Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis

R Harpaz[1], W DuMouchel[2,3], NH Shah[4], D Madigan[3,5], P Ryan[3,6] and C Friedman[1]

An important goal of the health system is to identify new adverse drug events (ADEs) in the postapproval period. Data-mining methods that can transform data into meaningful knowledge to inform patient safety have proven essential for this purpose. New opportunities have emerged to harness data sources that have not been used within the traditional framework. This article provides an overview of recent methodological innovations and data sources used to support ADE discovery and analysis.

BACKGROUND

Pharmacovigilance (PhV), also referred to as drug safety surveillance, has been defined as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug problem".[1] PhV starts at the pre-approval stage, when information about adverse drug events (ADEs) is collected during phases I–III of clinical trials (without necessarily establishing any causal relationship between the ADEs and either the investigational product or concomitant therapies) and continues in the postapproval stage and throughout a drug's life on the market. During the initial postapproval stage, PhV may continue through phase IV clinical trials, often mandated by regulatory agencies to obtain additional safety data on a product during routine use. Although clinical trials are used for evaluating safety issues, they are limited with respect to the number and characteristics of patients exposed, duration, and type of data collected. As a result, the complete safety profile associated with a new drug cannot be fully established through clinical trials.

Postapproval ADEs are a major global health concern, accounting for more than 2 million injuries, hospitalizations, and deaths each year in the United States alone,[2,3] and associated costs are estimated at $75 billion annually.[4] Hence, timely and accurate detection of ADEs in the postapproval period is an urgent goal of the public health system. Computational methods at the intersection of statistics, computer science, medicine, epidemiology, chemoinformatics, and biology that can translate data into meaningful knowledge to benefit patient safety have proven to be a critical component in PhV. These methods have commonly been referred to as data-mining algorithms (DMAs).

Historically, PhV has relied on a clinical review process of case reports collected at designated organizations. In response to the challenge posed by the vast quantities and complexity of data that needed to be examined, DMAs were originally designed to aid in this process, enabling evaluators to peruse large volumes of data and focus their attention on issues that might be more important to public health. Since then, however, the role, quality, and capabilities of DMAs have dramatically expanded in order to address new challenges, leverage new information sources, and bring about overall improvement in drug safety surveillance. In this article, the term DMAs refers to automated high-throughput methods to uncover hidden relationships of potential clinical significance to drug safety.

DMAs can be classified along several axes depending on the data source to which they are applied and the scientific function they are designed to perform. The main PhV data sources in current use are spontaneous reporting systems. The focus of research is currently shifting toward the use of large healthcare databases such as electronic health records and administrative claims. Other sources that have recently been investigated include the biomedical literature, chemical and biological information sources, and patient-generated data in health-related Internet forums. The main class of DMAs represents methods designed to generate measures of statistical association for large sets of drug–outcome pairs, which can be used to prioritize and identify risk signals warranting further attention.[5,6] Newer approaches have been designed to facilitate identification of higher-order or multivariate associations that represent more complex safety phenomena such as drug–drug interactions, syndromic events, and class effects. A large class of methods

[1]Department of Biomedical Informatics, Columbia University Medical Center, New York, New York, USA; [2]Oracle Health Sciences, Burlington, Massachusetts, USA; [3]Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, Maryland, USA; [4]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA; [5]Department of Statistics, Columbia University, New York, New York, USA; [6]Janssen Research and Development, Titusville, New Jersey, USA. Correspondence: R Harpaz (rave.harpaz@dbmi.columbia.edu)

Received 23 January 2012; accepted 7 March 2012; advance online publication 2 May 2012. doi:10.1038/clpt.2012.50

has been designed to address the issues of confounding. Other approaches have been designed to abstract the data in meaningful ways to uncover interesting patterns (such as clusters or networks of ADEs that may convey clinically important information), whereas a newer wave of methods is being designed to leverage non-traditional data sources and to link information from multiple data sources.

In recognition of the importance of DMAs, research into the application of DMAs to PhV has steadily grown in the past decade (**Figure 1**, DMA-related publication trends). This article provides an overview of recent innovations in DMA methodologies and the wide range of data sources used in support of PhV. Extending the theme of several related reviews,[6–8] we aim to cover a broader range of methods and data sources. Our overview is restricted to works published in the past five years. We do not seek to exhaustively list or critically examine all relevant work. Rather, we present a synopsis of basic concepts, contributions, and major findings. We begin our discussion with a description of data and information sources considered in this article, highlighting their strengths and limitations. The discussion regarding methodologies is organized according to the data source axis: spontaneous reports, health-care data, and other data sources. We also provide a brief overview of traditional approaches (disproportionality analysis (DPA)) as a comparator for discussion of other approaches.

DATA AND INFORMATION SOURCES USED IN SUPPORT OF PhV

Drug safety surveillance has relied predominantly on spontaneous reporting systems (SRSs). These are passive systems composed of reports of suspected ADEs collected from health-care professionals, consumers, and pharmaceutical companies, and maintained largely by regulatory and health agencies. Among the prominent SRSs are the US Food and Drug Administration (FDA) Adverse Event Reporting System (AERS) and the VigiBase maintained by the World Health Organization (WHO). Although the various SRSs may differ in structure and content, most are based on voluntary reporting (except for pharmaceutical companies, which are required to report suspected ADEs to regulators once they come to their attention) and typically capture information on the drug suspected to cause the ADE, as well as on concomitant drugs, indications, suspected events, and limited demographic information in a structured format directly amenable to data mining. The FDA uses a data-mining engine to compute signal scores (statistical reporting associations) for the millions of drug–event combinations in the AERS, which offers a "hypothesis-free" view of the safety characteristics in the underlying data. It should be stressed, however, that these signals by themselves do not establish a causal relationship between a drug and an adverse event; rather, they are considered initial warnings that require further assessment using other sources of support. Typically, after this initial signal-generation step, an intertwined process of signal strengthening and signal confirmation follows, with drug safety evaluators looking for signs such as a temporal relationship, coherence with published case reports, biological and clinical plausibility, similarity to other drugs, and supporting data from clinical trials, or by conducting epidemiologic studies in several large health-care databases to establish causality.[9,10]

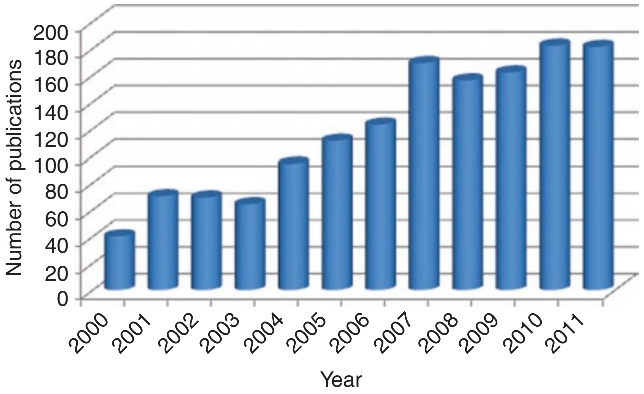

Figure 1 Evolution of PhV DMA research described by volume of publications per year indexed in PubMed. The volume for 2011 is effectively larger because of delayed indexing. PhV DMA, pharmacovigilance data-mining algorithm.

SRSs are prefocused on drug–adverse event relationships; the collection and processing is centralized; they communicate genuine health concerns, cover large populations, and are accessible for analysis; and since their inception they have supported regulatory decisions for a long list of marketed drugs.[11] Notwithstanding these advantages, SRSs suffer from a range of limitations including (i) overreporting (adverse events known to be linked to certain drugs are more likely to be reported than other adverse events), (ii) misattribution of causality in drug–event combinations, (iii) missing or incomplete data, (iv) duplication of reporting, and (v) unspecified causal links.[5,12]

Recent events pertaining to drug safety, such as the case of rofecoxib (Vioxx), a widely used anti-inflammatory drug estimated to have caused 88,000 episodes of myocardial infarction,[13] have highlighted the need to identify new data sources and improved analytic methods to create a more effective PhV system.[9,14–16] The US Congress recently gave the FDA a mandate to establish an active surveillance system.[17] Subsequently, several large-scale research initiatives, including the Sentinel Initiative[9,18] and the Observational Medical Outcomes Partnership (OMOP),[15,19] were established in the United States. A similar system, known as the EU-ADR (European Union–Adverse Drug Reaction) project, was initiated in Europe by the European Commission.[20] These new developments rely on the expanded secondary use of electronic health-care data (HCD) such as electronic health records and administrative claims that typically contain time-stamped interventions, procedures, diagnoses, medications, medical narratives, and billing codes. Unlike spontaneous reports, electronic HCD are representative of routine clinical care recorded over long periods of time. As such, they contain a more complete record of a patient's medical history, treatments, conditions, and potential risk factors. They are also not restricted to patients who experience ADEs. Consequently, electronic HCD offer several advantages that may be used to complement SRSs, especially confirmatory analysis and the potential for active surveillance.

Several retrospective studies have demonstrated that the Vioxx case could have been signaled earlier if this type of data had been used.[21–24] However, the secondary use of health-care data presents other challenges. The data often require complex preprocessing to support analysis. The data are not oriented to capture adverse events, which are typically not identified *per se* but as diagnoses (usually based on billing codes). There are logistical issues in storing, accessing, and sharing data across health-care providers, and these are compounded by legal and privacy issues concerning access to patient data.[9,15] Styles for data capture and documentation vary, as do standards for data encoding.[9,15] There is also a need for automated methods that can extract relevant information from free-text clinical narratives[25] and for methods that can address the pervasiveness of confounding inherent in observational studies.[10,26,27]

In the recent past, researchers have begun to focus on data and information sources that have not traditionally been used for PhV. Each of these sources offers unique prospects that may be leveraged to complement or augment existing approaches. Here we discuss several of these data sources.

The public availability of chemical and biological knowledge bases such as DrugBank,[28] which contains information on both chemical structure and drug targets, is opening new opportunities to bridge the gap between the molecular and clinical domains and further the study of ADEs.[29,30] By leveraging this type of knowledge, e.g., protein binding sites, biological pathways of drug action and metabolism, linkages between chemical substructures and specific toxicities, and chemical structural similarities between drugs, the molecular determinants of ADEs can be better understood. Moreover, predictive models can be created, thereby allowing a more proactive approach to PhV. A central premise in this domain is that ADEs are largely predictable consequences of certain molecular actors.[30] Several aspects of this premise have been validated by long experience and exploited for high-throughput screening of active compounds in computer-aided drug design and development. It has also been used by pharmaceutical companies in the preclinical drug design stage to predict toxicological effects, the main goal being to decrease late-stage attrition of new drugs due to toxicity.[31] In contrast to the preclinical stage, recent successful studies have linked this knowledge source with knowledge on post-marketed ADEs to create better predictive models and as a tool to augment existing ADE discovery methodologies.[30,32,33]

Mining the biomedical literature holds the promise of extracting new discoveries from the large amounts of biomedical knowledge available. This approach has been successfully used to discover new relationships among genes, biological pathways, and diseases, as well as for drug repurposing (discovering new indications for existing drugs).[34] The biomedical literature contains ADE-related information based on clinical studies and anecdotal observations. Current use of biomedical literature by drug safety researchers (to evaluate or confirm new ADEs) suggests that automated or data-mining approaches can supplement existing ADE discovery techniques. However, extracting information from the biomedical literature is nontrivial and requires elaborate natural-language processing (NLP) tools. Recent work has demonstrated its potential as a strategy for prioritizing ADE associations under consideration.[35]

Social networks and forums for patients on the Internet, such as Ask a Patient, DailyStrength, Yahoo Health and Wellness, and PatientsLikeMe, collect patient self-reports of drug side effects and provide a platform for patients to discuss and share their experiences with medications. Although the information provided by patients may be inaccurate or even questionable, such forums can provide valuable supplementary information on drug effectiveness and side effects because they cover large and diverse populations and contain unsolicited, uncensored data directly from patients. However, extracting such information is very challenging and requires deep statistical and linguistic methods to interpret colloquial language, correct grammatical and spelling errors, and distinguish real experiences from hearsay. Nonetheless, recent work has shown that the information contained in these forums is extractable and relevant to PhV.[36]

METHODS APPLIED TO SPONTANEOUS REPORTING SYSTEMS

DPA and basic concepts

DPA is the main driving force behind most computerized PhV methods involving SRSs. DPA methodologies use frequency analysis of 2 × 2 contingency tables to estimate surrogate measures of statistical association between specific drug–event combinations mentioned in spontaneous reports. The term "disproportionality analysis" conveys the purpose of the analysis, namely, to quantify the degree to which a drug–event combination occurs "disproportionally" as compared with what would be expected if there were no statistical association between the drug and the event.[5]

DPA methodologies differ with respect to the exact measures that are used and the statistical adjustments made to account for low counts; they can generally be classified into two categories: frequentist and Bayesian. Both approaches use the entries of Table 1 (or stratified versions thereof) to derive a statistical association/disproportionality measure. This table is usually computed for each drug–event pair in the SRS. The most widely discussed measure is the relative reporting ratio (RRR),[6] defined as the ratio of the observed incidence rate of a drug–event combination to its "baseline" expected rate under the assumption that the drug and event occur independently. Both the FDA and the WHO use a Bayesian version of RRR as a basis for monitoring safety signals in their SRSs.[37,38] Other widely used measures and their mathematical definitions are displayed in Table 2.

Table 1 Contingency table used in SRS-based DPA

	With target AE	Without target AE	Total
With target drug	a	b	$n = a + b$
Without target drug	c	d	c + d
Total	m = a + c	b + d	t = a + b + c + d

AE, adverse event; DPA, disproportionality analysis; SRS, spontaneous reporting system.

Reports are classified according to the presence/absence of specific drug–AE combinations. Each cell contains report counts.

A true value of close to 1 for any of these measures supports the hypothesis that there is no association between the drug and event. An RRR value of 3, for example, would indicate that there are three times as many drug–event reports in the database than would be expected and might support the hypothesis of a drug–ADE association.

Frequentist approaches use one of the measures listed in Table 2 to estimate associations and are typically accompanied by hypothesis tests of independence (χ^2-test or Fisher's exact test). The hypothesis tests are used as an extra precautionary measure to take into account the sample size used while computing an association. Bayesian approaches attempt to account for the uncertainty in the disproportionality measure associated with small observed and expected counts by "shrinking" the measure toward the baseline case of no association by an amount that is proportional to the variability of the disproportionality statistic. The result of this shrinkage is a reduction in spurious associations that have insufficient data to support them.

Table 2 Mathematical definitions of measures of association

Measure of association	Mathematical definition
Relative reporting ratio (RRR)	(t·a)/(m·n)
Proportional reporting ratio	(a·(t – n))/(c·n)
Reporting odds ratio	(a·d)/(c·b)
Information component	log2 (RRR)

Among the Bayesian approaches is the multi-item gamma Poisson shrinker (MGPS).[8,39] MGPS is the predominant DMA used in the United States and the UK, and it is currently used by the FDA[40] as well as several pharmaceutical companies to detect ADE signals in their databases. MGPS, which is based on a modeling framework called empirical Bayes, computes a measure known as the empirical Bayes geometric mean, a centrality measure of the posterior distribution of the true RRR in the population. Typically, the EB05 measure, which corresponds to the lower 5th percentile of the posterior RRR distribution, is used instead for extra conservatism. The WHO uses a Bayesian approach similar to MGPS, called the Bayesian Confidence Propagation Neural Network,[38] which estimates a Bayesian version of the information component.

Typically, *ad hoc* thresholds are applied to the association measures (regardless of the approach or measure) so as to highlight strong associations worthy of further investigation. The thresholds selected usually do not have theoretical or empirical justification. Rather, they are a preliminary means of filtering or sorting. Deshpande *et al.* provide a review of published threshold criteria for qualifying signals of disproportionate reporting in SRSs.[41] A graphical illustration of DMA output is shown in Figure 2.

As yet, there is no consensus on which DPA approach is best, and no gold standard has been established to evaluate the performances of the various approaches. It is widely accepted that none of the approaches is universally better than any other.[5,6]

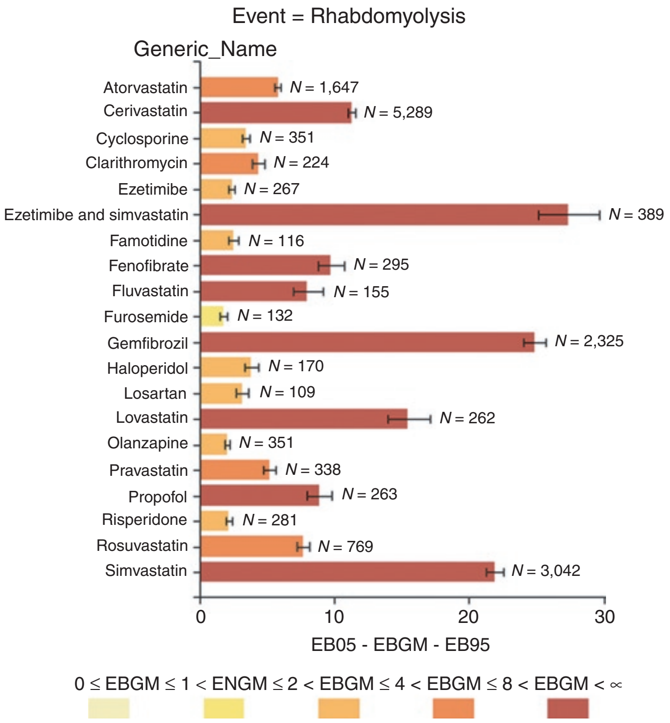

Figure 2 Bar plot of drugs statistically associated with rhabdomyolysis in the AERS, as an example of DPA output. Bar colors and lengths reflect statistical association strengths based on the EBGM scores. Each bar also includes the 90% confidence interval (EB05-EB95) and report count (*N*) for the corresponding drug. The plot and underlying DPA were performed using Oracle's Empirica Signal 7.3 based on AERS data up to and including the year 2011 Q2. Only the top 20 associations consisting of drugs reported as "suspected" with *N* ≥ 100 have been selected for display. As expected, the majority of the drugs belong to the class of statins known to cause rhabdomyolysis as a rare ADE. ADE, adverse drug event; AERS, Adverse Event Reporting System; DPA, disproportionality analysis; EBGM, empirical Bayes geometric mean.

As the number of reports of a specific drug–event combination increases, the different methods tend to give similar results. Some have argued that, for small counts, frequentist approaches are more prone to extreme values and therefore generate more false positives. Others have argued that the Bayesian approaches are too conservative, thereby delaying the detection of novel ADEs. Frequentist approaches are computationally more efficient than Bayesian approaches, but the latter offer the convenience of being able to sort associations along a single dimension because they incorporate information about both disproportionality and sample size. That said, none of the approaches can effectively address reporting biases or confounding in SRSs.

Multivariate methods

Although cumulative experience with DPA has shown it to be a promising adjunct in safety analysis, the reduction of ADE analysis to two dimensions may result in the loss of clinically crucial information. Two-dimensional DPA approaches do not support the discovery and/or analysis of more complex or higher-dimensional drug safety phenomena that involve more than just one drug and one event. The importance and difficulty associated with the detection of these more complex drug safety phenomena have been noted in several prominent PhV reports,[5,7,8] suggesting that more elaborate methods, henceforth collectively referred to as "multivariate" methods, are required.

More complex drug safety patterns may correspond to adverse events associated with drug–drug interactions, such as the pharmacodynamic drug interaction between tramadol and fluoxetine; in this interaction, tramadol (a pain reliever) can enhance the effect of fluoxetine (Prozac), thereby increasing serotonin levels and potentially leading to seizures. Recent studies have shown that many ADEs (close to 50% in hospital patients)[42] are due to drug interactions. This finding suggests that the plausible cause with respect to many of the ADEs reported to SRSs is drug interaction rather than the action of the single suspected drug that was reported. Other, more complex drug safety patterns of clinical interest are class effects and syndromic events (drug-induced syndromes). For example, the class of statins (cholesterol-lowering drugs) is known to cause rhabdomyolysis. The drug varenicline (indicated for smoking cessation) may cause a syndrome of sleeping disorders and other neuropsychiatric disorders. In addition, these patterns are important in highlighting the etiology of other ADEs, leading to the further probing of simpler associations and contributing to an overall better understanding of drug safety and risk factors.[7]

Another limitation of the two-dimensional DPA approaches is that they are not properly equipped to deal with confounding, which is a key factor in association analysis. A confounder is an extraneous variable, either observed or unobserved, that mediates an association between two other variables. If not properly accounted for, confounding may lead to the discovery of spurious associations and therefore erroneous study conclusions. Confounding can be addressed either through experimental design before data collection (e.g., selection of appropriate controls) or in the analysis stage when the data have already been collected (as in the case of SRSs). Simpler types of confounding, such as confounding by age, gender, and year, have been effectively dealt with within DPA approaches through stratification and Mantel–Haenszel type adjustments.[6,7] Although there are many types of confounding, such as confounding by indication, in which the reported event is associated with the indication for treatment, most SRS-related publications have focused on confounding by drug coadministration, where a drug is associated with an event just because it is frequently coprescribed or reported with another drug that is the real cause of the adverse event.

In recent years, several multivariate approaches have been proposed to address these issues while analyzing SRSs. They can generally be classified as (i) DPA extensions, (ii) multivariate logistic regression–based approaches, and (iii) unsupervised machine-learning approaches such as associations rule mining, clustering, and network analysis. DPA extensions to larger dimensions have been applied mostly to three-dimensional associations corresponding to drug–drug interactions,[43] for which observed-to-expected ratios are calculated in a similar manner but based on three elements (drug1–drug2–event). Logistic regression–based approaches have been applied mostly to eliminate confounding by comedication (given the lack of other confounding information in SRSs), whereas unsupervised machine-learning approaches have been used for the identification of more complex or higher-dimensional drug safety phenomena as well as for data abstraction and pattern discovery.

Logistic regression–based approaches. The traditional approach to handle confounding during the analysis stage is stratification, but this is not effective in situations in which a large number of potential confounders need to be examined.[44] A more appropriate approach to handling confounding is by the use of multiple logistic regression, which allows the estimation of a drug–event association by controlling or adjusting for the presence of other covariates (potential confounders).[44] Confounding by comedication can theoretically be addressed by using all drugs in an SRS as regression predictors for an event. Until recently, performing regression analyses of a specific event against all the thousands (>10,000) of drugs included in an SRS represented a significant computational as well a theoretical barrier. Currently, however, new extensions of logistic regression to very-large-dimensional data, known as regularized or Bayesian logistic regression (BLR), can carry out regression analyses with millions of covariates.[45] Caster *et al.*[46] describe an application of BLR to the WHO SRS, involving an attempt to address confounding caused by comedication and a "masking" effect, the latter being relevant in cases in which an increase in background reporting of a specific event (e.g., due to media influences) can disproportionately attenuate measures of true associations toward lower values of no association, thereby masking the true association. The authors describe several real examples of false-positive associations due to confounding by comedication that were corrected using their method and true ADEs that were masked by media focus on the withdrawal of a drug causing rhabdomyolysis. In earlier work, Solomon and DuMouchel[47] applied standard DPA along with BLR to the AERS as part of a study to estimate associations between several contrast media agents and events related to contrast-induced nephropathy. All the methods were adjusted for demographic variables. BLR was also adjusted for 200 drugs coreported with contrast media as potential confounders and as proxies for unobserved confounders. The authors found that the results were consistent among the different methods (including the rank order of associations) but that BLR odds ratio estimates were generally 50% larger. The authors explain this difference as stemming from the differences in the comparators used in the various methods and a masking effect related to the agents investigated.

Unsupervised machine-learning approaches. Multi-item ADE associations are ones that relate multiple drugs to possibly multiple adverse events. Association rule mining (ARM)[48] is a well-established data-mining method for discovering interesting relationships among variables in large databases. ARM can be applied to discover multi-item ADE associations—a special case of association rules. For example,

Chantix, darvocet → memory impairment, abnormal dreams, fatigue, insomnia
(Chantix may interact with darvocet, a pain reliever, and cause various sleeping or mental disorders.)

The computation of association rules is inherently a very hard combinatorial problem that can easily become intractable. The Apriori algorithm[48] can alleviate the problem but does not completely resolve the computational challenge. Rouane *et al.*[49] applied ARM to the SRS of the French Medicines Agency to identify rules related to anti-HIV drugs. The authors proposed the use of formal concept analysis as a means of reducing the computational complexity, but their approach was restricted to only three-item associations. In a recent study, Harpaz *et al.*[50] applied ARM to the AERS with rules of up to six items. Noting the inappropriateness of standard ARM scores for ADE applications, the authors used the RRR score instead, with the additional constraint that each rule must have an RRR larger than any of its subsets. The latter constraint was used to exclude rules that can better be explained by smaller sets of drugs or events.[43] The authors showed that ~66% of the rules corresponded to known associations, thereby demonstrating the potential value of SRSs for discovering multi-item, clinically relevant ADE associations. A promising Bayesian approach to ARM was recently proposed by McCormick *et al.*[51] It has direct application to ADE discovery and can address the sparseness of SRS data when computing association rules.

Currently, the main bottleneck in efforts to widen the application of ARM to SRSs is its requirement for intensive computation. It is likely that ARM will be more widely adopted as computing power increases. An alternative, described in a recent pilot study by Fan *et al.*,[52] is the potential applicability of the highly parallelized and distributed computing paradigm—MapReduce—to the problem.

Clustering is routinely used in many biomedical areas, but until recently its potential was not investigated in the context of ADE analysis. Harpaz *et al.*[53] proposed a nonstandard clustering approach suitable for dealing with the high-dimensional nature and sparseness of SRS data. The method, called biclustering, was applied to the AERS; it defines an ADE cluster as a group of drugs that are all statistically associated with the same group of adverse events. The authors demonstrated that biclustering can be used as an exploratory tool in PhV, and that the underlying large and complex structures of SRSs can be summarized and described in a macroscopic manner (e.g., 40% of ADEs in the AERS are cancer related). They demonstrated that biclustering can be used to highlight class effects (e.g., bisphosphonates) and syndromic events (e.g., sleeping disorders), and showed how it could be used to support the discovery of potentially new ADEs. They found that a large proportion (41%) of the clustered relationships contained associations that are currently unrecognized, signaling potentially new ADEs by allowing these unrecognized associations to borrow support from confirmed drug–event associations within the same cluster. Examples include the following associations: chlorpromazine–hepatotoxicity, methotrexate–pancytopenia, and bosentan–hepatic steatosis, all of which are supported by published case reports.

Ball *et al.*[54] proposed the use of network analysis to facilitate the identification of clinically interesting multidimensional patterns of adverse events. The authors applied network analysis to the FDA's Vaccine Adverse Event Reporting System, in which nodes in the network correspond to vaccines and events. They focused on identifying "hubs," tightly clustered elements within the network that reveal strong informative structures. The authors found patterns linking the vaccine HPV4 with syncope and syncope with seizures in adolescents. They also found patterns of serious gastrointestinal adverse events with the rotavirus vaccine. Last, they demonstrated that the Vaccine Adverse Event Reporting System has the characteristics of a "scale free" (nonrandom) network, with certain vaccines and events acting as hubs.

METHODS APPLIED TO ELECTRONIC HCD

Methods applied to electronic HCD can generally be classified as those based on modified DPA ported from spontaneous reporting and those based on epidemiologic study designs such as the cohort, case–control, and self-controlled study designs. One of the major challenges in the use of HCD is the pervasiveness of confounding. Although DPA approaches are simpler, methods based on epidemiologic study designs may be better equipped to deal with confounding. However, they present challenges in scaling to high-throughput settings and require many design decisions to be made. Other major challenges in the use of HCD are the definition and ascertainment of exposures and outcomes. Because HCD are not collected for PhV purposes it must be ensured that the data contain sufficient clinical information to correctly capture and validate the exposures and outcomes of interest. Outcomes can be defined in various ways, each of which may have different operating characteristics.[55] Often the exposures and outcomes of interest may not be captured or may not reflect actual experience; e.g., over-the-counter or dietary supplements may not be captured because they are not prescribed or associated with reimbursement. Mild, untreated symptoms and extreme situations such as death without medical care may also not be captured. Actual ingestion, dosage, and filling of a prescription for a drug are hard to ascertain. Even after they have been defined, the actual identification of exposures and outcomes may present challenges when portions of the data are in unstructured, uncoded format, as is the case with health record medical narratives, which may require the use of NLP. A key distinguishing feature of HCD-based methods is the use of temporal information to identify time frames (known as surveillance windows, drug/condition eras, or hazard periods) in which drug–outcome pairs are identified and analyzed, e.g., outcomes recorded 30 days after drug exposure. Essentially, all HCD-based methods define and use some form of time frames to detect ADEs. Using this temporal information, drug safety analysis with HCD can be visualized and analyzed using patient time-line graphs (**Figure 3**).

DPA

There are several ways in which drug–outcome pairs can be counted and mapped into 2 × 2 contingency tables. Zorych *et al.*[56] of the OMOP discuss three approaches that they refer to as "distinct patients," "SRS," and "modified SRS." The first counts the number of distinct patients who experience an outcome within a drug era (even when the same patient experiences multiple outcomes in several drug eras). The second approach attempts to mimic SRSs and treats each drug–outcome occurrence as a spontaneous report. The third approach attempts to take advantage of other information and to augment SRS-like reporting with added denominator information—counting exposures without outcomes as well as outcomes without exposure (SRSs count only exposures reported with an event). Another distinction is made between incident conditions, in which only the first occurrence of an event is counted, and prevalent conditions, in which all occurrences are counted, giving a total of six ways to map the data into 2 × 2 tables. Based on a large-scale systematic evaluation of these counting approaches using the DPA metrics described under "DPA and Basic Concepts," the authors conclude that the SRS and modified SRS approaches using Bayesian metrics provide the best performance.

Some have hypothesized that metrics based on person-time rather than person-counts could produce more accurate association estimates because length of exposure is a more granular and information-carrying quantity than number of persons. Exploiting the temporal information available in HCD, Schuemie[57] proposed an approach known as longitudinal GPS (LGPS). This approach, a modification of the original MGPS approach, uses person-time rather than person-counts to estimate the expected number of events. In LGPS the expected number of occurrences of an event is calculated as the total time during which patients are exposed to a specific drug multiplied by the number of occurrences of the event per unit time when the patients are not exposed. Schuemie also proposed a heuristic approach in conjunction with LGPS to remove spurious associations caused by protopathic bias. The heuristic is based on the assumption that an increase in the number of prescriptions after an event as compared with before the event is an indication of protopathic bias. LGPS has been shown to outperform related methods, including MGPS, and was the winner of the 2010 OMOP Cup competition based on simulated data.[19] A similar approach to LGPS was proposed by Noren *et al.*, who argue that their approach has several important advantages over LGPS and offers better protection against confounding.[58] Applying their method to longitudinal HCD from the UK, the authors were able to demonstrate the timely identification of the association between terbinafine and angioedema.

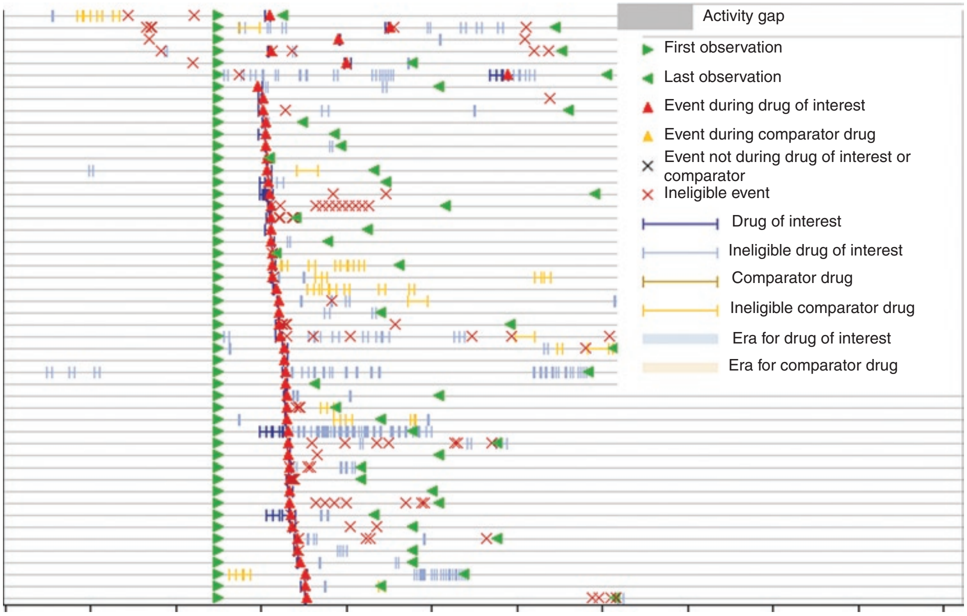

Figure 3 Patient time lines used for visualizing and analyzing health-care data for purposes of drug safety. Each patient is represented by a horizontal line capturing time, and the symbols on the line represent clinical events including diagnoses, test results, treatments, and drug treatment periods. The figure shows time lines of patients who experienced various events such as headache (red triangle) within the eligible periods of acetaminophen administration. The data are sorted according to the first occurrence of an event (red triangle).

Cohort designs

Although there are many variants, the basic concept underlying cohort designs is to partition the subject population into two groups: those who are "exposed" (taking a specific drug) and those who are "unexposed" (taking a comparator drug(s)). The relationship between the exposure and the outcome is then examined by comparing the prevalence of the outcome in the two groups. An association is identified when the outcome occurs more often in the exposed group than in the unexposed group. Comparators may be subjects who are not taking the drug or those taking another drug(s) from the same therapeutic class. However, given the nonrandom assignment of groups, increased attention must be given to the selection of appropriate comparators. Inappropriate selection of comparators may lead to confounding and biases such as channeling (commonly observed when comparing drugs with similar indications), in which imbalances of risk or prognostic factors between groups results in biased effect estimates and thereby lead to unreliable conclusions. To address and minimize these issues, matching (matching of two groups based on a set of covariates such as gender, age, length of exposure, and comorbidities) or propensity scores (PSs) are often employed.

PS methods have become a common analytic approach to control confounding in cohort designs by impersonating the role of randomization in clinical trials.[10,27] A PS is the conditional probability that a subject will receive treatment, given a set of measured preselected covariates (potential confounders). Among subjects with the same propensity to receive treatment, the treatment is conditionally independent of the confounders. This suggests that, within groups of subjects with the same PS, any difference in outcome between the treated and untreated cannot be attributed to the confounders. A PS can also be viewed as a one-dimensional (scalar) value that summarizes a large number of covariates. Treatment–outcome effects can then be estimated by using the PS for matching, stratification, or as an adjustment factor in regression models.[59] A central challenge in the use of PS is the selection of covariates to be included in the model. Incorrect selection may introduce bias into the analysis. There are differing views as to the type of covariates that should be included, i.e., whether the covariates should be related to exposure only, both outcome and exposure, or outcome regardless of exposure.[60]

Schneeweiss *et al.*[61] proposed an algorithm for PS covariate selection that has received much attention lately. It is known as high-dimensional PS (HDPS). The method automatically identifies and selects empirical confounders, estimates propensity scores, and integrates them into an exposure–outcome, PS-based, confounder adjustment model. The authors claim that adjusting for large numbers of covariates ascertained from patients' health-care claims data may improve control of confounding because these variables may collectively be proxies for unobserved factors. Empirical confounders are automatically identified on the basis of a function that incorporates both the prevalence of a covariate and its association with the outcome. Covariates are then ranked on the basis of this function, and the top *k* covariates are selected as the final set of empirical confounders (in addition to the usual demographic covariates). Based on the empirical confounders, a logistic regression is used to estimate a PS for each subject. These PS values are converted into indicator variables on the basis of PS deciles before being used in the final logistic regression model of exposure–outcome to estimate confounder-adjusted associations. In an experiment conducted by the OMOP, HDPS achieved a sensitivity of 56%, specificity of 82%, and positive predictive value of 38% in the detection of 53 associations corresponding to true ADEs and negative controls.[19]

Case–control designs

In a case–control study, the subject population is divided into those experiencing the outcome under investigation ("cases") and a comparator group ("controls"). The relationship between the exposure and the outcome is then examined by comparing the prevalence of the exposure in the two groups. An association is identified when the exposure occurs more often in the cases than in the controls. In case–control designs, the controls will typically consist of subjects who did not experience the outcome being studied but who are otherwise similar. Other options for selection of controls would be to include those who experienced a different set of conditions of interest or those with conditions indicated for the same types of drugs. The main advantage of case–control studies as compared with alternative study designs such as cohort designs is their data efficiency, which permits the study of rare events.[44] Matching is often employed to control for potential confounding factors. In a matched case–control study, each case is matched to one or more controls based on a set of predetermined covariates. Overmatching may introduce bias and should be discouraged.[62] As part of the OMOP's experiment in which HDPS was evaluated, the implementation of a case–control design achieved close to 100% sensitivity, but at the expense of extremely low (15%) specificity.[19]

Self-controlled designs

A self-controlled design can be viewed as a special variant of the case–control design wherein subjects are used as their own controls and outcome rates for periods when a subject is exposed to the drug are compared against periods when the subject is not exposed to the drug. Because the two sets of exposure data relate to the same individual, these designs implicitly control for all time-invariant and subject-invariant confounders (e.g., comorbidities, smoking status, and chronic use of drugs) without the need for confounders to be measured. They also eliminate selection bias. Another advantage of this design approach is that only "cases" need to be included in the analysis. Self-controlled designs can be used when subject data include multiple periods of exposure risk.

The self-controlled case series (SCCS)[63] is a type of self-controlled design. SCCS assumes that adverse events arise according to a nonhomogeneous Poisson process, with each subject having an individual baseline (non-exposure) event rate that is constant over time, and with periods of exposure resulting in a multiplicative effect on the baseline rate. The goal is to estimate the multiplicative effect, which corresponds to the relative risk of an adverse event during an exposure period. Simpson *et al.*[21] of the OMOP were able to demonstrate that if SCCS had been applied to one of the OMOP's observational data sources (i3 claims data for ~50 million subjects) it would have led to detection of the Vioxx–myocardial infarction association 3 years before the drug was withdrawn (2004), whereas AERS-based DPA failed to detect the association. It was also demonstrated that SCCS outperformed DPA methods on most performance metrics.

Although much progress has been made, methodologic research into the use of HCD is currently in its early stages.[64] Ultimately, it is unlikely that the optimal solution will be a one-size-fits-all approach; instead, a process may be developed to refine the analysis to the characteristics of the medical product, outcomes, and databases in question. Gagne *et al.*[65] have laid out a taxonomy of study design considerations based on anticipated drug safety questions, but substantial research is required to validate such recommendations. Establishing best practices requires further empirical evaluation to measure performance of alternative methods across the continuum of expected scenarios. Absent such information, heuristics are applied using expert subjective assessment without supporting empirical evidence.

METHODS INVOLVING NONSTANDARD DATA SOURCES OR LINKED MULTIPLE DATA SOURCES

Chemical and biological information

In a series of articles, Matthews *et al.*[32,33] from the FDA's Center for Drug Evaluation and Research discuss implementation of a system based on quantitative structure activity relationship (QSAR) models to predict ADEs and possible mechanisms of action responsible for adverse events. QSARs are mathematical models used to predict measures of toxicity from physical characteristics of the structures of chemicals. Drug candidates for QSAR modeling were indentified from the AERS using standard DPA approaches and were supplemented with findings from the literature. Commercial QSAR software was then applied to the drug candidates to identify chemical properties of molecules that may correlate with adverse events. The authors built separate QSAR models for each of several adverse events, including cardiac-, liver-, and urinary-related ones. They report an average of 78% specificity and 56% sensitivity for this application, noting that, after data based on the literature were added, there was usually a substantial improvement in performance. They remark that approximately half of the drugs related to hepatobiliary and urinary tract ADEs that were missed in premarket clinical trials could have been predicted using QSAR models. They also found that cardiac ADEs correlate with mechanisms of action (such as the α/β-adrenoceptors and the dopamine and hydroxytryptomine receptors) that affect cardiovascular and cardioneurological functions, and that the screening of new drugs on the basis of these mechanisms of action could predict the majority of cardiac ADEs. The QSAR models are now being used internally by the FDA to provide decision-support information for a variety of regulatory activities.

Villar *et al.*[66] proposed a SAR modeling technique to prioritize ADE associations generated from the AERS. The authors compiled, from the literature, a reference set of drugs related to rhabdomyolysis and mapped them to two-dimensional molecular fingerprints (bit vectors that represent the presence/absence of specific structural features) using information available in DrugBank and commercial software. The initial drug candidates, generated from the AERS by the MGPS algorithm, were then screened by comparing their structural fingerprints with the reference set of fingerprints. Highly similar candidates were then retained as the final set of drug candidates. Using this approach, the authors achieved 70% sensitivity, 45% positive predictive value, and a greater than twofold enrichment of AERS signals.

Publicly available preclinical molecular screening assays such as those in PubChem can be mined to correlate a drug's bioactivity with postmarketing ADEs. Pouliot *et al.*[30] created models to correlate a drug's propensity to cause specific system organ class (SOC) ADEs. They used data from more than 487,000 drug activity screens from the National Center for Biotechnology Information's PubChem BioAssay database and SOC-specific ADE information from the Canadian Adverse Drug Reaction database to create logistic regression models for nine SOCs. They validated these models by performing retro-prediction for eight individual drugs and found that 75% of the predicted adverse reactions in humans could be substantiated by information in the literature or in the drug label. Using these validated models, they predicted not-yet-recognized ADEs associated with three drugs that had recently been approved or were awaiting approval in the United States. The authors note that making such predictions can generate testable hypotheses for the identification of ADEs in the clinical setting, thereby shortening the period during which new ADEs go unrecognized.

Biomedical literature

Shetty *et al.*[35] describe an approach for collecting, filtering, and analyzing biomedical literature as a complementary strategy for prioritizing ADE associations generated from SRSs. First, all articles mentioning drug–outcome pairs from a predefined set of drugs and events were retrieved from PubMed. Next, NLP was used to identify and exclude the articles that mentioned irrelevant pairs (e.g., pairs that capture a treatment relationship). Finally, DPA was applied to the pairs mentioned in the rest of the retrieved articles to highlight statistically significant drug–event associations. The authors show that this method discovered true associations with over 70% sensitivity and 40% positive predictive value as assessed against a reference set of true ADE associations obtained from the "Warnings" section of drug labels. They also demonstrate that, through the use of this approach, 54% of the associations analyzed could have been detected before the FDA warnings about them and that the Vioxx–myocardial infarction association could have been identified using literature published before 2002.

User-generated content in health forums

In a recent study, Leaman *et al.*[36] demonstrated that information posted by users on health-related websites contain extractable information relevant to PhV. The authors also described a prototype system to mine this type of information source. In their approach, raw data was automatically collected with a Web crawler from the website DailyStrength. NLP techniques were used to process the raw data and extract clinical concepts related to ADEs. Special procedures were used to deal with colloquial phrases (e.g., "zoned out" to mean "somnolent") and spelling errors. The system was evaluated using an expert-annotated set of 3,600 user-generated posts corresponding to six drugs. The system achieved 78% precision and 70% recall in correctly labeling the user-generated data. Importantly, the authors found that the incidence of ADEs reported by users is highly correlated with documented incidence rates listed by the FDA. They also noted that the ADEs most frequently identified through this method corresponded to well known ADEs. The website PatientsLikeMe recently published[67] a study based on its data. The user community of this website essentially self-assigned themselves to assess the efficacy of lithium as a treatment for patients with amyotrophic lateral sclerosis. While this study does not constitute safety surveillance, it is one that illustrates the promise of patient-initiated observational studies.

Linking of multiple knowledge sources

By linking information from multiple sources, Cami *et al.*[68] proposed a network-based model to predict ADEs. The authors first constructed a network representation of drug–event associations that were known as of the year 2005. Then, using network topological indices (e.g., node degree), supplemented with ontological features (e.g., distance between two events in the MedDRA hierarchy) and molecular descriptors (e.g., the drug's molecular weight and melting point), the authors trained a logistic regression model to predict the probability of an unknown ADE association (edge in the network graph). The predictive performance of the model was prospectively validated by predicting ADEs reported in the years 2006–2010. The model achieved an area under the receiver operating characteristic curve of 0.87, sensitivity of 42%, and specificity of 95%. The authors were also able to predict seven of eight ADEs that emerged after 2005, including those between the seizure drug zonisamide and suicidal thoughts, the antibiotic norfloxacin and ruptured tendons, and the controversial diabetes drug rosiglitazone (Avandia) and heart attacks. The authors claim that, unlike related work, the prospective characteristics of their model make it a realistic method for predicting future ADEs.

By integrating information from the AERS and several HCD sources, Tatonetti *et al.*[69] discovered a potentially new drug interaction between two widely used drugs—the antidepressant paroxetine and the cholesterol-lowering medication pravastatin—that can lead to unexpected increases in blood glucose levels. The finding that motivated their data-mining approach was the observation that side effects are not independent of each other and latent evidence for an (unreported) adverse event can be found by examining other (reported) side effects. By scanning the AERS for pairs of drugs that have matching side-effect profiles when taken together but not when taken individually, the authors created a candidate set of drug–drug interactions. The list of candidates was then narrowed down to the paroxetine–pravastatin interaction by conducting retrospective studies using electronic health records from Stanford University Hospital, Vanderbilt University Hospital, and Partners Healthcare. Finally, the interaction was confirmed by a prospective study in an insulin-resistant mouse model.

CONCLUDING REMARKS, FUTURE PERSPECTIVES, AND CHALLENGES

We have shown that a rich and diverse portfolio of data-mining approaches aligned to different strategies and objectives is now available for the analysis and detection of postapproval ADEs. Each approach may offer unique prospects that collectively can advance the science of drug safety surveillance.

Renewed interest and new opportunities have emerged to harness data that have not traditionally been used in PhV, allowing for new active and proactive paradigms of surveillance. Although methodologic research is now shifting away from the use of SRSs, this will not diminish the important role or value of spontaneous reports. That said, the use of spontaneous report narratives to enhance SRS-based discovery is yet to be explored. It is also evident that a new trend is emerging of data mining being used to link human safety information with experimental platforms that have been traditionally used in the preclinical drug discovery phase. The diversity of approaches highlights the value of systems that can span data and expertise across multiple domains. Nonetheless, to fully realize this potential, new and creative methods will be required to integrate these disparate sources in a more synergistic manner.

It has been suggested that a revisit of randomized clinical trials data, by synthesis or pooling of several related trials, should be employed to augment findings from other sources. The main benefit to this approach is that it enjoys the scientific and statistical benefits of randomization.[70] Although beyond the scope of this article, it is important to emphasize the role of pharmacogenomics research[29,71] and the utility of knowledge bases such as PharmGKB[72] that could further enhance our understanding of ADEs by correlating human genetic variations with drug toxicity.

Although much progress has been made in utilizing healthcare data, a substantial amount of further empirical assessment is required, using both real and simulated data sets. A key OMOP finding is that the heterogeneity of data sources and methods strongly affects results. Therefore, consistent methods that can be applied to multiple data sources will be required. There are also opportunities to improve data quality, coding standards, sharing, and access. The relative value of information contained in the electronic health records as compared with that in administrative claims databases needs to be further explored. The continued development of simulated data for which ground-truth is available is critically important to further the understanding of method efficacy.

A central challenge in PhV research is the lack of established standards to evaluate DMAs. One of the main contributors to this problem is the lack of a gold standard because the set and nature of all possible drug safety issues are unknown. Although suggestions have been made, there is an ongoing debate about the evaluation strategies and what should constitute reference standards for DMA evaluation.[7,73] With that said, there is currently little empirical evidence to support or prefer the use of one method or data source over another. This is why efforts such as those by the OMOP and EU-ADR are of paramount importance so that methods and data sources can be assessed on a solid scientific footing.

Typically, DMA activity is conducted at certain levels of granularity of medical terminologies that are not optimally designed to support PhV. Often, similar medical concepts are fragmented across distinct terms, weakening the potential for statistical discovery. Therefore, methods that can make better use of and integrate knowledge from lexical resources[74] may prove beneficial. As a related step, Bayesian approaches allowing for information borrowing[75] across similar terms and drugs should be further developed and evaluated. In addition, improved NLP methods to process unstructured textual data, whether from clinical narratives, literature, or health forums, will continue to play an important role.[25]

Last but not least, it is important to recognize that, at its core, data mining is a tool to formulate or refine new hypotheses and thus will not eliminate the important role of medical review, which will always be required for final adjudication of causality.

ACKNOWLEDGMENTS

This work was supported in part by grants 1R01LM010016, 3R01LM010016-01S1, 3R01LM010016-02S1, and 5T15-LM007079-19(HS) from the National Library of Medicine; U54-HG004028 from the National Center for Biomedical Ontology; and 2U54LM008748 from the i2b2 National Center for Biomedical Computing. We thank Nicholas Tatonetti of Stanford University and Roee Sa'adon of First Life Research for their valuable comments. We are grateful to Oracle's Health Sciences Division for supplying data and figures.

CONFLICT OF INTEREST

D.M. has served as a consultant to lawyers representing plaintiffs in litigation involving Vioxx . The authors declared no conflict of interest.

1. World Health Organization. The Importance of Pharmacovigilance—Safety Monitoring of Medicinal Products. World Health Organization: Geneva, 2002.
2. Lazarou, J., Pomeranz, B.H. & Corey, P.N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* **279**, 1200–1205 (1998).
3. Classen, D.C., Pestotnik, S.L., Evans, R.S., Lloyd, J.F. & Burke, J.P. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. *JAMA* **277**, 301–306 (1997).
4. Ahmad, S.R. Adverse drug event monitoring at the Food and Drug Administration. *J. Gen. Intern. Med.* **18**, 57–60 (2003).
5. Bate, A. & Evans, S.J. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol. Drug Saf.* **18**, 427–436 (2009).
6. Hauben, M., Madigan, D., Gerrits, C.M., Walsh, L. & Van Puijenbroek, E.P. The role of data mining in pharmacovigilance. *Expert Opin. Drug Saf.* **4**, 929–948 (2005).
7. Hauben, M. & Bate, A. Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov. Today* **14**, 343–357 (2009).
8. Almenoff, J.S., Pattishall, E.N., Gibbs, T.G., DuMouchel, W., Evans, S.J. & Yuen, N. Novel statistical tools for monitoring the safety of marketed drugs. *Clin. Pharmacol. Ther.* **82**, 157–166 (2007).
9. Platt, R., Wilson, M., Chan, K.A., Benner, J.S., Marchibroda, J. & McClellan, M. The new sentinel network - improving the evidence of medical-product safety. *N. Eng. J. Med.* **361**, 645–647 (2009).
10. Schneeweiss, S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol. Drug Saf.* **19**, 858–868 (2010).
11. Wysowski, D.K. & Swartz, L. Adverse drug event surveillance and drug withdrawals in the United States, 1969-2002: the importance of reporting suspected reactions. *Arch. Intern. Med.* **165**, 1363–1369 (2005).
12. Stephenson, W.P. & Hauben, M. Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiol. Drug Saf.* **16**, 359–365 (2007).
13. Graham, D.J. *et al.* Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet* **365**, 475–481 (2005).
14. Avorn, J. & Schneeweiss, S. Managing drug-risk information—what to do with all those new numbers. *N. Engl. J. Med.* **361**, 647–649 (2009).
15. Stang, P.E. *et al.* Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann. Intern. Med.* **153**, 600–606 (2010).
16. McClellan, M. Drug safety reform at the FDA–pendulum swing or systematic improvement? *N. Engl. J. Med.* **356**, 1700–1702 (2007).
17. US Food and Drug Administration Amendments Act (FDAAA) of 2007 <http://www.fda.gov/RegulatoryInformation/Legislation/federalfooddrugandcosmeticactfdcact/significantamendmentstothefdcact/foodanddrugadministrationamendmentsactof2007/default.htm>. Accessed February 2012.
18. The Sentinel Initiative: a national strategy for monitoring medical product safety <http://www.fda.gov/downloads/Safety/FDAsSentinelInitiative/UCM124701.pdf>. Accessed February 2012.
19. Observational Medical Outcomes Partnership (OMOP) <http://omop.fnih.org/>. Accessed February 2012.
20. Coloma, P.M. *et al.* Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol. Drug Saf.* **20**, 1–11 (2011).
21. Simpson, S.E. Self-controlled methods for postmarketing drug safety surveillance in large-scale longitudinal data. Ph.D. dissertation, Columbia University, 2011.
22. LePendu, P., Iyer, S.V., Fairon, C. & Shah, N.H. Annotation analysis for testing drug safety signals using unstructured clinical notes. *J. Biol. Sem.* **3** (suppl. 1), S5 (2012).
23. Brownstein, J.S., Sordo, M., Kohane, I.S. & Mandl, K.D. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. *PLoS ONE* **2**, e840 (2007).
24. Brown, J.S. *et al.* Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiol. Drug Saf.* **16**, 1275–1284 (2007).
25. Chapman, W.W., Nadkarni, P.M., Hirschman, L., D'Avolio, L.W., Savova, G.K. & Uzuner, O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J. Am. Med. Inform. Assoc.* **18**, 540–543 (2011).
26. Brookhart, M.A., Stürmer, T., Glynn, R.J., Rassen, J. & Schneeweiss, S. Confounding control in healthcare database research: challenges and potential approaches. *Med. Care* **48**, S114–S120 (2010).
27. Schneeweiss, S. & Avorn, J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J. Clin. Epidemiol.* **58**, 323–337 (2005).
28. Wishart, D.S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**, D901–D906 (2008).
29. Chiang, A.P. & Butte, A.J. Data-driven methods to discover molecular determinants of serious adverse drug events. *Clin. Pharmacol. Ther.* **85**, 259–268 (2009).
30. Pouliot, Y., Chiang, A.P. & Butte, A.J. Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clin. Pharmacol. Ther.* **90**, 90–99 (2011).
31. Bender, A. & Glen, R.C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2**, 3204–3218 (2004).
32. Matthews, E.J. *et al.* Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part B. Use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. *Regul. Toxicol. Pharmacol.* **54**, 23–42 (2009).
33. Frid, A.A. & Matthews, E.J. Prediction of drug-related cardiac adverse effects in humans–B: use of QSAR programs for early detection of drug-induced cardiac toxicities. *Regul. Toxicol. Pharmacol.* **56**, 276–289 (2010).

34. Deftereos, S.N., Andronis, C., Friedla, E.J., Persidis, A. & Persidis, A. Drug repurposing and adverse event prediction using high-throughput literature analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **3**, 323–334 (2011).
35. Shetty, K.D. & Dalal, S.R. Using information mining of the medical literature to improve drug safety. *J. Am. Med. Inform. Assoc.* **18**, 668–674 (2011).
36. Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J. & Gonzalez, G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts in health-related social networks. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, 117–125 (2010).
37. Szarfman, A., Machado, S.G. & O'Neill, R.T. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf.* **25**, 381–392 (2002).
38. Bate, A. *et al*. A Bayesian neural network method for adverse drug reaction signal generation. *Eur. J. Clin. Pharmacol.* **54**, 315–321 (1998).
39. DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System. *Am. Stat.* **53**, 177–190 (1999).
40. Szarfman, A. Safety Data Mining. FDA Advisory Committee Meeting Briefing Document (2006) <http://www.fda.gov/ohrms/dockets/ac/06/briefing/2006-4266b1-02-06-FDA-appendic-f.pdf>. Accessed December 2011.
41. Deshpande, G., Gogolak, V. & Sheila, W.S. Data mining in drug safety: review of published threshold criteria for defining signals of disproportionate reporting. *Pharmaceut. Med.* **24**, 37–43 (2010).
42. Ramirez, E. *et al*. A pharmacovigilance program from laboratory signals for the detection and reporting of serious adverse drug reactions in hospitalized patients. *Clin. Pharmacol. Ther.* **87**, 74–86 (2010).
43. Almenoff, J.S., DuMouchel, W., Kindman, L.A., Yang, X. & Fram, D. Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. *Pharmacoepidemiol. Drug Saf.* **12**, 517–521 (2003).
44. Jewell, N.P. *Statistics for Epidemiology* (Chapman and Hall, Boca Raton, FL, 2003).
45. Genkin, A., Lewis, D.D. & Madigan, D. Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49**, 29–-304 (2007).
46. Caster, O., Noren, G.N., Madigan, D. & Bate, A. Large-scale regression-based pattern discovery: the example of screening the WHO global drug safety database. *Stat. Analy. Data Mining* **3**, 197–208 (2010).
47. Solomon, R. & DuMouchel, W. Contrast media and nephropathy: findings from systematic analysis and Food and Drug Administration reports of adverse effects. *Invest. Radiol.* **41**, 651–660 (2006).
48. Agrawal, R., Imielinski, T. & Swami, A. Mining association rules between sets of items in large databases. *SIGMOD*, 207–216 (1993).
49. Rouane, H., Toussaint, Y. & Valtchev, P. *Mining Signals in Spontaneous Reports Database Using Concept Analysis*. AIME 2009 (Springer, Berlin, Germany, 2009).
50. Harpaz, R., Chase, H.S. & Friedman, C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics* **11** (suppl. 9), S7 (2010).
51. Zorych, I., Madigan, D. Ryan, P. & Bate, A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Stat Methods Med Res.* 30 August 2011 (doi:10.1177/0962280211403602).
52. Fan, K. *et al*. High-performance signal detection for adverse drug events using MapReduce paradigm. *AMIA Annu. Symp. Proc.* **2010**, 902–906 (2010).
53. Harpaz, R., Perez, H., Chase, H.S., Rabadan, R., Hripcsak, G. & Friedman, C. Biclustering of adverse drug events in the FDA's spontaneous reporting system. *Clin. Pharmacol. Ther.* **89**, 243–250 (2011).
54. Ball, R. & Botsis, T. Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS? *Clin. Pharmacol. Ther.* **90**, 271–278 (2011).
55. Stang, P.E. *et al*. Health outcomes of interest in observational data: issues in identifying definitions in the literature. *Health Outcomes Res. Med.* **3**, e37–e44 (2012).
56. Zorych, I., Madigan, D., Ryan, P. & Bate, A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Stat Methods Med Res* (in press).
57. Schuemie, M.J. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiol. Drug Saf.* **20**, 292–299 (2011).
58. Norén, G.N., Hopstadius, J., Bate, A. & Edwards, I.R. Safety surveillance of longitudinal databases: methodological considerations. *Pharmacoepidemiol. Drug Saf.* **20**, 714–717 (2011).
59. D'Agostino, R.B. Jr & D'Agostino, R.B.Sr. Estimating treatment effects using observational data. *JAMA* **297**, 314–316 (2007).
60. Patrick, A.R. *et al*. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiol. Drug Saf.* **20**, 551–559 (2011).
61. Schneeweiss, S., Rassen, J.A., Glynn, R.J., Avorn, J., Mogun, H. & Brookhart, M.A. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20**, 512–522 (2009).
62. Bloom, M.S., Schisterman, E.F. & Hediger, M.L. The use and misuse of matching in case-control studies: the example of polycystic ovary syndrome. *Fertil. Steril.* **88**, 707–710 (2007).
63. Madigan, D., Ryan, P., Simpson, S.E. & Zorych, I. Bayesian methods in pharmacovigilance (with discussion). In *Bayesian Statistics 9* (eds. Bernardo, J.M. *et al.*) 421–438 (Oxford University Press, Oxford, UK,, 2010).
64. Madigan, D. & Ryan, P. What can we really learn from observational studies?: the need for empirical assessment of methodology for active drug safety surveillance and comparative effectiveness research. *Epidemiology* **22**, 629–631 (2011).
65. Gagne, J.J. *et al*. Design considerations in an active medical product safety monitoring system. *Pharmacoepidemiol. Drug Saf.* **21** (suppl. 1), 32–40 (2012).
66. Vilar, S., Harpaz, R., Chase, H.S., Costanzi, S., Rabadan, R. & Friedman, C. Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. *J. Am. Med. Inform. Assoc.* **18** (suppl. 1), i73–i80 (2011).
67. Wicks, P., Vaughan, T.E., Massagli, M.P. & Heywood, J. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nat. Biotechnol.* **29**, 411–414 (2011).
68. Cami, A., Arnold, A., Manzi, S. & Reis, B. Predicting adverse drug events using pharmacological network models. *Sci. Transl. Med.* **3**, 114ra127 (2011).
69. Tatonetti, N.P. *et al*. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clin. Pharmacol. Ther.* **90**, 133–142 (2011).
70. Gibbons, R.D. *et al*. Post-approval drug safety surveillance. *Annu. Rev. Public Health* **31**, 419–437 (2010).
71. Becquemont, L. Pharmacogenomics of adverse drug reactions: practical applications and perspectives. *Pharmacogenomics* **10**, 961–969 (2009).
72. Pharmacogenomics Knowledge Base (PharmGKB) <http://www.pharmgkb.org>. Accessed February 2012.
73. Hochberg, A.M. *et al*. An evaluation of three signal-detection algorithms using a highly inclusive reference event database. *Drug Saf.* **32**, 509–525 (2009).
74. Musen, M.A. *et al*. The National Center for Biomedical Ontology. *J. Am. Med. Inform. Assoc.* **19**, 190–195 (2012).
75. Berry, S.M. & Berry, D.A. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics* **60**, 418–426 (2004).