. 2005]. In addition, studies that include more than one EHR database—such as EU-ADR—face challenges, such as dealing with different coding systems [Valkhoff et al. 2014], dealing with free-text search with potentially different languages, and different health care

[7]http://www.thin-uk.com/.

[8]Informatics for Integrating Biology and the Bedside: https://www.i2b2.org/.

systems, as well as cultural, ethical, governance, and political issues [Coloma et al. 2011].

4.4. Medical Literature

Medical literature has been used for a long time to communicate the findings of medical researchers. Medical literature is usually free text written in a formal language. Other forms of medical literature include publicly accessible resources such as DrugBank,[9] Side Effect Resource (SIDER) [Kuhn et al. 2010], and Hazardous Substances Data Bank (HSDB).[10] Since this literature is easily accessible and growing, various analyses have been performed on it, focusing on medical articles [Wang et al. 2011], letters to the editor in medical journals [Yang et al. 2012b], and article metadata, such as the MeSH terms used when indexing them [Wang et al. 2011; Avillach et al. 2013], and integrating them in novel ways.

Medical literature along with other publicly available data sources has been used to generate useful resources for identifying and predicting ADEs. For example, to create a public database of drug side effects called SIDER, Kuhn et al. [2010] curated data from five public sources: the British Columbia Cancer Agency, Facts@FDA, FDA Center for Drug Evaluation and Research (CDER), FDA MedWatch, and Health Canada (DPD). SIDER covers 888 drugs and 1,450 distinct side effects including 62,269 drug/side effect pairs. Using SIDER and other published data sources including DrugBank and HSDB, Atias and Sharan [2011] proposed an algorithmic framework and showed that ADEs of new drugs can also be predicted.

Although many structured datasets have been developed from the medical literature, they are often not sufficient to be used for training, optimization, or evaluation of techniques to detect drug side effects. In order to tackle this problem, Gurulingappa et al. [2012] have annotated 2,972 randomly selected reports from PubMed case reports for drugs, adverse effects, doses, and their relationships and used them to evaluate different methods for identifying ADE sentences [Gurulingappa et al. 2011].

The main disadvantage of this type of data source is its mostly unstructured nature and thus, as for clinical notes, the complexities of dealing with natural language.

4.5. Social Media

The Web 2.0 has changed the way people interact and how information is being disseminated. People spontaneously discuss their thoughts, opinions, and details of their lives, including their medical situations (e.g., how they are feeling, the medications they are taking, how they react to them, etc.) on public websites such as medical forums, Facebook, Twitter, or other social media sites. This results in the generation of a vast amount of information that could be harnessed for early detection of ADRs. The public, in fact, already harnesses this information pool for a variety of purposes. Already in 2009, a Pew survey [Fox and Jones 2009] found that 61% of American adults looked for health information online, including for information about specific diseases and treatments, and the majority of these *e-patients* accessed user-generated health information. For example, 41% had read someone else's commentary or experience on an online news group, website, or blog. Furthermore, the study found that 30% of them were "actively writing or creating new health content." These numbers are likely to increase with the growth of social media sites [Chee et al. 2011]. In our previous work [Colineau and Paris 2010], we found that 85% of our survey participants were seeking information about their medical condition online.

[9]http://www.drugbank.ca/.

[10]http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB.

This is an important development because information about ADRs, although noisy, is likely to appear on these sites a long time before it reaches other data sources mentioned earlier. People using social media often seek advice on topics that they may be hesitant to discuss with their doctor, including adverse effects of some medications, especially those prescribed for serious conditions such as cancer, where the patient can experience high levels of anxiety due to the characteristics of the disease and the long-term exposure to potentially toxic drugs [Benton et al. 2011; Leaman et al. 2010]. Another advantage is that it provides rich data that can be used to learn more about a given population without relying on costly methods of data collection, such as focus groups [Benton et al. 2011]. Researchers have looked at mining these types of datasets, including Twitter data, message boards, consumer reviews in drug-related social media sites, and medical forums. They are briefly reviewed next.

Twitter is a microblogging site that allows its users to publicly send small messages. In a study of Twitter data, about 0.08% of Twitter messages (or *tweets*) were identified as health-related tweets, pointing at the potential of a microblogging service such as Twitter in applications related to public health, such as syndromic surveillance [Paul and Dredze 2011].

Consumer reviews of drug social media sites provide a forum for consumers to publicly discuss medical products. Liu et al. [2011] envision a future in which users would be able to explore the associations of prescription drugs with possible side effects, where the associations would be drawn from harnessing patient-submitted reviews, and thus using patients' terms (potentially in addition to other sources).

Using this data, they created a hierarchical ontology of adverse effects and analyzed the associations of side effects with statin drugs. Their results were similar to those found in previous studies such as Cable [2009].

Also using consumer reviews on social media sites, Yates and Goharian [2013] extracted ADEs of breast cancer drugs to identify whether the extracted ADRs were expected or unexpected. In a similar work, Mao et al. [2013] looked for mentions of side effects in cancer message boards. In this work, over 1 million posts were extracted from 12 cancer message boards between 2002 and 2010, with the majority of posts coming from http://breastcancer.org. These data were used to perform a quantitative analysis of the frequency and associations between aromatase inhibitor drugs and side effects. They note that the observed relative frequencies of key side effects mostly reflected those reported from large clinical trials. In their study, thus, the analysis of social media did not so much discover new side effects as provide more evidence of already known ones. Yet, their work shows that people do discuss side effects on online message boards.

Medical Forums. Medical forums are online sites where people discuss specifically their health concerns in the form of posted messages. These are typically hierarchical and contain several categories, each of which may have several topics. Within each topic, each discussion is called a thread, which may involve many participants. Each thread contains posts, which are user-submitted messages that contain the user's details and the date and time they were submitted. Unlike other types of social media such as Twitter, an additional advantage of medical forums is that they are specifically health related.

Leaman et al. [2010] proposed to mine patients' comments on health-related web sites, specifically DailyStrength,[11] to find relationships between drugs and adverse side effects. Similar to others, they note that texts from social media has irregularities, including misspellings, colloquial phrases, and even novel phrases, where users would

[11]http://www.dailystrength.org.

deliberately exercise a high level of linguistic creativity. Following from their work, we previously proposed to mine patient forums for *both* adverse and beneficial drug side effects, along with background information on the patients that could contribute to their positive/negative experience [Karimi et al. 2011].

In a related work, Chee et al. [2011] hypothesized that drugs that have undergone regulatory actions (which they called a *watchlist* drug—i.e., drugs that might be dangerous) are talked about in similar ways on medical forums—particularly with regard to sentiments.

Challenges. There are several main challenges with social media data sources. One is, again, dealing with unstructured data, in particular free text. While analyzing other data sources also had this challenge, as noted before, it is heightened for social media. There is a large language gap between medical documents and patient vocabulary, which can cause confusion and misunderstanding [Zeng et al. 2002]. In addition, the language employed in social media often uses colloquial language, nonstandard abbreviations or grammatical forms, misspellings, and a wide variation in syntax, vocabulary, and spelling [Leaman et al. 2010].

A second challenge lies in the level of noise that exists in social media, so that obtaining useful information is challenging. It might be the case that medical forums and consumer reviews constitute the best social media sources of information for the detection of adverse effects, while microblog services such as Twitter are more useful to find population-level health signals, as noted by Paul and Dredze [2011], who argue that the aggregation of millions of posts, as opposed to individual posts, has informational value. Another level of noise comes from the fact that ADE reports in social media are often statistically noisy, "full of biases and confounding factors that may not be easily identifiable" [Savage 2012, p. 12]. For example, the *Vioxx effect* captures the phenomenon in which people reported an increasing number of heart-related side effects for the painkiller Vioxx when a link was discovered between Vioxx and heart attacks. This problem also exists in ADE reporting systems, but it is more significant in social media. Some additional challenges include the process of de-identification, which in this case has to deal with user names instead of proper names and dealing with temporal relationship between posts [Benton et al. 2011].

Finally, interpreting the results might be challenging, as the volume of posts may not necessarily reflect prevalence of the side effect, but rather how pertinent it was to the individuals involved [Mao et al. 2013] or other factors (similar to the Vioxx effect). In addition, many factors could contribute to a side effect (e.g., specific demographics or other drugs taken at the same time), so that isolating the adverse effect symptoms for a specific drug might be very difficult.

4.6. Search Engine Logs

Search engines such as Google, Yahoo, and PubMED collect large logs of queries entered by their users. These logs represent a big picture of information needs of millions of users, complete with their interactions with the search systems to fulfill those needs. These logs have been studied in the area of information retrieval for a number of years for different purposes, such as improving their search engine results.

Query logs can provide information about health matters, and drug interactions in particular. Given the popularity of search engines among people of different demographics, their logs represents what information people around the world are looking for, including their search for specific drugs and reactions. In Section 5.4.4, we review the existing studies that use search logs for ADR detection. It is worth noting that access to such data is restricted to mostly commercial search engine owners, making it difficult for other researchers to study this data source.

5. METHODOLOGIES

A number of data and text mining methods have been developed to assist in the discovery of drug adverse effects from the data sources described previously. Some of these methodologies deal with spontaneous reporting systems for passive ADR monitoring. Increasingly, however, newer methods are being developed for *active monitoring* based on other types of data, such as administrative databases, medical literature, drug discussions in social media, and electronic health records. As the amount of computer-accessible ADR-related data grows, methods that integrate different types of data for ADR discovery become increasingly important. In this section, we review these methods.

5.1. Signal Detection Techniques for Spontaneous Reporting Systems

This section covers two major topics in ADR signal detection: measures of signal detection in SRS data sources and duplicate detection techniques in SRS data sources.

5.1.1. Overview of Signal Detection in SRS. The main approach for identifying drug safety signals from the data in a reporting system such as FAERS is to detect the *disproportionality* of reports about a given drug's adverse effect among all reports of the same adverse effect. The following three categories of methods are commonly used for this purpose:

(1) *Direct disproportionality measurement,* which uses the observed frequency in the database to detect potential signals. Proportional reporting ratios (PRRs) [Evans et al. 2001] and reporting odds ratios (RORs) [Stricker and Tijssen 1992; Sakaeda et al. 2013] belong to this type of methods.
(2) *Information theory-based measurement,* which calculates the association strength between drugs and adverse reactions. The Bayesian confidence propagation neural network (BCPNN) [Bate et al. 1998; Lindquist et al. 2000] uses mutual information [Pearl 1988] in information theory to define an association strength and identify potential signals.
(3) *Baseline measurement,* which attempts to establish the baseline or null hypothesis frequency for a drug–ADE combination and use the baseline to detect potential signals using relative report rate, statistical significance test, and empirical Bayes [DuMouchel 1999; Slade et al. 2009].

All of these methods deal with the following two aspects:

—How to identify the frequency difference of a drug–ADE pair in the database and
—How to deal with the sampling variability of drug–ADE pairs.

We use n_{ijk} to denote the number of reports in the database about the adverse effect j of drug i in report category k. A category is a subset of reports in the database. As an example, k may represent a set of reports about a particular age and gender group. For simplicity, we discuss signal detection methods within a single category later. We use n to denote the total number of reports in the database that fall into the category under concern. We use n_{ij} to denote the number of reports about the adverse effect j of drug i in this category. n_i is the total number of reports about drug i, and n_j is the total number of reports about adverse effect j in this category. This is summarized in Table II.

5.1.2. Direct Disproportionality Measure. PRRs measure the proportional ADR reporting ratio as follows:

$$PRR = \frac{n_{ij}/n_i}{(n_j - n_{ij})/(n - n_i)}.$$

Table II. A 2 × 2 Table for Disproportionality Calculation

	Reports with ADE j	Reports Without ADE j	Total
Reports with drug i	n_{ij}	$n_i - n_{ij}$	n_i
Reports without drug i	$n_j - n_{ij}$	$n - n_i - n_j + n_{ij}$	$n - n_i$
Total	n_j	$n - n_j$	n

The PRR value shows the disproportionality of a given (i, j) pair in the database. To characterize the sampling variability, the 95% confidence interval of PRR is given as follows [Agresti 2007]:

$$e^{\ln(PRR) \pm 1.96\sqrt{\frac{1}{n_{ij}} - \frac{1}{n_i} + \frac{1}{n_j - n_{ij}} - \frac{1}{n - n_i}}}.$$

ROR is also a direct measurement of disproportionality. It is defined as follows:

$$ROR = \frac{n_{ij}/(n_j - n_{ij})}{(n_i - n_{ij})/(n - n_i - n_j + n_{ij})}.$$

The following is the 95% confidence interval of ROR:

$$e^{\ln(ROR) \pm 1.96\sqrt{\frac{1}{n_{ij}} + \frac{1}{n_i - n_{ij}} + \frac{1}{n_j - n_{ij}} + \frac{1}{n - n_i - n_j + n_{ij}}}}.$$

5.1.3. Information Theory Based Measure. BCPNN measures the strength of association between a drug and an adverse effect using the Bayesian approach. It uses *mutual information* in information theory to characterize the association between drug i and ADE j. BCPNN introduces the information component (IC) as follows to measure the amount of information obtained about i when knowing information about j:

$$IC = \log_2 \frac{P(i, j)}{P(i)P(j)},$$

where $P(i)$ is the prior probability that drug i is in a report, $P(j)$ is the prior probability that adverse effect j is in a report, and $P(i, j)$ is the probability that drug i and adverse effect j are in the same report. As the real prior probabilities are not known, BCPNN estimates them based on observed data in reporting systems and then uses the estimated probabilities to calculate IC. Bate et al. [1998] assume a beta distribution for each probability. The confidence level of the estimation can be derived from the variance of IC as

$$var(IC) \approx \left(\frac{1}{\log 2}\right)^2 \left[\frac{n - n_{ij} + \gamma - \gamma_{11}}{(n_{ij} + \gamma_{11})(1 + n + \gamma)} + \frac{n - n_i + \alpha - \alpha_1}{(n_i + \alpha_1)(1 + n + \alpha)} + \frac{n - n_j + \alpha - \alpha_1}{(n_j + \alpha_1)(1 + n + \alpha)}\right],$$

where α_0 and α_1 are parameters of the beta distribution of $P(i)$ and $P(j)$, respectively, and γ_{11} and γ are parameters for the joint probability $P(i, j)$. These parameters are learned through an artificial neural network (ANN). The network contains two layers, with one layer representing drugs and the other representing adverse effects in the database. The interconnections between the two layers represent suspected associations between drugs and adverse effects.

BCPNN is used by the WHO to find potential signals in its VigiBase [Lindquist 2008]. In a typical operation scenario, the database is scanned on a quarterly basis. First, a set of cases excluding the reports received in the latest quarter is selected from VigiBase. This set contains typical reports received in a quarter. The IC and its associated 95% confidence limit are calculated. Second, the corresponding IC and

associated 95% confidence limit are calculated for drugs and adverse drug reactions of reports received in the latest quarter. The two ICs and lower 95% confidence limits of ICs are then compared. A signal is found if the lower 95% confidence limit of the IC of a drug–ADE pair changes from a negative to a positive in the latest quarter.

This method can also be used to examine the relationship between a group of drugs and the same ADE [Bate et al. 2002]. These groupings can be established, for example, using the WHO Anatomical Therapeutic Chemical (ATC) classification.

5.1.4. Baseline Measure. A baseline can be defined as the expected count of a drug–ADE pair (i, j) as

$$E_{ij} = \frac{n_i n_j}{n}.$$

The definition assumes i and j are independent. Even though i and j are in fact associated, the comparison between n_{ij} and E_{ij} is still useful for identifying potential signals. Relative risk (RR), also known as relative report rate [DuMouchel 1999], is such a measurement:

$$RR_{ij} = \frac{n_{ij}}{E_{ij}}.$$

This measure suffers from sampling variability when the baseline and observed counts are small. To address the problem, it is common to calculate the probability that the observed count n_{ij} is within a given range surrounding E_{ij} according to a given distribution.

Multi-item Gamma Poisson Shrinker (MGPS) [DuMouchel and Pregibon 2001; Fram et al. 2003] uses an empirical Bayes approach to compare n_{ij} and the baseline frequency for identifying potential signals. It attempts to achieve two goals simultaneously: making the relative risk measures easy to interpret and being able to adjust for data sampling variation. MGPS consists of two steps: computing a baseline frequency between i and j and comparing n_{ij} and the baseline frequency. When n_{ij} is sufficiently greater than E_{ij}, the association between i and j may be interesting.

Empirical Bayes assumes each observed n_{ij} is drawn from a Poisson distribution with unknown mean μ_{ij} and $\lambda_{ij} = \mu_{ij}/E_{ij}$. Furthermore, λ_{ij} is assumed to be drawn from a common prior distribution. The common distribution mixes two gamma distributions. The probability density function (PDF) of λ is as

$$\pi(\lambda, \alpha_1, \beta_1, \alpha_2, \beta_2, P) = Pg(\lambda; \alpha_1, \beta_1) + (1 - P)g(\lambda; \alpha_2, \beta_2),$$

in which g is the PDF of gamma distribution; α_1 and β_1 are the shape and rate parameter of the first gamma distribution, respectively; and α_2 and β_2 are the shape and rate parameter of the second gamma distribution, respectively. P is the probability that λ is from the first gamma distribution. The posterior distribution of λ after observing n reports is

$$\lambda | X = n \sim \pi(\lambda; \alpha_1 + n, \beta_1 + E, \alpha_2 + n, \beta_2 + E, Q_n), \tag{1}$$

in which Q_n is the posterior probability that λ is from the first gamma distribution after observing $X = n$. The expectation of λ can be calculated as follows based on Equation (1):

$$E[\lambda | X = n] = \frac{Q_n(\alpha_1 + n)/(\beta_1 + E) + (1 - Q_n)(\alpha_2 + n)}{(\beta_2 + E)}.$$

For the convenience of comparison, the expectation of $\log(\lambda)$ is computed as follows:

$$E[\log(\lambda) | X = n] = Q_n[\psi(\alpha_1 + n) - \log(\beta_1 + E)] + (1 - Q_n)[\psi(\alpha_2 + n) - \log(\beta_2 + E)],$$

in which $\psi(x)$ is the digamma function.

Table III. A Comparison of ADR Signal Detection Methods

Method	Disproportionality Measure	Sampling Variability
PRR	The disproportionality is easy to interpret.	Significance test of observed frequencies.
ROR	The disproportionality is easy to interpret.	Significance test of observed frequencies.
BCPNN	The IC value does not directly show disproportionality, but the change of IC may indicate a signal.	Beta distribution is assumed for prior probabilities and the variance of IC can be derived from this assumption.
RR	The disproportionality is easy to interpret.	The baseline is assumed to be the mean value of a given distribution, and a statistical test is used to compute the significance of difference between observed frequencies and the baseline.
MGPS	The EBlog2 value does not directly show disproportionality, but the value can be used to rank the ADEs for signal assessment.	A mixture probability model to make the EBlog2 value adjustable to sampling variability.

An empirical Bayes measure in MGPS is therefore defined as follows:

$$EBGM_{ij} = 2^{EBlog2_{ij}},$$

and $EBlog2_{ij}$ is

$$EBlog2_{ij} = E[\log_2(\lambda)|X = n_{ij}].$$

The parameter vector $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, P)$ is estimated using maximum likelihood that maximizes $L(\theta)$ as

$$L(\theta) = \prod_{ij} Pf(n_{ij}; \alpha_1, \beta_1, E_{ij}) + (1 - P)f(n_{ij}; \alpha_2, \beta_2, E_{ij}),$$

in which f is defined as follows:

$$f(n; \alpha, \beta, E) = (1 + \beta/E)^{-n}(1 + E/\beta)^{-\alpha} \times \Gamma(\alpha + n)/\Gamma(\alpha)n!$$

5.1.5. Comparison of SRS Signal Detection Methods. Table III compares the methods described previously from the following two perspectives: disproportionality measurement and sampling variability handling. PRR, ROR, and RR are direct indicators of disproportionality, while BCPNN and MGPS produce an indirect disproportionality metric. Direct indicators make the results easy to interpret and are popularly used to quickly produce potential safety signals. Indirect measurements often have relatively complex ways of dealing with sampling variability and are often used for signal detection in large SRS databases such as WHO and FDA reporting systems.

5.1.6. Duplicate Detection in SRS Databases. Duplicates in spontaneous reports are the multiple reports of the same ADE that either are generated by reports from different sources or are artifacts of follow-up reports that are not linked to the earlier ones.

Duplication is a significant problem in SRS databases. For example, Nkanza and Walop [2004] report a 5% duplication rate in vaccine adverse event data. Duplication affects the quality of signal detection using the methods described earlier as it biases the estimates of drug–event associations. Duplicate detection is therefore an important step for improving data quality in SRS databases. Norén et al. [2005, 2007] give a method to address the problem. The match score for a field in two reports i and j is defined as the following log-likelihood ratio:

$$W_{ij} = \log_2 \frac{p_{ij}}{p_i p_j},$$

in which p_i and p_j denote the probability that the field has a value i and j, respectively.

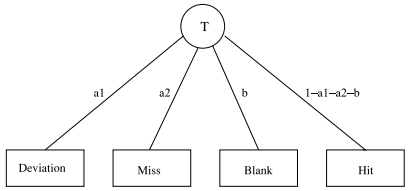

Fig. 3. The hit–miss model for duplicate detection.

The *hit–miss model* shown in Figure 3 is given to calculate the score. The model assumes that a true but unobserved value T may generate a random variable X in a report through a reporting process. The process has a probability a_1 of generating X that deviates from T, a probability a_2 of generating X that is totally different from T, and a probability b of generating X that is blank (no content). The probability that X is equal to T is therefore $(1 - a_1 - a_2 - b)$.

Let $P(T = t) = \beta_t$; the match score can be calculated as follows on this model:

$$W_{ij} = \begin{cases} \log_2 c - 2\log_2(1-b) & i \neq j \\ \log_2\{1 - c(1-\beta_t)(1-b)^{-2}\} - \log_2 \beta_t & i = j \\ 0 & i \text{ or } j \text{ blank} \end{cases} \quad (2)$$

in which $c = (a_1 + a_2)(2 - a_1 - a_2 - 2b)$. When comparing i and j, there are the following six scenarios:

(1) i and j are both equal to the true value t.
(2) i is equal to t, but j deviates from t.
(3) Both i and j deviate from t.
(4) i is equal to t, but j is different from t.
(5) Both i and j are different from t.
(6) i deviates from t and j is different from t.

The difference between i and j in each scenario follows a certain distribution; for example, scenario 1 uses Dirac's delta function, while scenarios 2 and 3 use a normal distribution. The probability density function for scenarios 4 to 6 is estimated from the data. b is estimated by the relative frequency of blanks in the whole database. Other parameters are estimated using the EM algorithm.

Norén et al. [2005, 2007] also proposed a modification of the hit–miss model to improve model fitting for ADR data and handle correlated fields in a report. Handling duplications in SRS data is not widely studied. One avenue of research would be combining the hit–miss model with duplication detection methods in text. While most information in SRS databases is either numerical or categorical, there can be *narrative* information that contains textual data requiring specific text processing methods.

5.2. Signal Detection Techniques for Electronic Health Records

Electronic health records contain both structured and unstructured data. The structured sections have been used directly to detect drug safety signals. Park et al. [2011] used the laboratory reports in an EHR database to identify abnormal laboratory results. The laboratory result abnormalities are defined as hypertype and hypotype. For hypertype abnormalities, the maximum value of the test results was used for the comparison, whereas for hypotype abnormalities, the minimum value was used. They compared the laboratory results from before and after the use of a medication. The paired t-test and McNumar's test were performed on the data.

Liu et al. [2013] applied statistical measures commonly used in AERS to medication order data and abnormal laboratory results. The goal of the study was to correlate abnormal laboratory results with specific drug administration. The six signal detection algorithms used were Chi squared test, PRR, ROR, Yule's Q, Bayesian confidence propagation neural networks, and Gamma Poisson Shrinker. The authors created an evidence-based reference standard and also used an existing reference standard to evaluate the different algorithms. The performance was calculated in terms of precision, recall, and F-score. The ROR method had the best performance in both datasets with F-scores of 62% and 68%.

Patel and Kaelber [2013] also used medication order data and abnormal laboratory results but focused only on the prescription of the drug azathioprine (AZA). The data were obtained from the Explorsys platform, a system that collects EHR data from different health care systems and maps it to ontologies. The laboratory measurements were used to assess the presence of side effects. The authors then built a side effect network using the conditional probabilities of developing a side effect. The analysis of this network revealed some results consistent with existing literature and others that were missing from previous work.

The unstructured sections of the EHR, mainly the free-text narratives, have also been used to detect drug safety signals. The main approach followed by most researchers is to apply Natural Language Processing (NLP) techniques to transform the free text into some form of structured data and then use the same statistical signal detection methods described in Section 5.1.

The transformation from free text to a structured representation suitable for statistical analysis is typically composed of several steps. First, the free text is annotated with concepts in a medical terminology. LePendu et al. [2013] created a lexicon from several existing medical terminologies. They annotated the clinical text using regular expressions. Wang et al. [2009a, 2009b], Chen et al. [2008], and Friedman [2009] used the MedLee NLP system for annotations. Then, the resulting annotations are filtered to exclude unwanted items such as negated terms or terms that apply to the family history. The remaining annotations are then normalized using the underlying terminologies. For example, in LePendu et al. [2013], "Rofecoxib 12.5 mg oral tablet" and "Vioxx" are normalized to the active ingredient rofecoxib using the relationships from RxNorm.[12] Finally, in some approaches, the timestamps in the EHRs are used to induce a temporal ordering over the recognized concepts for each patient [LePendu et al. 2013].

5.3. Signal Detection Techniques for Administrative Databases

A number of data mining techniques have been proposed to specifically capture infrequent and unexpected patterns in the data from administrative databases. We divide these techniques broadly into those that use unsupervised clustering algorithms and those that use supervised classification methods and rule mining techniques.

Clustering. He et al. [2004] proposed a clustering algorithm to signal for possible ADRs from a linked database of hospital admissions from Queensland Health and the Australian PBS data (details on data in Section 4.2). The idea was to compare temporal sequences of drug usage events using general dissimilarity and an event-based similarity measure based on standard drug classification hierarchy. They transformed the data into temporal sequences of the form $S = <(e_1, t_1), \ldots, (e_i, t_i), \ldots, (e_n, t_n)>$, where e_i represents an event that occurred at time t_i. Each event e_i was dispensing a given prescribed drug to a patient. They pointed out that it is important to have the data in

[12]http://www.nlm.nih.gov/research/umls/rxnorm/.

the sequence format so that the drug usage sequence is not lost if a static data format is used.

An important part of a clustering algorithm is the similarity measure used to define the clusters. For drugs, there are a number of parameters to consider: drugs can be the same but have different names due to different manufacturers, or they can have similar active ingredients, or they can belong to the same category. He et al. [2004] used the WHO classification called the ATC classification, where all the drugs are mapped to a tree. Each level of the tree presents a coding for drugs and divides drugs based on the organ or system on which they act or their chemical or therapeutic characteristics. Using this hierarchical coding, they defined similarities between events based on the assigned codes in each level of the ATC hierarchy. They then applied a uniform kernel KNN clustering [Cover and Hart 1967] and tested their proposed method on data from 222 patients with the disease angioedema. The technique showed promising results in discovering potential relationships between drugs and resulting hospital admissions due to adverse effects.

Classification and Rule Mining. Chen et al. [2005] proposed a modification of the optimal class association rule mining algorithm [Li et al. 2001] to discover the minimal set of rules that could potentially help medical professionals to identify the group of patients prone to experience an ADR more than others. The idea behind this rule mining algorithm is finding those rules that satisfy certain requirements, such as the minimum support in the dataset and the minimum confidence, rather than finding all the rules from a database and then pruning them [Li et al. 2001]. Chen et al. [2005] introduced two modifications in order to find patterns of a minority class, which was the patients with the potential to experience an ADR, rather than a majority class, which would not experience the ADR for a particular drug. To achieve this, two factors were introduced: local support and risk ratio.

Local support was defined as $lsup(A \rightarrow C) = \frac{sup(A \rightarrow C)}{sup(C)}$, where $sup(C)$ is the support or proportion of class C in the entire population, and $sup(A \rightarrow C)$ is the support or proportion of pattern A in class C. The risk ratio is defined based on the local support as

$$RiskRatio(A \rightarrow C) = \frac{lsup(A \rightarrow C)sup(\overline{A})}{lsup(\overline{A} \rightarrow C)sup(A)},$$

where $\overline{A}$ denotes lack of presence of pattern A. Risk ratio shows the relative risk of belonging to class C. Each rule was defined based on a list of parameters used for the classification. These parameters were extracted based on the linked data (the same data used by He et al. [2004]) and included patient-specific features such as age and gender and drug-specific features. Patients were considered if they had taken a specific drug of interest in a time window of 180 days.

Once all the rules were extracted, the set of mined rules was reduced and a threshold was set based on the *RiskRatio* value of each rule to be included. Chen and colleagues further evaluated the remaining rules using the Fleming-Harrington survival analysis statistical test [Fleming and Harrington 1992]. This analysis is designed for censored survival data in cohort studies where some data may be absent due to being lost, becoming unavailable, or subject withdrawal. They also suggested a tree-based presentation of the rules to help in clinical decision making.

Jin et al. [2006], following Chen et al. [2005], proposed a new data mining method called the Mining Unexpected Temporal Association Rules Algorithm (MUTARA). Its main advantage over previous methods was that it handled unexpectedness by excluding expected patterns for a given drug from the time period of interest. The criticism over the previous method by Chen et al. [2005] was that, because it heavily relied on

the risk ratio, which was already used in premarket testing, it failed to pick up new adverse effects in administrative databases.

MUTARA introduced a new knowledge representation called unexpected temporal association rules (UTARs). UTARs are rules created on patterns of two sets of event types: antecedent (A) and consequent (C). These event types are separate from each other in a time period. If such pattern combination is observed in the data, then a potential adverse drug reaction is detected.

A temporal association rule (TAR) is presented as $A \xrightarrow{T} C$, which is similar to the work by Chen et al. [2005] for the definition of A (a drug is prescribed) and C (an adverse event happens), and it explicitly introduces a time constraint T. The introduction of T ensures that A and C occur within the time window T, implementing the real-life implication that ADRs are normally short acting. To account for the unexpectedness of ADRs and to distinguish them from expected adverse events, UTAR is defined as $A \overset{T}{\hookrightarrow} C$, which means an unexpected event C happens after A in timeframe T. They then defined *unexpected leverage* as

$$supp(A \overset{T}{\hookrightarrow} C) - supp(A \overset{T}{\hookrightarrow}) \times supp(\overset{T}{\hookrightarrow} C).$$

This formula was used to rank UTARs. Only those with high value of unexpected leverage were considered to be inspected for ADRs.

Jin et al. [2008, 2010] further modified MUTARA and called the improved method HUNT (Highlighting UTARs, Negating TARs). Unlike MUTARA, HUNT accounts for *therapeutic failures*, which were not distinguished using MUTARA. Therapeutic failures appear in high ranks if only unexpected leverage is used for ranking ADRs. To compensate for this, a rank ratio metric is introduced as

$$RankRatio(A \overset{T}{\hookrightarrow} C) = \frac{\text{rank based on leverage of } A \xrightarrow{T} C}{\text{rank based on unexpected leverage of } A \overset{T}{\hookrightarrow} C}.$$

Chen et al. [2010] proposed the Multiple Occurrence of Target events Mining (MOTM) algorithm that takes the previous work (MUTARA and HUNT) one step further to mine for multiple occurrences of events. These events can be both adverse and beneficial. They presented an interestingness measure called *consequency,* which can be used to determine if a newly introduced drug was beneficial (negative value), made no difference (zero), or caused adverse effects (positive value). This measure combines the Before-After-Control-Impact (BACI) [McDonald et al. 2000] designs and frequent pattern mining techniques. BACI designs are popular with epidemiologists who study the environmental impacts of events such as accidents. However, standard BACI designs cannot be readily used in ADR discovery as they require existing hypotheses and do not accommodate multiple occurrences of the target events. The consequency measure, therefore, was required to account for these shortcomings.

Norén et al. [2010] proposed a method called temporal pattern discovery (TPD) that uses similar concepts of the disproportionality measures for SRS databases (Section 5.1). They define an IC measure as IC_{Δ}[13] that is calculated for every adverse event of the drug of interest, in order to rank them for further investigation. Although this approach is different from the methods we explained before that use temporal association rules, the concept of events (taking a drug and experiencing an adverse event) that happen in a given period of time is similar. We therefore use the same notation for easier comparison. If A represents the drug of interest and C the

[13]Note that this measure is different to the IC measure explained for SRS data (Section 5.1).

event of interest, which happens during the time interval T after the drug is administered, then $n_{A\xrightarrow{T}C}$ is the number of prescriptions of the drug A followed by event C in time interval T. They then define the expected value for $n_{A\xrightarrow{T}C}$ as

$$E_{A\xrightarrow{T}C} = \frac{n_{A\xrightarrow{T}} n_{\xrightarrow{T}C}}{n_{\xrightarrow{T}}},$$

where $n_{A\xrightarrow{T}}$ is the number of prescriptions for the drug A in time internal T, $n_{\xrightarrow{T}C}$ is the number of occurrences of the event C in the time interval T, and $n_{\xrightarrow{T}}$ is the number of prescriptions in the time interval T.

The logarithm of the ratio of observed events to expected events, $log_2 \frac{n_{A\xrightarrow{T}C}}{E_{A\xrightarrow{T}C}}$, is then a disproportionality measure that, if positive, means that C often occurs and, if negative, means that C rarely occurs. This measure, however, presents a challenge in large-scale databases as it can become too sensitive and generate noise. To reduce this sensitivity, Norén et al. [2010] proposed to use the shrinkage technique, which led to the following information IC measure:

$$IC = log_2 \frac{n_{A\xrightarrow{T}C} + \frac{1}{2}}{E_{A\xrightarrow{T}C} + \frac{1}{2}}.$$

While IC measures the temporal association of a drug to an event, it only refers to one time interval, not accounting for time variations. To compensate for this, IC_Δ is proposed, which contrasts two different time intervals U and V. The expected number of prescriptions of A leading to C in time period U is

$$E_{A\xrightarrow{U^*}C} = E_{A\xrightarrow{U}C} \times \frac{n_{A\xrightarrow{V}C}}{n_{\xrightarrow{V}}}.$$

The assumption is that the observed-to-expected ratio between A and C is the same in U and V. Finally, IC_Δ is calculated as

$$IC_\Delta = log_2 \frac{n_{A\xrightarrow{U}C} + 1/2}{E_{A\xrightarrow{U^*}C} + 1/2}.$$

Norén et al. [2010] evaluated their proposed TPD method on the UK IMS Disease Analyzer data, which contains electronic health records of patients in the United Kingdom.

All the previous methods are evaluated on different databases, for different drugs and ADRs, and with different metrics, making it difficult to compare them. Admittedly, though, due to the nature of such work, it is not easy to draw conclusions as to which method generates more reliable signals with high coverage of common and rare ADRs. Recently, though, Reps et al. [2013] conducted a comprehensive experimental comparison of a number of ADR detection algorithms, including modified ROR, MUTARA, HUNT, and TPD, using data from The Health Improvement Network (THIN) database (details in Section 4.2). They evaluated these methods for 27 drugs that belong to seven different drug families. The main conclusions were that none of these algorithms are able to detect rare cases of ADRs and that there was no obvious superior method.

Fuzzy Decision Making. Pharmacovigilance practice often lacks causality assessment of a drug and potential ADRs [Anderson and Borlak 2011]. Such a relationship is not easy to establish because a number of parameters contribute to the observation of an event, and information on all these factors may not be present in the data itself.

For instance, information about over-the-counter drugs is often missing. Nonetheless, there is a small number of recent studies that intend to take the ADR detection process closer to causality assessment.

Ji et al. [2007, 2011, 2013] proposed a method to discover the possible *causal* relationship between a drug and an ADR. They argued that temporal association, as explained for the previous algorithms, is not sufficient for assessing the causal relationship of a drug and an ADR. They therefore proposed an experience-based model that incorporates multiple signals—including a temporal association, a dechallenge, a rechallenge, and other explanation—and produces a decision as *very likely*, *probable*, *likely*, and *unlikely*. Dechallenge is a clinical decision to withdraw or discontinue a medication after occurrence of a possible ADR [Alghabban 2004]. Rechallenge is to reintroduce a drug that was previously stopped for the suspicion of being responsible for an ADR [Alghabban 2004]. Ji et al. [2011] introduce a new knowledge representation as $A \xrightarrow{S} C$, where S is the potential causal relationship between the drug A and the event C.

The causal relationship between a drug and an adverse event is defined based on all the factors as

$$S_{<A,C>} = \sum_{i=1}^{m} \mu_i \times w_i,$$

where μ_i is the membership of the i^{th} causality category with the weight w_i. They then introduce an interestingness measure to rank the potential causal relationships. Evaluations in this work point out the same problem found in all the other previous studies. Apart from some existing established casual relationships, there is no other gold standard data to test the new methodology, and even if it finds new relationships, these will need further investigation.

5.4. Signal Detection Techniques for Unstructured Data

We review text mining methods used to process unstructured data sources such as medical literature and social media in this section. One important aspect of working with text in the medical domain is dealing with the domain-specific terminology. Medical terminology presents a number of difficulties, including ambiguity and variability [Tsuruoka et al. 2008]. This problem has been dealt with mainly by normalizing the terminology and mapping it to standard dictionaries.

5.4.1. Data Annotation and Semantic Mapping. Unstructured or semistructured data sources that are useful for ADR discovery include medical literature, plain text parts of various medical records and reports, and discussions on adverse drug effects in social media. The processing of this type of data often involves data annotation and semantic mapping to well-defined concepts and terms.

There are a few standard terminology systems commonly used in the biomedical domain. The Unified Medical Language System (UMLS) Metathesaurus[14] [Bodenreider 2004] is such a database of biomedical vocabularies developed by the U.S. National Library of Medicine. It intends to facilitate the data exchange and interoperability between biomedical information systems. UMLS contains a component called *semantic network* that categorizes these vocabularies and defines their relationships. It also provides a natural language processing tool, SPECIALIST, for application developers to analyze text and process lexical variations. MetaMap [Aronson 2001] is another tool that maps arbitrary biomedical text to concepts in the UMLS.

[14]http://www.nlm.nih.gov/research/umls/.

RxNorm contains normalized names for clinical drugs and links between these names and commonly used drug vocabularies in sources such as the ATC Classification System, Micromedex Red Book (MMX), Medical Subject Headings (MeSH), and U.S. Edition of Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT). A few tools use RxNorm for drug name extraction. Levin et al. [2007] built a drug name mapping and extraction system using RxNorm and Metaphone [Philips 1990], a phonetic algorithm that maps an input string to a string that sounds similar to the input string. Metaphone was used to produce a list of variants of a drug name, and these were matched with the input string to identify the drug name. Peters et al. [2011] employed a similar approach to identify drug names in input text, using the lexical variant generation tool available in SPECIALIST of UMLS [McCray et al. 1994] with additional rules to generate drug name variations. Once drug names in the input text are identified, Peters et al. [2011] further compare candidate strings in the RxNorm dataset with the tokenized input text to determine its terminology concept, with a string similarity measure based on the Jaccard coefficient.

In addition to extracting drug names from clinical notes, MedEx [Xu et al. 2010] also extracts structured contextual information associated with the medication. To do this, MedEx also exploits RxNorm to tag drug names in the input clinical notes and then uses regular expressions to capture information, such as frequency. A set of rules is defined to disambiguate tags. A top-down chart parser [Kay 1986] combined with regular expressions extracts the structured information from medication sentences.

There are some efforts on defining ADR-related terms. MedDRA[15] and its early version, Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART), are among them. Leaman et al. [2010] uses a lexicon that combines COSTART and a few other sources to extract ADR-related information from unstructured data.

Extracting ADR-related information from electronic health records involves a step that maps clinical terminologies to MedDRA terminologies. Bodenreider [2009] reports a method for mapping SNOMED CT to MedDRA terms. The mapping is done through UMLS. A term in SNOMED CT is directly mapped to a term in MedDRA through using either synonymy in the UMLS or explicit mapping relations such as descendants given in the UMLS. However, the overall mapping rate of MedDRA PT terms to SNOMED CT terms is only 58.3%, and only 12,843 concepts in SNOMED CT have a mapping to MedDRA terms.

5.4.2. Signal Detection from Medical Literature. Most techniques used to extract ADR information from medical literature are based on supervised machine learning. Wang et al. [2011] developed an algorithm to identify ADEs from articles in PubMed. First, documents are retrieved from PubMed by looking for keywords that represent a drug and a target adverse event. The retrieved documents are then classified as denoting an ADE or not using logistic regression with different ontological and textual features. Finally, all the positively classified documents are used to determine if the ADE is likely or not. This is done by computing a ratio of articles that were positively classified over all articles retrieved. If this percentage is high based on a threshold calculated automatically using ROC analysis, then the drug is considered to cause the target event.

Yang et al. [2012b] developed an algorithm to extract ADEs from letters to the editor. First, letters are retrieved using a PubMed search specifying "adverse effects" as the MeSH term, restricted to the ones published in the *Lancet* and the *New England Journal of Medicine* and excluding the ones labeled as "comment." A binary classifier is used to determine if a letter to the editor contains an ADE or not. Two sets of

[15]http://www.meddra.org/.

features are used as inputs to the classifier. The first set, referred to as statistical features, includes the syntactic and semantic labels created by MetaMap from the text. The second set includes n-grams created using dummy words to replace certain concepts in the sentences. For example, "In 1970 furosemide was reported to cause. . ." would be replaced by "In <num><drug> was reported to cause. . . ." The authors tested several machine-learning algorithms including ZeroR, Naive Bayes, Decision Tree, and Support Vector Machines (SVM) using a combination of features and found that the best performance was obtained using SVM and a combination of statistical features and n-grams up to 3.

Avillach et al. [2013] did not use a machine-learning method because they focused exclusively on the structured metadata from MEDLINE. First, the drugs and events are mapped to MeSH terms. Then, a query using these terms is executed and a list of matching documents is returned. A threshold number of publications is defined to inform the presence of the drug–event association in the literature. The mappings to MeSH are achieved using UMLS and ATC codes. The key aspect of the method is the way the query is constructed. Each query looks for the co-occurrence of four elements: the drug (in either the "substances" or "MeSH heading" fields), the adverse event, and two subheadings, anti-infective agents and chemically induced. Drugs in the "substances" fields are only considered if their pharmacological action is qualified by the heading AE, guaranteeing a link between the drug and the event in the context of drug safety.

Gurulingappa et al. [2012] used MEDLINE case reports to create an annotated corpus of text with information about the drug, adverse effect, dosage, and their relationships. In order to restrict the scope of the corpus to drug-related adverse events among 1.5 million case reports, a PubMed search with "drug therapy" and "adverse effect" as MeSH terms was performed limiting the language to English and excluding the documents with a title only and no abstract. The precise PubMed query performed was

```
"adverse effects"[sh] AND (hasabstract[text] AND Case Reports[ptyp])
AND "drug therapy"[sh] AND English[lang] AND (Case Reports[ptyp] AND.
("1"[PDAT]: "2010/10/07"[PDAT])).
```

Gurulingappa et al. [2011] used the following methods to generate features to use with various classifiers:

—All words: features were made of all words occurring in a sentence excluding special characters.
—Lemmatized tokens: Lemmatization is used to treat different variants of a word as the same word. Here the words in a sentence were lemmatized before being used as features. The authors used the Genia[16] tokenizer and MorphAdorner[17] English lemmatizer.
—Lexicon token match: a single word lexicon was created from DrugBank and MedDRA. The presence of these lexicons was checked in a sentence. The counts of drug matches and condition matches were used as special features.
—Lemmatized token bigrams: pairs of adjoining lemmatized tokens were used as features. The tokens that match a drug name or a condition name were renamed as DRUG and CONDITION, respectively.

[16]http://www.nactem.ac.uk/genia/.
[17]http://morphadorner.northwestern.edu/.

—Lemmatized token trigrams: similar to bigrams, three consecutive tokens were used together as features. The drug and condition names were treated in the same way as in lemmatized token bigrams.
—Noun character affixes: all the noun forms of the words occurring in a sentence were used to extract two-, three-, and four-character suffixes and prefixes to be used as features.
—Lemmatized verbs: whenever a drug name or condition name match was found, their immediate preceding and succeeding lemmatized verbs were used as features.
—Lemmatized tokens in window: for a drug name or condition name match, a window of size $n = 5$ lemmatized words around the match was extracted. These words were then used as features.
—Stanford token dependencies: the Stanford dependencies[18] parser was used to create a dependency pair in a sentence. The words were then lemmatized, and the pairs were used in the order based on their relationships.

Gurulingappa and colleagues used Naïve Bayes, Decision Tree, Maximum Entropy, and Support Vector Machine (SVM) classifiers and found that the Maximum Entropy classifier outperforms others.

5.4.3. Signal Detection from Social Media Data. To obtain useful health-related information from social media data, researchers typically employ a combination of natural language processing and machine-learning techniques (e.g., Bian et al. [2012]). Similar to dealing with semistructured and unstructured data, ADR discovery from social media faces the problems associated with processing free text and mismatching terminology. Natural language processing techniques are often first used to preprocess the data (e.g., Leaman et al. [2010]; Bian et al. [2012]). These include, for example, sentence splitting, tokenization, and part-of-speech tagging. As already mentioned, the noise in the text of social media renders the text processing tasks even more challenging than processing from other textual sources such as medical literature.

To address this challenge, additional preprocessing steps are often adopted, including [Liu et al. 2011] filtering out stop words, filtering out phrases that contain only stop words, filtering out phrases that have frequency counts less than five in the whole review dataset, and group phrases containing the same set of non-stop-words. To address the issue of misspellings, some have adopted a string similarity measure [Leaman et al. 2010].

To extract words related to drugs and symptoms, controlled vocabularies are constructed using existing resources, such as UMLS, the FDA's COSTART corpus, the SIDER database [Kuhn et al. 2010], and the Canada Drug Adverse Reaction Database (MedEffect). The ADR extraction methods are similar to those described earlier in this section to extract medical terms from text. To address the large gap between medical and layman language, these controlled vocabularies (which have low coverage of colloquial adverse effect expressions commonly seen in social media) are then often augmented with a set of colloquial phrases, created through automated scripts (i.e., through manually crafted patterns, e.g., Yates and Goharian [2013]), manual extraction (e.g., Leaman et al. [2010]), or both (e.g., Benton et al. [2011]).

Liu et al. [2011] built their own ADR ontology through a manual annotation process. The result is a hierarchical classification of ADRs that contains three levels: ADR class, ADR group, and ADR synonyms. The process they employed is as follows:

[18]http://nlp.stanford.edu/software/stanford-dependencies.shtml.

(1) Build an ADR ontology by grouping phrases into a hierarchical ontology.
(2) Identify and group synonyms; for example, "elevated blood pressure," "increase in blood pressure," and "higher blood pressure" are clustered into the same group.
(3) Organize synonym groups into broad classes; for example, "achy legs," "muscle pain," and "joint pain" are clustered into the class of "pain."

The resulting vocabularies or ontologies guide the extraction of terms as well as the association analysis of drugs and adverse reactions (e.g., Benton et al. [2011]; Yates and Goharian [2013]).

When dealing with generic social media (e.g., Twitter), some researchers have first built a classifier to identify which post was health related, before doing any further analysis [Paul and Dredze 2011]. Or, if focusing on a specific drug, they would first filter out all the tweets that did not mention the name of the drugs [Bian et al. 2012]. For the analysis, some looked for co-occurrences of specific words to identify correlations (e.g., Scanfeld et al. [2010]). Other researchers developed text mining systems using machine-learning tools to classify the remaining tweets as mentioning a side effect or not [Leaman et al. 2010; Bian et al. 2012], sometimes with limited results, but still demonstrating that generic social media such as Twitter could contain useful information on drug side effects.

Mao et al. [2013] analyze the adverse reactions of aromatase inhibitors, a specific class of cancer drug, among postmenopausal women with hormone-receptor-positive breast cancer. Having collected posts from breast cancer message boards, they associated ADRs and drugs based on co-occurrences of the following terms [Benton et al. 2011]. The approach first calculates the frequency count of each term in the corpus and then extracts association rules between pairs of terms. Pairs of terms are considered correlated if they co-occur more frequently than would be expected if they were distributed independently. These pairs are chosen if they are within a window of 20 tokens apart. For each correlated pair (A, C), a two-by-two table of the occurrence of A and C is constructed and a one-tailed Fisher's exact p-value is computed. The p-value indicates whether the two terms co-occur independently by chance. This automatic identification of associations is complemented with a manual qualitative analysis. From their results, the authors list the most commonly reported side effects for a specific class of cancer drugs and people's reactions to these side effects (such as switching drug and adherence behavior).

Some methods, such as log-likelihood ratios, can be used to extract summary information derived from biases in word and phrase distributions and to quantify associations between drugs and symptoms. The method by Liu et al. [2011] is based on the following observation on disproportionality: the number of reviews of a given drug containing a given ADR phrase is more than that of reviews of other drugs containing the same ADR phrase. A p-value test can be used to validate if an ADR term is significantly more likely to appear in the reviews of a given drug.

Isolating the adverse reaction symptoms to the drug by excluding factors that may contribute to the same symptoms can help address the *Vioxx effect* [Savage 2012]. As an example, a person who takes an arthritis drug may report complications that are caused by being elderly and not by the drug. The approach reported by Savage [2012] is to take patient information into account. For each drug, it identifies a group of patients who take the drug and another group of patients who are similar to the group but do not take the drug. By comparing the symptoms of the two groups, side effects most likely caused by the drug can therefore be inferred.

Bian et al. [2012] studied identifying ADEs in 2 billion tweets collected from May 2009 to October 2010. The process was divided in two stages: identifying potential users of the drug and finding possible adverse effects in the user's timeline. Five cancer

drugs were used to test the approach, and overall, 239 users were identified with a total of 489 tweets. All of the users' tweets related to the drugs were collapsed into a single document. The first stage was implemented using an SVM, trained and tested using manually labeled data. The second stage was also implemented using an SVM, but it used a different dataset: only the users identified as using the drugs in the previous stage were used, and the list of tweets was augmented with tweets before and after the ones mentioning the drugs. The features used in both stages included textual and semantic features. The latter were generated with MetaMap.

Chee et al. [2011] employed a set of classifiers organized as an ensemble to identify dangerous drugs, which are the drugs that might undergo regulatory actions based on their side effects, or *watchlist drugs*. Their work is based on the hypothesis that the language used to talk about such drugs is similar across drugs. They developed classifiers, using both Naïve Bayes classifiers and SVM, and organized them as an ensemble. Their system focuses on false positives, that is, drugs that the ensemble classifier identifies as a watchlist drug but that are not yet present in an official list of watchlist drug from the FDA.[19] Their work exploited 27,290 public Health and Wellness Yahoo! groups with a total of 12,519,807 messages, spanning 7 years. For their evaluations, they purposefully removed four drugs from the official watchlist list, labeling them as negative. Their system was able to identify them as watchlist drugs.

Yang et al. [2012a] studied signal detection from a medical forum called MedHelp using data mining approaches that are more popular in SRS or EHR data. They modified association rule mining algorithms that are based on interestingness and impressiveness. These are in turn measured by metrics such as support, leverage, and confidence (see Section 5.3 for similar work in administrative databases).

In their work, Yang et al. [2012a] defined support and confidence for the rule $A \rightarrow C$ or drug A being responsible for the presence of adverse effect C as

$$supp(A \rightarrow C) = P(A \cup C) = \frac{count(A \cup C)}{\text{total number of data records}},$$

where $count(A \cup C)$ is the number of co-occurrences of A and C in the data. Low support means low association between the drug and the ADE. Confidence is then defined as

$$confidence(A \rightarrow C) = P(C|A) = \frac{supp(A \rightarrow C)}{supp(A)},$$

where $supp(A)$ is the proportion of records with at least one occurrence of drug A. Leverage in turn is defined as

$$leverage = supp(A \rightarrow C) - supp(A) \times supp(C).$$

Similar to the previous work, leverage is used to rank the associations found in the data.

To process the forum data and calculate confidence and leverage, the authors had to find mentions of ADEs in the text. They used the Consumer Health Vocabulary (CHV), which is an electronic collection of health expressions by consumers rather than by health professionals. This makes it more suitable to process forum data than a thesaurus such as UMLS. Evaluations were done for five different drugs, and they used already known ADRs for those drugs as listed by the FDA. This work, compared to the algorithms applied to other data sources, is in a preliminary stage. It is also evaluated in a limited setting, which makes it hard to judge its value in gaining new knowledge. However, it shows again that useful information can potentially be extracted from social media.

[19]The watchlist can be found at http://www.fda.gov/Drugs/DrugsSafety.

Liu and Chen [2013] designed the AZDrugMiner system to identify ADEs from medical posts. The system first crawls the web looking for posts from patient forums. These are then preprocessed to remove noise and detect sentence boundaries. MetaMap is used to detect drugs and event names, and the ones that never appear in FAERS are filtered out. Also, the UMLS medical terms identified in the process are extended with the consumer terms available in the CHV. Once the events and the drugs have been extracted, the system identifies the relations between them using a two-step approach based on a kernel-based learning method. First, a dependency graph of each sentence is built and the shortest dependency path kernel-based machine-learning method is used to detect relations between drugs and events. Then, the relations are classified in one of the following classes: negated ADE, drug indication, and true ADE. One of the key aspects of the proposed system is the identification of the source of the report. This is useful because ideally, only ADEs that are true patient experiences should be considered (as opposed to hearsay, for example). The system classifies the reports using Transductive SVM (a semisupervised learning classification method) using different sets of features including bag of words, bigrams, and POS tags. The system was evaluated on data extracted from a well-known diabetes online community. The medical entity extraction task achieved a 92.5% F-measure for drugs and 83.6% for events. The ADE extraction task using relation extraction combined with relation classification achieved the best performance with an F-measure of 66.9% compared to 55.6% achieved using co-occurrence analysis. The source classification task achieved the best results using the bag-of-words approach with an F-measure of 84.1%. The reported ADEs were compared to the ones in FAERS. The authors found several advantages of ADEs found in social media. First, patients in social media are not biased to severe ADEs. Also, patient reports in social media are not biased to drugs with well-known events. Finally, the reports found in social media can capture the emotional reactions of patients to treatments.

5.4.4. Signal Detection from Search Engine Logs. White et al. [2013] used web search logs from millions of users who opted to share search activities from Google, Bing, and Yahoo with Microsoft through a browser add-on during the whole of 2010. First, the log data was mined to find a specific adverse drug–drug interaction that was unknown at the time, the interaction between paroxetine and pravastatin, which causes hyperglycemia. Queries were identified as relevant by using a list of terms related to hyperglycemia built manually based on a review of medical literature and the names of the drugs (paroxetine and pravastatin) plus a list of their trade name variants. The queries were divided into the following groups: searches of paroxetine, searches of pravastatin, and searches of both. Disproportionality analysis was used to assess the increased chance of a user searching for terms related to hyperglycemia terms given that he or she had searched for both pravastatin and paroxetine. Reporting ratios were computed based on observed versus expected adverse reports.

Yom-Tov and Gabrilovich [2013] used web search logs from Yahoo in the United States for 6 months in 2010. The authors propose an approach that works by calculating a query log reaction score (QLRS). One hundred and seventy-six million users were identified using a unique signature from the browser, and 195 symptoms from ICD-10 (International Classification of Diseases) were studied. These were filtered using the Wikipedia List of Medical Symptoms. Additional synonyms were added based on two query expansion methods (an example of the end result is diplopia → double vision). The basic method, in summary, works like this: for each drug, a list of users that searched for the drug and the users that did not search for the drug is created, and the Day Zero is calculated (first day when a user searched for the drug or the midpoint of the query history otherwise). For each drug–symptom pair, a two-way contingency

table is built counting the number of times a symptom was searched for before and after Day Zero. For each symptom, the QLRS is the result of the Pearson's goodness-of-fit test against the drug.

Both these studies highlight that information on potential adverse drug reactions exists in search engine logs. However, similar to other methods reviewed for social media, they are not sufficient to establish a causality relationship for a drug–ADE pair, and they remain in the level of raising alerts.

5.5. Data Integration Techniques for ADR Discovery

Currently, the discovery of drug–ADR associations is mainly based on spontaneous reporting systems in practice. As mentioned earlier, information extracted from these systems alone is not sufficient to identify a signal. Integrating data from other sources is likely to improve signal detection in spontaneous reporting systems. As an example, Harpaz et al. [2013] combine potential signals from FAERS with signals extracted from the narratives of EHRs to improve the accuracy of signal detection. The motivation behind this work is to validate the idea that the combination of the top k signals produced by two independent sources should be better than each source on its own. The system was evaluated by comparing the top signals produced by the FAERS data on its own and the signals from the intersection of the FAERS data and the EHR data. The results were evaluated against a reference standard manually created by a pharmacological expert and reviewed by three physicians. The evaluation showed that the F-measure of the combination of signals was better than the FAERS signals on their own, especially when looking for plausible ADRs.

Data integration can also help in establishing the cause and effect between a drug and an adverse reaction. In practice, data from different sources are often taken into account through a manual drug safety review process. It is desirable to have a system that integrates data from various sources to assist decision making. In this section, we discuss technologies along this direction.

5.5.1. Schema-Level Data Integration. Data integration has been studied for many years in the database area [Halevy et al. 2006]. Work in this area focuses on providing a uniform query interface to multiple data sources. Most of these efforts are based on relational databases and work at the schema level. In *schema-level data integration*, the system provides a mediated data schema for users to pose queries. A mapping between the mediated schema and schemas of data sources is needed to answer these queries. The main techniques for achieving this goal are Local-as-View and Global-as-View [Lenzerini 2002; Levy 2000]. The former describes the source schema as a view over the mediated schema, and the latter describes the mediated schema as a view over the source schema. In this approach, an essential step is to specify which elements in a schema correspond to which elements in another schema semantically. It is largely a manual process. There are many efforts attempting to automate the process by exploiting knowledge representation, machine learning, and information retrieval techniques [Rahm and Bernstein 2001].

5.5.2. Data-Level Data Integration. Another approach for data integration is at the data level, with an aim to match and aggregate records in different databases that refer to the same entity [Christen 2008, 2012]. The initial motivation for work in this area was mainly to improve the quality of integrated data from independently managed databases. *Duplicate record detection* is a major issue targeted by these studies. A database record often consists of multiple fields. Duplicate record detection techniques can therefore be divided into two levels: field matching and record matching. Field matching mainly concerns string similarity. Most similarity metrics are defined on a syntactical level, such as the Edit Distance [Levenshtein 1966], Hamming Distance

[Hamming 1950], and Jaccard Coefficient [Chaudhuri et al. 2006]. Record matching involves machine-learning techniques to learn similarity measures. The field similarities between two records form a comparison vector. Supervised decision tree induction and unsupervised clustering are used to determine if a record is likely to match a set of other data records [Elfeky et al. 2002]. Active learning is also used for record matching to reduce the amount of training data [Sarawagi and Bhamidipaty 2002].

5.5.3. Linked Data. There are significant efforts in using the web to create more meaningful links between data from different sources. *Linked Data* [Bizer et al. 2009] promotes linking data from the web through typed links rather than untyped hyperlinks. The representation of types relies on the *Resource Definition Framework (RDF)*. In the RDF model, data is represented as *subject, predicate*, and *object* triples. The predicate specifies how the subject and object are related. RDF triples can form sophisticated networks of data. A query language called SPARQL [Quilitz and Leser 2008] is used to query the interlinked data. In the medical domain, *Linking Open Drug Data (LODD)* [Samwald et al. 2011] is an effort to create *Linked Data representation* of publicly available data about drugs. However, interlinked data often suffer from the difficulty of composing queries on them due to the lack of a unified data model [Samwald et al. 2011]. The mapping of heterogeneous datasets can be done using the Linked Data framework, but understanding the mapping intuitively for a meaningful information search is complicated for both end-users and search algorithm designers.

5.5.4. Data Integration for Active ADR Monitoring. Data integration techniques are increasingly used in various platforms and applications such as Google Fusion Table [Gonzalez et al. 2010] and Google Public Data. In the ADR discovery application domain, there are some projects attempting to integrate data for drug safety monitoring. The Vaccine Safety Datalink (VSD) [Baggs et al. 2011] allows data sharing between the Centers for Disease Control and Prevention (CDC) and eight managed care organizations (MCOs) in the United States for postmarketing evaluation of vaccine safety. Each MCO maintains its own data including patient demographic and medical service information. The data sharing is performed through an intermediate called a "hub," where the CDC sends statistical programs to the hub, and MCO computers periodically retrieve these programs to run on their local data. The results are sent back to the hub for retrieval by the CDC. The VSD also links data from FAERS for newly licensed vaccines. Mini-Sentinel [Curtis et al. 2012] is a distributed data system sponsored by the FDA for active safety surveillance of medical products.

6. CONCLUSIONS AND OPEN RESEARCH AVENUES

Adverse drug events impose substantial costs on health care systems around the globe. Manual review of safety signal reports on drugs released to the market is not plausible, with its drawbacks reiterated in various studies. Computing technology is increasingly developed and applied to automate different stages of safety signal management.

At the core of ADR discovery is the assessment of the causal relationship between a drug and unexpected harm caused by the normal use of the drug at the normal dosage. Many technologies we discussed go as far as detecting potential ADR signals based on disproportionality. These potential signals are then examined in medication safety review and assessment meetings. The main task of these meetings is to establish the causality between a drug and its adverse reactions. This is mainly a manual process and often generates wide variability in assessment. It is challenging to fully automate the entire process. However, as ADR-related data becomes increasingly accessible, it becomes crucial to introduce more carefully designed computing algorithms to assist the causality reasoning process that may replace some manual steps in the assessment.

There are two important approaches to achieve this goal. First, it is essential to understand and capture well-designed reasoning processes in the existing practice. A good reasoning process tends to minimize the variability and inconsistency in assessment, as shown decades ago by Naranjo et al. [1981]. Second, integrating data from various sources is essential for reaching reliable conclusions in the reasoning process.

We reviewed different data sources and their corresponding data processing technologies for ADR discovery as well as emerging data integration technologies in this area. While the existing statistical and data mining methodologies for passive signal detection seem to have reached a level of maturity, active detection methods have room for improvement. Knowledge discovery methodologies in this area should efficiently handle large-scale data sources and be effective in discovering potential signals without producing much noise that requires manual assessment. At the same time, missing signals that exist in the data can cost lives and should be avoided. All these emphasize the importance of this field of study and the fact that both computer scientists and domain experts should collaborate for further advancement.

Evaluation of signal detection techniques is not straightforward. The methods we surveyed often rely on the current knowledge of adverse effects listed in the product information as the gold standard. Often such lists consist of the most common adverse effects already discovered in the clinical trials. Whether other drug–ADE relationships found by these methods are valid or not is not easy to assess.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

We thank Dr. Bronwen Harvey from the TGA for her very constructive comments on the draft of this article.

REFERENCES

Australian Commission on Safety and Quality in Health Care (ACSQHC). 2011. National Safety and Quality Health Service Standards. Sydney.

A. Agresti. 2007. *An Introduction to Categorical Data Analysis*. Vol. 423. John Wiley & Sons.

A. Alghabban. 2004. *Dictionary of Pharmacovigilance*. Pharmaceutical Press.

N. Anderson and J. Borlak. 2011. Correlation versus causation? Pharmacovigilance of the analgesic flupirtine exemplifies the need for refined spontaneous ADR reporting. *PloS One* 6, 10 (2011), e25221.

A. R. Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *AMIA Annual Symposium Proceedings*. 17–21.

N. Atias and R. Sharan. 2011. An algorithmic framework for predicting side effects of drugs. *Journal of Computational Biology* 18, 3 (2011), 207–218.

P. Avillach, J.-C. Dufour, G. Diallo, F. Salvo, M. Joubert, F. Thiessard, F. Mougin, G. Trifirò, A. Fourrier-Réglat, and A. Pariente. 2013. Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: A contribution from the EU–ADR project. *Journal of the American Medical Informatics Association* 20, 3 (2013), 446–452.

J. Baggs, J. Gee, E. Lewis, G. Fowler, P. Benson, T. Lieu, A. Naleway, N. P. Klein, R. Baxter, and E. Belongia. 2011. The vaccine safety datalink: A model for monitoring immunization safety. *Pediatrics* 127, Supplement 1 (2011), S45–S53.

A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas. 1998. A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology* 54, 4 (1998), 315–321.

A. Bate, M. Lindquist, I. R. Edwards, and R. Orre. 2002. A data mining approach for signal detection and analysis. *Drug Safety* 25, 6 (2002), 393–397.

D. W. Bates, R. S. Evans, H. Murff, P. D. Stetson, L. Pizziferri, and G. Hripcsak. 2003. Detecting adverse events using information technology. *Journal of the American Medical Informatics Association* 10, 2 (2003), 115–128.

A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard, and J. H. Holmes. 2011. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of Biomedical Informatics* 44, 6 (2011), 989–996.

J. Bian, U. Topaloglu, and F. Yu. 2012. Towards large-scale twitter mining for drug-related adverse events. In *2012 International Workshop on Smart Health and Wellbeing*. 25–32.

C. Biriell and I. R. Edwards. 1997. Reasons for reporting adverse drug reactions – some thoughts based on an international review. *Pharmacoepidemiology and Drug Safety* 6, 1 (1997), 21–26.

C. Bizer, T. Heath, and T. Berners-Lee. 2009. Linked data – the story so far. *International Journal on Semantic Web and Information Systems* 5, 3 (2009), 1–22.

O. Bodenreider. 2004. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research* 32, Supplement 1 (2004), D267–D270.

O. Bodenreider. 2009. Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. In *AMIA Annual Symposium Proceedings*. 45–49.

C. L. Burgess, C. D. Holman, and A. G. Satti. 2005. Adverse drug reactions in older australians, 1981–2002. *Medical Journal of Australia* 182, 6 (2005), 267–270.

J. Cable. 2009. Adverse events of statins - an informal internet-based study. *Journal of Independent Medical Research* 7, 1 (2009), 24 pages.

S. Chaudhuri, V. Ganti, and R. Kaushik. 2006. A primitive operator for similarity joins in data cleaning. In *22nd International Conference on Data Engineering*. 5–17.

B. W. Chee, R. Berlin, and B. Schatz. 2011. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*. 217–226.

E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman. 2008. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: An initial study. *Journal of the American Medical Informatics Association* 15, 1 (2008), 87–98.

J. Chen, H. He, J. Li, H. Jin, D. McAullay, G. Williams, R. Sparks, and C. Kelman. 2005. Representing association classification rules mined from health data. In *9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*. 1225–1231.

J. Chen, H. Jin, H. He, D. Mcaullay, C. M. O'Keefe, R. Sparks, and C. Kelman. 2010. Mining consequence events in temporal health data. *Intelligent Data Analysis - Knowledge Discovery in Bioinformatics* 14, 2 (2010), 245–261.

P. Christen. 2008. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 151–159.

P. Christen. 2012. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering* 24, 9 (2012), 1537–1555.

CIOMS. 2010. Practical Aspects of Signal Detection in Pharmacovigilance. (2010).

D. C. Classen, S. L. Pestotnik, R. S. Evans, and J. P. Burke. 1991. Computerized surveillance of adverse drug events in hospital patients. *Journal of the American Medical Association* 266, 20 (1991), 2847–2851.

D. C. Classen, S. L. Pestotnik, R. S. Evans, and J. P. Burke. 1992. Description of a computerized adverse drug event monitor using a hospital information system. *Hospital Pharmacy* 27, 9 (1992), 776–779.

D. C. Classen, S. L. Pestotnik, R. S. Evans, J. F. Lloyd, and J. P. Burke. 1997. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. *JAMA* 277, 4 (1997), 301–306.

N. Colineau and C. Paris. 2010. Talking about your health to strangers: Understanding the use of online social networks by patients. *New Review of Hypermedia and Multimedia* 16, 1–2 (2010), 141–160.

P. M. Coloma, M. J. Schuemie, G. Trifiró, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, C. Giaquinto, G. Corrao, L. Pedersen, J. van der Lei, and M. Sturkenboom. 2011. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiology Drug Safety* 20, 1 (2011), 1–11.

P. M. Coloma, G. Trifiró, V. Patadia, and M. Sturkenboom. 2013. Postmarketing safety surveillance. *Drug Safety* 36, 3 (2013), 183–197.

T. M. Cover and P. E. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27.

L. H. Curtis, M. G. Weiner, D. M. Boudreau, W. O. Cooper, G. W. Daniel, V. P. Nair, M. A. Raebel, N. U. Beaulieu, R. Rosofsky, and T. S. Woodworth. 2012. Design considerations, architecture, and use of the mini-sentinel distributed data system. *Pharmacoepidemiology and Drug Safety* 21, S1 (2012), 23–31.

C. M. Cusack. 2008. Electronic health records and electronic prescribing: Promise and pitfalls. *Obstetrics and Gynecology Clinics of North America* 35, 1 (2008), 63–79.

S. Doan, N. Collier, H. Xu, P. H. Duy, and T. M. Phuong. 2012. Recognition of medication information from discharge summaries using ensembles of classifiers. *BMC Medical Informatics and Decision Making* 12, 36 (2012), 10 pages.

W. DuMouchel. 1999. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *American Statistician* 53, 3 (1999), 177–190.

W. DuMouchel and D. Pregibon. 2001. Empirical Bayes screening for multi-item associations. In *7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 67–76.

I. R. Edwards and J. K. Aronson. 2000. Adverse drug reactions: Definitions, diagnosis, and management. *The Lancet* 356, 9237 (2000), 1255–1259.

J. P. Ehsani, T. Jackson, and S. J. Duckett. 2006. The incidence and cost of adverse events in victorian hospitals 2003-04. *Medical Journal of Australia* 184, 11 (2006), 551–555.

M. G. Elfeky, V. S. Verykios, and A. K. Elmagarmid. 2002. TAILOR: A record linkage toolbox. In *18th International Conference on Data Engineering*. 17–28.

EMA. 2012. *The European Medicines Agency Guidelines on Good Pharmacovigilance Practices (GVP)*. Technical Report EMA/827661/2011. The European Medicines Agency.

S. J. W. Evans, P. C. Waller, and S. Davis. 2001. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety* 10, 6 (2001), 483–486.

FDA. 2014. FAERS Reporting by Healthcare Providers and Consumers by Year. Retrieved from http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070456.htm.

T. R. Fleming and D. P. Harrington. 1992. Counting processes and survival analysis. *Statistics in Medicine* 11, 13 (1992), 1792–1793.

S. Fox and S. Jones. 2009. *The Social Life Of Health Information.* Technical Report. Pew Internet & American Life Project (2009).

D. M. Fram, J. S. Almenoff, and W. DuMouchel. 2003. Empirical Bayesian data mining for discovering patterns in post-marketing drug safety. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 359–368.

C. Friedman. 2009. Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In *12th Conference on Artificial Intelligence in Medicine: Artificial Intelligence in Medicine*. 1–5.

C. D. Furberg, A. A. Levin, P. A. Gross, R. S. Shapiro, and B. L. Strom. 2006. The FDA and drug safety: A proposal for sweeping changes. *JAMA Internal Medicine* 166, 18 (2006), 1938–1942.

H. Gonzalez, A. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, and W. Shen. 2010. Google fusion tables: Data management, integration and collaboration in the cloud. In *1st ACM Symposium on Cloud Computing*. 175–180.

M. Gordon. 2008. Improving post-approval risk surveillance for drugs: Active post-market risk identification. *Michigan Telecommunications and Technology Law Review* 15, 297 (2008), 297–314.

H. Gurulingappa, J. Fluck, M. Hofmann-Apitius, and L. Toldo. 2011. Identification of adverse drug event assertive sentences in medical case reports. In *1st International Workshop on Knowledge Discovery and Health Care Management*. 16–27.

H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics* 45, 5 (2012), 885–892.

A. Y. Halevy, A. Rajaraman, and J. J. Ordille. 2006. Data integration: The teenage years. In *32nd International Conference on Very Large Data Bases*. Seoul, Korea, 9–16.

R. W. Hamming. 1950. Error detecting and error correcting codes. *Bell System Technical Journal* 29, 2 (1950), 147–160.

T. J. Hannan. 1999. Detecting adverse drug reactions to improve patient outcomes. *International Journal of Medical Informatics* 55, 1 (1999), 61–64.

R. Harpaz, S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase, and C. Friedman. 2013. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association* 20, 3 (2013), 413–419.

C. Hawn. 2009. Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care. *Health Affairs* 28, 2 (2009), 361–368.

L. Hazell and S. A. Shakir. 2006. Under-reporting of adverse drug reactions: A systematic review. *Drug Safety* 29, 5 (2006), 385–396.

H. He, G. J. Williams, J. Chen, S. Hawkins, and C. Kelman. 2004. Exploring possible adverse drug reactions by clustering event sequences. In *Data Warehousing and Knowledge Discovery*. 199–208.

A. B. Hill. 1965. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine* 58, 5 (1965), 295–300.

B. Honigman, J. Lee, J. Rothschild, P. Light, R. M. Pulling, T. Yu, and D. W. Bates. 2001. Using computerized data to identify adverse drug events in outpatients. *Journal of the American Medical Informatics Association* 8, 3 (2001), 254–266.

B. L. Hug, C. Keohane, D. L. Seger, C. Yoon, and D. W. Bates. 2012. The costs of adverse drug events in community hospitals. *Joint Commission Journal on Quality and Patient Safety* 38, 3 (2012), 120–126.

R. K. Hulse, S. J. Clark, J. C. Jackson, H. R Warner, and R. M. Gardner. 1976. Computerized medication monitoring system. *American Journal of Hospital Pharmacy* 33, 10 (1976), 1061–1064.

S. H. Hwang, S. Lee, H. K. Koo, and Y. Kim. 2008. Evaluation of a computer-based adverse-drug-event monitor. *American Journal of Health-System Pharmacy* 65, 23 (2008), 2265–2272.

ICH Working Group. 1996. ICH harmonised tripartite guideline: Guideline for good clinical practice E6 (R1). In *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use*, Vol. 10.

International Organization for Standardization (ISO). 2005. *Health Informatics – Electronic Health Record – Definition, Scope and Context*. Technical Report ISO/TR 20514.

A. K. Jha, G. J. Kuperman, J. M. Teich, L. Leape, B. Shea, E. Rittenberg, E. Burdick, D. L. Seger, M. Vander Vliet, and D. W. Bates. 1998. Identifying adverse drug events: Development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *Journal of the American Medical Informatics Association* 5, 3 (1998), 305–314.

Y. Ji, R. M. Massanari, J. Ager, J. Yen, R. E. Miller, and H. Ying. 2007. A fuzzy logic-based computational recognition-primed decision model. *Information Sciences* 177, 20 (2007), 4338–4353.

Y. Ji, H. Ying, P. Dews, A. Mansour, J. Tran, R. E. Miller, and R. M. Massanari. 2011. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Transactions on Information Technology in Biomedicine* 15, 3 (2011), 428–437.

Y. Ji, H. Ying, J. Tran, P. Dews, A. Mansour, and R. M. Massanari. 2013. A method for mining infrequent causal associations and its application in finding adverse drug reaction signal pairs. *IEEE Transactions on Knowledge and Data Engineering* 25, 4 (2013), 721–733.

H. Jin, J. Chen, H. He, C. Kelman, D. McAullay, and C. M. O'Keefe. 2010. Signaling potential adverse drug reactions from administrative health databases. *IEEE Transactions on Knowledge and Data Engineering* 22, 6 (2010), 839–853.

H. Jin, J. Chen, H. He, G. J. Williams, C. Kelman, and C. M. O'Keefe. 2008. Mining unexpected temporal associations: Applications in detecting adverse drug reactions. *IEEE Transactions on Information Technology in Biomedicine* 12, 4 (2008), 488–500.

H. Jin, J. Chen, C. Kelman, H. He, D. McAullay, and C. M. O'Keefe. 2006. Mining unexpected associations for signalling potential adverse drug reactions from administrative health databases. In *10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. 867–876.

S. Karimi, S. N. Kim, and L. Cavedon. 2011. Drug side-effects: What do patient forums reveal? In *2nd International Workshop On Web Science and Information Exchange in the Medical Web*. 14–15.

M. Kay. 1986. Algorithm schemata and data structures in syntactic processing. In *Readings in Natural Language Processing*. Morgan Kaufmann Publishers Inc., San Francisco, CA. 35–70.

C. W. Kelman, S. Pearson, R. O. Day, C. D. J. Holman, E. V. Kliewer, and D. A. Henry. 2007. Evaluating medicines: Let's use all the evidence. *Medical Journal of Australia* 186, 5 (2007), 249–252.

D. M. Kirking, J. W. Thomas, F. J. Ascione, and E. L. Boyd. 1986. Detecting and preventing adverse drug interactions: The potential contribution of computers in pharmacies. *Social Science and Medicine* 22, 1 (1986), 1–8.

J. Koch-Weser, E. M. Sellers, and R. Zacest. 1977. The ambiguity of adverse drug reactions. *European Journal of Clinical Pharmacology* 11, 2 (1977), 75–78.

M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork. 2010. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology* 6, 1 (2010), Article 343.

R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. 2010. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the Workshop on Biomedical Natural Language Processing*. Uppsala, Sweden, 117–125.

M. Lenzerini. 2002. Data integration: A theoretical perspective. In *21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. 233–246.

P. LePendu, S. V. Iyer, A. Bauer-Mehren, R. Harpaz, J. M. Mortensen, T. Podchiyska, T. A. Ferris, and N. H. Shah. 2013. Pharmacovigilance using clinical notes. *Clinical Pharmacology and Therapeutics* 93, 6 (2013), 547–555.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, Vol. 10. 707–710.

M. A. Levin, M. Krol, A. M. Doshi, and D. L. Reich. 2007. Extraction and mapping of drug names from free text to a standardized nomenclature. In *AMIA Annual Symposium Proceedings*. 438–442.

A. Y. Levy. 2000. Logic-based techniques in data integration. In *Logic-Based Artificial Intelligence, Kluwer International Series in Engineering and Computer Science*, Vol. 597. Kluwer Academic Publishers, Norwell, MA. 575–595.

J. Li, H. Shen, and R. Topor. 2001. Mining optimal class association rule set. In *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. 364–375.

M. Lindquist. 2008. VigiBase, the WHO Global ICSR database system: Basic facts. *Drug Information Journal* 42, 5 (2008), 409–419.

M. Lindquist, M. Ståhl, A. Bate, I. R. Edwards, and R. Meyboom. 2000. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. *Drug Safety* 23, 6 (2000), 533–542.

J. Liu, A. Li, and S. Seneff. 2011. Automatic drug side effect discovery from online patient-submitted reviews: Focus on statin drugs. In *International Conference on Advances in Information Mining and Management*. 91–96.

M. Liu, E. R. McPeek Hinz, M. E. Matheny, J. C. Denny, J. S. Schildcrout, R. A. Miller, and H. Xu. 2013. Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *Journal of the American Medical Informatics Association* 20, 3 (2013), 420–426.

X. Liu and H. Chen. 2013. AZDrugMiner: An information extraction system for mining patient-reported adverse drug events in online patient forums. In *International Conference on Smart Health*. 134–150.

J. J. Mao, A. Chung, A. Benton, S. Hill, L. Ungar, C. E. Leonard, S. Hennessy, and J. H Holmes. 2013. Online discussion of drug side effects and discontinuation among breast cancer survivors. *Pharmacoepidemiology and Drug Safety* 22, 3 (2013), 256–262.

A. T. McCray, S. Srinivasan, and A. C. Browne. 1994. Lexical methods for managing variation in biomedical terminologies. In *The Annual Symposium on Computer Application in Medical Care*. 235–239.

T. L. McDonald, W. P. Erickson, and L. L. McDonald. 2000. Analysis of count data from before-after control-impact studies. *Journal of Agricultural, Biological, and Environmental Statistics* 5, 3 (2000), 262–279.

C. A. Naranjo, U. Busto, E. M. Sellers, P. Sandor, I. Ruiz, E. A. Roberts, E. Janecek, C. Domecq, and D. J. Greenblatt. 1981. A method for estimating the probability of adverse drug reactions. *Clinical Pharmacology and Therapeutics* 30 (1981), 239–245.

J. Nelson, A. Cook, and O. Yu. 2009. *Evaluation of Signal Detection Methods for Use in Prospective Post-licensure Medical Product Safety Surveillance*. Technical Report. FDA Sentinel Initiative Safety Signal Identification Contract.

J. N. Nkanza and W. Walop. 2004. Vaccine associated adverse event surveillance (VAAES) and quality assurance. *Drug Safety* 27, 12 (2004), 951–952.

G. N. Norén, J. Hopstadius, A. Bate, K. Star, and I. R. Edwards. 2010. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery* 20, 3 (2010), 361–387.

G. N. Norén, R. Orre, and A. Bate. 2005. A hit-miss model for duplicate detection in the WHO drug safety database. In *11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. 459–468.

G. N. Norén, R. Orre, A. Bate, and I. R. Edwards. 2007. Duplicate detection in adverse drug reaction surveillance. *Data Mining and Knowledge Discovery* 14, 3 (2007), 305–328.

S. Olsson, S. Pal, A. Stergachis, and M. Couper. 2010. Pharmacovigilance activities in 55 low- and middle-income countries. *Drug Safety* 33, 8 (2010), 689–703.

S. N. Pal, C. Duncombe, D. Falzon, and S. Olsson. 2013. WHO strategy for collecting safety data in public health programmes: Complementing spontaneous reporting systems. *Drug Safety* 36, 2 (2013), 75–81.

M. Y. Park, D. Yoon, K. Lee, S. Y. Kang, I. Park, S. H. Lee, W. Kim, H. J. Kam, Y. H. Lee, J. H. Kim, and R. W. Park. 2011. A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. *Pharmacoepidemiology and Drug Safety* 20, 6 (2011), 598–607.

V. N. Patel and D. C. Kaelber. 2013. Using aggregated, de-identified electronic health record data for multivariate pharmacosurveillance: A case study of azathioprine. *Journal of Biomedical Informatics* 52 (2013), 36–42.

M. Paul and M. Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *International AAAI Conference on Weblogs and Social Media*. 265–272.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

L. Peters, J. E. Kapusnik-Uner, T. Nguyen, and O. Bodenreider. 2011. An approximate matching method for clinical drug names. In *AMIA Annual Symposium Proceedings*, Vol. 2011. 1117–1126.

L. Philips. 1990. Hanging on the metaphone. *Computer Language* 7, 12 (1990), 39–44.

B. Quilitz and U. Leser. 2008. Querying distributed RDF data sources with SPARQL. In *Proceedings of the 5th European Semantic Web Conference on the Semantic Web: Research and Applications*. Canary Islands, Spain. 524–538.

E. Rahm and P. A. Bernstein. 2001. A survey of approaches to automatic schema matching. *VLDB Journal* 10, 4 (2001), 334–350.

E. Ramirez, A. J. Carcas, A. M. Borobia, S. H. Lei, E. Piñana, S. Fudio, and J. Frias. 2009. A pharmacovigilance program from laboratory signals for the detection and reporting of serious adverse drug reactions in hospitalized patients. *Clinical Pharmacology and Therapeutics* 87, 1 (2009), 74–86.

J. M. Reps, J. M. Garibaldi, U. Aickelin, D. Soria, J. E. Gibson, and R. B. Hubbard. 2012. Comparing data-mining algorithms developed for longitudinal observational databases. In *12th UK Workshop on Computational Intelligence*. 1–8.

J. M. Reps, J. M. Garibaldi, U. Aickelin, D. Soria, J. Gibson, and R. Hubbard. 2013. Comparison of algorithms that detect drug side effects using electronic healthcare databases. *Soft Computing* 17, 12 (2013), 2381–2397.

E. E. Roughead and S. J. Semple. 2009. Medication safety in acute care in Australia: Where are we now? Part 1: A review of the extent and causes of medication problems 2002–2008. *Australia and New Zealand Health Policy* 6, 1 (2009), 18.

T. Sakaeda, A. Tamon, K. Kadoyama, and Y. Okuno. 2013. Data mining of the public version of the FDA adverse event reporting system. *International Journal of Medical Sciences* 10, 7 (2013), 796–803.

M. Samwald, A. Jentzsch, C. Bouton, C. Kallesøe, E. Willighagen, J. Hajagos, M. Marshall, E. Prud'hommeaux, O. Hassanzadeh, and E. Pichler. 2011. Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics* 3, 19 (2011), 6 pages.

S. Sarawagi and A. Bhamidipaty. 2002. Interactive deduplication using active learning. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 269–278.

N. Savage. 2012. Digging for drug facts. *Communications of the ACM* 55, 10 (2012), 11–13.

D. Scanfeld, V. Scanfeld, and E. Larson. 2010. Dissemination of health information through social networks: Twitter and antibiotic. *American Journal of Infection Control* 3, 38 (2010), 182–188.

B. A. Slade, L. Leidel, C. Vellozzi, E. J. Woo, W. Hua, A. Sutherland, H. S. Izurieta, R. Ball, N. Miller, and M. M. Braun. 2009. Postlicensure safety surveillance for quadrivalent human papillomavirus recombinant vaccine. *Journal of the American Medical Association* 302, 7 (2009), 750–757.

P. E. Stang, P. B. Ryan, J. A. Racoosin, J. M. Overhage, A. G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock. 2010. Advancing the science for active surveillance: Rationale and design for the Observational Medical Outcomes Partnership. *Annals of Internal Medicine* 153, 9 (2010), 600–606.

M. Stephens, D. Buckingham, J. C. Talbot, and P. Routledge. 1985. *The Detection of New Adverse Drug Reactions*. Stockton Press.

B. H. Stricker and J. G. Tijssen. 1992. Serum sickness-like reactions to cefaclor. *Journal of Clinical Epidemiology* 45, 10 (1992), 1177–1184.

M. Suling and I. Pigeot. 2012. Signal detection and monitoring based on longitudinal healthcare data. *Pharmaceutics* 4, 4 (2012), 607–640.

The Yellow Card Scheme. 2013. Homepage. Retrieved from https://yellowcard.mhra.gov.uk/.

Petra A. Thrmann. 2001. Methods and systems to detect adverse drug reactions in hospitals. *Drug Safety* 24, 13 (2001), 961–968. DOI:http://dx.doi.org/10.2165/00002018-200124130-00003

A. Tinoco, R. S. Evans, C. J. Staes, J. F. Lloyd, J. M. Rothschild, and P. J. Haug. 2011. Comparison of computerized surveillance and manual chart review for adverse events. *Journal of the American Medical Informatics Association* 18, 4 (2011), 491–497.

Y. Tsuruoka, J. McNaught, and S. Ananiadou. 2008. Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics* 9, 3 (2008), 1–10.

V. E. Valkhoff, P. M. Coloma, G. M. Masclee, R. Gini, F. Innocenti, F. Lapi, M. Molokhia, M. Mosseveld, M. S. Nielsson, M. Schuemie, F. Thiessard, J. van der Lei, M. C. Sturkenboom, G. Trifiró, and EU-ADR Consortium. 2014. Validation study in four health-care databases: Upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. *Journal of Clinical Epidemiology* 67, 8 (2014), 921–931.

W. Wang, K. Haerian, H. Salmasian, R. Harpaz, H. Chase, and C. Friedman. 2011. A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from Pubmed citations. In *AMIA Annual Symposium Proceedings*. Washington, DC, 1464–1470.

X. Wang, G. Hripcsak, and C. Friedman. 2009a. Characterizing environmental and phenotypic associations using information theory and electronic health records. *BMC Bioinformatics* 10, 9 (2009), S13.

X. Wang, G. Hripcsak, M. Markatou, and C. Friedman. 2009b. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *Journal of the American Medical Informatics Association* 16, 3 (2009), 328–337.

R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz. 2013. Web-scale pharmacovigilance: Listening to signals from the crowd. *Journal of the American Medical Informatics Association* 20, 3 (2013), 404–408.

G. Williams, R. Baxter, C. Kelman, C. Rainsford, H. He, L. Gu, D. Vickers, and S. Hawkins. 2002. Estimating episodes of care using linked medical claims data. In *Advances in Artificial Intelligence*. 660–671.

A. M. Wilson, L. Thabane, and A. Holbrook. 2004. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology* 57, 2 (2004), 127–134.

R. L. Woosley. 2013. One hundred years of drug regulation: Where do we go from here? *Annual Review of Pharmacology and Toxicology* 53 (2013), 255–273.

L. Wuerdeman, L. Volk, L. Pizziferri, R. Tsurikova, C. Harris, R. Feygin, M. Epstein, K. Meyers, J. S. Wald, and D. Lansky. 2005. How accurate is information that patients contribute to their electronic health record? In *AMIA Annual Symposium Proceedings*, Vol. 2005. 834.

H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny. 2010. MedEx: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association* 17, 1 (2010), 19–24.

C. C. Yang, L. Jiang, H. Yang, and X. Tang. 2012a. Detecting signals of adverse drug reactions from health consumer contributed content in social media. In *ACM SIGKDD Workshop on Health Informatics*.

C. Yang, P. Srinivasan, and P. M. Polgreen. 2012b. Automatic adverse drug events detection using letters to the editor. In *AMIA Annual Symposium Proceedings*. 1030–1039.

A. Yates and N. Goharian. 2013. ADRTrace: Detecting expected and unexpected adverse drug reactions from user reviews on social media sites. *Advances in Information Retrieval* 7814 (2013), 816–819.

E. Yom-Tov and E. Gabrilovich. 2013. Postmarket drug surveillance without trial costs: Discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of Medical Internet Research* 15, 6 (2013), e124.

Q. Zeng, S. Kogan, N. Ash, R. A. Greenes, and A. A. Boxwala. 2002. Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine* 41, 4 (2002), 289–298.

Received March 2014; revised November 2014; accepted January 2015