Curr Epidemiol Rep (2014) 1:110–119
DOI 10.1007/s40471-014-0016-2

EPIDEMIOLOGIC METHODS (T LASH, SECTION EDITOR)

Data Mining Methods to Detect Sentinel Associations and Their Application to Drug Safety Surveillance

Preciosa M. Coloma · Sandra de Bie

Published online: 22 June 2014

Abstract Data mining techniques identify hitherto unsuspected patterns in data using an algorithmic approach. Such techniques have been used for a long time to support day-to-day operations of organizations handling large volumes of data, including banks, airlines, and retail organizations. In the context of healthcare and medicine, data mining as a field flourished with the establishment of vast repositories collecting biomedical literature, genomics data, as well as population-based electronic medical records and reimbursement claims for health services. Although many of the statistical techniques used in data mining are similar to conventional methods of interrogating data, the datasets that are being mined are massive and multi-dimensional. In healthcare, this translates to millions of patients—far beyond the domain of clinical trials or standard epidemiologic studies—and their associated patient characteristics with all its permutations. This report provides an overview of data mining methods for detection of sentinel associations, with a specific focus on their applicability to surveillance of drug (or vaccine)-related sentinel associations.

Keywords Data mining · Drug-related sentinel associations · Drug safety monitoring · Drug safety surveillance · Electronic healthcare databases · Epidemiology · Pharmacoepidemiology · Pharmacovigilance · Signal detection · Vaccine surveillance

P. M. Coloma (✉) · S. de Bie
Department of Medical Informatics, Erasmus MC University Medical Center, 's Gravendijkwal 230, Rotterdam, The Netherlands
e-mail: p.coloma@erasmusmc.nl

S. de Bie
Department of Internal Medicine, Erasmus MC University Medical Center, Rotterdam, The Netherlands

Introduction

Data mining is a process that involves an algorithmic and database-oriented approach to find previously unsuspected patterns in data [1]. Data mining approaches have been used for a long time in the financial industry, supporting day-to-day operations of organizations handling large volumes of data such as banks, airlines, and retail organizations. It emerged as a result of huge developments in the field of machine learning; in the context of healthcare and medicine, data mining as a field flourished upon the establishment of vast repositories collecting biomedical literature, genomics (and other omics) data, as well as electronic medical records and reimbursement claims for health services covering several geographic regions or entire nations. Many of the statistical techniques used in data mining are similar to conventional methods of interrogating data and are aimed towards finding correlations, links, or dissimilarities between groups of data to generate new information. What makes data mining both interesting and challenging is its scale and complexity: the datasets that are being referred to are massive (in millions of terabytes and growing by the minute) and multi-dimensional. In healthcare, this translates to millions of patients—far beyond the domain of clinical trials or standard epidemiologic studies—and their associated patient characteristics, with all their permutations. In epidemiology, it is primarily statistical power that drives the pursuit of data mining, which enables the investigation of large populations that represent an entire spectrum of exposures and outcomes of interest and extensive observation periods that permit long-term follow-up of patients. These features become particularly important when conducting studies involving outcomes that are rare or have a long latency. Increase in study sample size and power, in turn, provide the opportunity for timely assessment of health-related issues, including the safety and effectiveness of medical treatments and public health interventions. Disease predisposition and

manifestations often vary in populations because of ethnicity or exposures peculiar to a group, and patients respond differently to healthcare interventions. Data from a wide variety of sources allows a broader scope for investigation as well as evaluation of generalizability of results.

This report provides an overview of data mining methods for detection of sentinel associations, with a specific focus on their applicability to surveillance of drug (or vaccine)-related sentinel associations. The term '*sentinel*' reflects the need for immediate investigation of, and response to, such associations. A sentinel event, as defined by the Joint Commission, constitutes an " unexpected occurrence involving death or serious physical or psychological injury, or the risk thereof" (http://www.jointcommission.org/Sentinel_Event_Policy_and_Procedures). A drug-related sentinel association, more commonly referred to as a '*signal*' in pharmacovigilance, represents an association that is novel and important and demands further investigation and confirmation and, when necessary, remedial actions [2]. In contrast to the Joint Commission's definition, an event within the context of a signal in pharmacovigilance may be either adverse or beneficial. Mining data sources of different configurations and settings and originating from different geographical areas is a field of research that is particularly gaining ground in pharmacoepidemiology and pharmacovigilance. Knowledge of the safety profile of medications (and vaccines) before marketing is limited because of the small and selective groups of individuals included in clinical trials. The prevailing system of post-marketing surveillance is passive and relies on spontaneous reporting systems, which, although sometimes effective, are not always efficient. Surveillance using population-based electronic healthcare databases representing real-world circumstances offers an additional route that can facilitate earlier detection and earlier management of potential safety issues.

Methodological Approach to Detection of Drug-Related Sentinel Associations

It was in the aftermath of the thalidomide tragedy in the late 1960s [3] that the US Food and Drug Administration (FDA), the World Health Organization (WHO) and the UK's Medicines and Healthcare products Regulatory Agency (MHRA) independently set up voluntary adverse drug reaction (ADR) reporting systems that collect and subsequently analyze post-marketing safety information. Establishment of other national ADR reporting databases soon followed. More than 70 countries, including a number of developing countries, currently have their own reporting systems, which attempt to ensure that potential signals are detected as soon as possible after licensing [4••]. The largest spontaneous reporting databases available worldwide include the FDA's Adverse Event Reporting System (FAERS; http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects), the Vaccine Adverse Event Reporting System (VAERS; http://vaers.hhs.gov/), EudraVigilance (https://eudravigilance.ema.europa.eu/), and the WHO Uppsala Monitoring Center's VigiBase (http://www.who-umc.org/).

The adoption of a more proactive approach to detection of drug-related sentinel associations has resulted in exploration of accessible data resources, whether or not the data are collected for the primary purpose of drug safety monitoring. These potential additional resources include electronic medical records with detailed clinical information such as patients' symptoms, physical examination findings, laboratory procedures, diagnostic test results, and prescribed medications or other interventions. Automated recording of filled prescriptions, laboratory and ancillary tests as well as hospitalizations are increasingly collected routinely for audit and reimbursement of various health services. These electronic healthcare databases (medical records databases and administrative/claims databases) have previously been employed to characterize healthcare utilization patterns, monitor patient outcomes, and carry out formal pharmacoepidemiological studies [5–7]. With regard to drug safety surveillance, such databases have been commonly used to confirm or refute potential signals detected initially from analysis of individual case safety reports. The longitudinal nature of routinely collected healthcare data may allow identification of clinical (especially adverse) outcomes that have a long delay between exposure and clinical manifestations, such as pulmonary/cardiac fibrosis or cancer. While most spontaneous reports usually involve newly marketed drugs, electronic healthcare data may be able to highlight new risks associated with old drugs that are a consequence of new indications of use or a new generation of users, as well as adverse events that are not pharmacologically predictable and less likely to be suspected as drug induced, and thus less likely to be reported.

The choice of data mining technique and strategy for use in the detection of drug-related (as well as other types of) sentinel associations is driven by the properties of the dataset and such that the utility of the data is maximized while its weaknesses are minimized. Akin to searching for the proverbial needle in the haystack, data mining methods are better able to find the needle when the size of the haystack increases [8]. Although this may seem paradoxical, it actually makes sense since the more data available, the greater is the chance to identify an important but rare outcome.

Prospective Versus Retrospective Surveillance

Prospective surveillance entails monitoring as the data accrue ('real-time' surveillance). Retrospective surveillance, on the other hand, involves analyzing signals using data collected from a previous time period. The methods most commonly employed for prospective surveillance

include sequential analysis techniques that allow testing of pre-specified hypotheses at multiple points in time throughout the course of a study as data accumulate. Group sequential monitoring involves analyzing data after a group of individuals accrues in the study, while continuous sequential monitoring involves analyzing the data after each individual accrues. The maximized sequential probability ratio test (maxSPRT) has been successfully used (on a weekly basis) for prospective vaccine safety monitoring within the Vaccine Safety Datalink (VSD) in the USA [9], while group sequential methods have been used to monitor medical product efficacy/safety in clinical trials on a less frequent basis (i.e., quarterly, or semi-annually, throughout the trial) [10]. Other sequential analysis techniques include control charts originally intended for detecting changes in manufacturing and process control. Bayesian-based methods have likewise been explored in sequential monitoring, especially in the context of clinical trials [11]. Important issues to consider when selecting methods for prospective surveillance using automated databases include establishing the appropriate (1) frequency of sequential testing; (2) signaling threshold (i.e., 'stopping rules'); (3) test statistic to quantify risk between the groups being compared; and (4) adjustment for bias and confounding. Minimizing the time to signal detection is a key objective when performing prospective surveillance.

Assessment of Beneficial Versus Adverse Associations

Methodologies appropriate for the quantitative analysis of drug-related adverse outcomes are not as well-developed or as prevalent as those for evaluation of benefits (i.e., efficacy) [12]. A fundamental concept in safety evaluation is the estimation of incidence rates and increased risks that are time related; this concept requires methods that account for differential patient exposure time to distinguish between acute and chronic exposures. While it is a given that to have a reasonable statistical chance to detect rare events and to demonstrate increased risk, drug safety surveillance must involve large numbers of individuals, the number of cases needed to trigger a signal is not well-defined and often depends on the outcome of interest and corresponding background incidence rates. An attempt to address this issue of adequate statistical power has been carried out by investigators working with a network of European healthcare databases assembled for the purpose of drug safety surveillance. In this study, the investigators provided estimates of the number and types of drugs that can be monitored as a function of actual exposure, minimal detectable relative risks, and empirically derived incidence rates for a wide range of adverse events [13]. Data simulation was also performed in order to find out to what extent expansion of database size would affect the power to detect signals.

Methods of Evaluation in the Context of Drug-Related Sentinel Associations

Table 1 shows the most important factors to consider when selecting a suitable method for detecting sentinel associations (whether specifically drug related or not). All methods should have the ability to confirm acknowledged 'positive' associations (i.e., sensitive) and at the same time to not detect known spurious 'negative' associations (i.e., specific). Striking a balance between sensitivity and specificity remains a challenge for all methodologies used in the detection of sentinel associations and especially because an appropriate gold standard is often lacking. In drug safety signal detection, a definitive list of all known associations, which drugs can cause them, and the date on which such causality was confirmed does not exist. Surrogate reference standards have been constructed ad hoc and, although still far from being complete or comprehensive, they have been quite useful for evaluation purposes [14, 15••, 16••, 17••]. There is no ideal method that can assess with high precision every newly discovered signal. In practice, this is done by manual evaluation by human experts, utilizing clinical judgment based on a series of cases involving similar conditions for each new signal. The extensive application of automated methods to sift through 'big data' needs to be approached with caution, and the aim should be to better define the predictive value of these techniques as well as their added value as adjuncts to traditional methods. Difficulties in distinguishing between limitations of the methods and limitations of the data themselves can sometimes further complicate the evaluation process.

Appraisal and comparison of signal detection methods can be done by looking at concordance of results achieved by different methods when implemented on specific datasets, including artificial (simulated) data, and performance vis-à-vis some definition of true or false associations (using surrogates such as changes in labeling or safety withdrawals) [18]. However, it is important to remember that while a method may successfully detect *known* associations, this is not a guarantee that such a method will also be able to detect signals, i.e., new, currently unknown, drug-related adverse events [19].

Overview of Methods

This section provides a brief discussion of the following methods and their application to signal detection in pharmacovigilance: Bayesian approach; disproportionality

Table 1 Factors to consider when choosing data mining methods in the context of detection of sentinel associations using large databases

Statistically sound	It takes into account sampling variation and multiple comparisons It has mechanisms to investigate confounding, bias, and effect modification
Timeliness	It employs judicious sequencing of methodological steps to enable timely generation of signals
Flexibility	It accepts a generic dataset that allows compatibility with the broadest possible range of sources and permits simultaneous analyses within each dataset
Transparency and interpretability of results	The process is easy to track, is reproducible, and does not constitute a 'black box.' The parameters used to evaluate the data enable intuitive interpretation of results. Summarized graphical displays can provide insight into multivariate relationships, enabling better understanding and visualization of event timelines as well as changes in patterns of exposures
Robustness	It performs well over a variety of datasets, and even if its assumptions are somewhat violated by the true model from which the data were generated
Computational efficiency, portability	It is able to manage large volumes of data within an acceptable running time and could be implemented in a distributed mode

analyses; combined data-and-knowledge mining; biosurveillance methods; chart-based methods; and algorithm-based methods.

Methods with a Bayesian Framework

Bayesian methods interpret data in the light of external evidence and judgment and provide a formal statistical framework to combine domain knowledge (expressed as probability distributions) with data. Prior probability of hypotheses is taken into account and the form in which conclusions are drawn contributes intuitively to decision-making [20, 21]. Detection of sentinel associations—and particularly monitoring drug-related signals—involves a constant need to update posterior probabilities as data accumulates, for which the Bayesian approach is very suitable because repeated testing is fundamental to the Bayesian concept rather than an additional issue to be considered in the analysis. Additional advantages of this approach include (1) shrinkage of imprecise outliers toward the mean or null difference, with high-variance differences shrinking the most, thus resulting in a reduction of the number of false positives that would otherwise be flagged for further investigation; and (2) calculation of (reliable) estimates despite missing data [22]. The most widely used Bayesian-based methods for detecting drug-related outcomes include two methods that are described in the "Disproportionality-Based Methods" section [Bayesian Propagation Confidence Neural Network (BCPNN) and Multi-item Gamma Poisson Shrinker (MGPS)]. Two other frequently used methods, Bayesian updating and Bayesian hierarchical modeling, have been applied mostly in the context of clinical trials [23]. Bayesian updating formally derives estimates of risk of adverse events following drug use or other intervention through combination of prior risk estimates with data from current experience [24]. An often-tricky issue in Bayesian updating is the requirement of seemingly subjective estimates of prior probabilities. But what is considered subjective information often comes from previously observed data, as in the case of published studies. Even expert opinion that appears subjective may, in fact, be based on empiric data (i.e., clinical experience); what makes it subjective is that incorporation of observed data into the 'expert opinion' is not quantitative. However, this should not be so much of a problem since the effect of prior probabilities is large only when the data are weak and unconvincing. Bayesian hierarchical modeling, on the other hand, is a method designed to accommodate dependence in multivariate data to allow many types of adverse events (or many types of exposures) to be simultaneously addressed and provides for 'borrowing information' across data from different sources. One application of this method is the idea that while adverse events involving the same body system may or may not be related, rates of adverse events are more likely to be similar within than across body systems (a hierarchical model allows for this possibility but does not impose it); this model is predicated on the assumption of exchangeability of an adverse event within the same body system [25]. The hierarchical nature of the model gives rise to a regression effect and, in the context of multiple comparisons, modulates extremes. Overall, Bayesian hierarchical modeling seems more useful as a supplementary method to account for multiple comparisons than as a primary signaling tool.

Disproportionality-Based Methods

Disproportionality-based methods have their origins in spontaneous ADR reporting systems and have been employed, among others, in the WHO's Uppsala Monitoring Centre VigiBase (BPCNN) [26], the FDA's FAERS (MGPS) [27, 28], the UK's MHRA Yellow Card database and the European Medicine Agency's EudraVigilance [method estimating the proportional reporting ratio (PRR)] [29], and the Netherlands

Pharmacovigilance Foundation Lareb [method estimating the reporting odds ratio (ROR)] [30]. Disproportionality techniques have also been used in prescription event monitoring database systems [31]. In general, these methods look for unexpected frequencies of reports in the dataset in comparison to general reporting frequencies. The number of reports of a particular drug–event combination is compared with an estimate of the expected number based on other reports in the same dataset. When a high number of cases is observed relative to the number expected, then such combinations are flagged for further investigation. While disproportionality calculations using these methods are routinely performed against a comparison of the whole dataset, secondary analyses are also often done against subsets of the data, sometimes using certain exclusion criteria and performing stratification and sensitivity analyses and 'best case–worst case analyses' of the results. Such sub-analyses help account for potential confounding and/or bias. With BCPNN it is possible to do time scans, which are retrospective investigation of the change in the disproportionality measure IC (information component) for a specific drug–adverse event combination over time. To evaluate how a drug association compares in unexpectedness to related drugs that might be used for the same clinical indication, the method is extended to consideration of groups of drugs. Compared with BCPNN, MGPS [which has the Empirical Bayes Geometric Mean (EBGM) as its disproportionality measure of interest] is generally more flexible and can possibly accommodate better confounder adjustment by complex stratification and can also handle not only single drug–event combinations but also triplets/quadruplets/quintuplets. Covariates such as age, sex, and calendar year are taken into account by computing the expected value (expected number of reports having the event of interest multiplied by the overall proportion of reports having the drug of interest) separately for each covariate stratum and the results summed across strata. Additionally, drug–drug interactions may be identifiable.

Two methods that are based on a similar concept of disproportionality but adapted specifically for use in longitudinal healthcare data were developed by Schuemie [32•]. The first method, Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs (LEOPARD), screens off false positive associations that are likely due to protopathic bias. Protopathic bias occurs when a drug is prescribed to treat the early manifestations of the event of interest before the event itself is identified in the database. The second method, Longitudinal Gamma Poisson Shrinker (LGPS), is a variation of MGPS that utilizes the exposure information available in longitudinal databases, with the number of conditions occurring during non-exposed days employed in the estimation of the expected number of conditions during exposure. A similar method employing Bayesian shrinkage in the context of longitudinal data was proposed by Norén and colleagues. This method, called temporal pattern discovery, compares event rates in different time periods to filter out indications for treatment and provides a graphical statistical approach to depict temporal patterns that can lead to better interpretation within a clinical context [33•]. In addition, it provides for control of time-constant confounders through a self-controlled design while incorporating information on unexposed patients separately to account for systematic variability in event rates over time. Several recent publications provide a very good review on the suitability of disproportionality-based methods for drug safety surveillance [15••, 34•] as well as for occupational health surveillance [35].

Methods that Combine Data and Knowledge Mining

Siadaty and Knaus [36] developed an automated mining method that can systematically identify new and interesting patterns from a large pool of mined patterns by comparing the strength of patterns mined from a repository of clinical/healthcare data with the strength of equivalent patterns mined from a biomedical literature citation knowledgebase (i.e., PubMed). When such estimates of pattern strength do not match, a high 'surprise score' is assigned to the pattern, identifying such as potentially interesting. The surprise score thus represents the degree of novelty or interestingness of the mined pattern, shown simultaneously for the two databases graphically as a scatter plot. *P* values are computed for each surprise score, thus filtering out noise and attaching statistical significance. This method was used by the authors to determine the association between specific clinical diagnoses and laboratory findings. An important advantage of this dual mining method is that it obviates the need to develop a separate filtering method to incorporate evidence from the literature. In addition, associations that are very weak initially, but of interest and potential importance, may be captured.

Methods Designed for Biosurveillance

Methods for biosurveillance are specifically designed to detect outbreaks of disease, whether naturally occurring or caused by intentional release of bioterrorism agents. In the past, biosurveillance was implemented by manual reporting of suspicious and notifiable clinical and/or laboratory data from clinical practitioners, hospitals, and laboratories to public health officials. More recent systems have moved towards automated extraction and analysis of data routinely collected for other purposes. Patterns of illness in time and space are investigated and unusual clusters are flagged for further evaluation. The setting for these methods are mostly hospital- or community-based in scope and a wide variety of data sources are employed, including electronic medical records [37], emergency medical service dispatch systems, emergency department (ED) visit logs from hospital billing databases,

school records of absenteeism, laboratory and prescription records databases, as well as retail sales of over-the-counter healthcare products or groceries [38, 39]. In these systems, data may be acquired at the level of an individual visit (with either primary or multiple diagnoses), of a patient (in either a snapshot or longitudinal view), or of the population aggregate (e.g., ED visits per day for gastrointestinal complaints). Many of these methods can be adapted for use in monitoring drug (and particularly vaccine) safety.

In contrast to surveillance for drug-related signals, however, the principal underlying premise of biosurveillance systems is that the first signs of a (covert) biological warfare attack will be clusters of victims who change their behavior because they begin to become symptomatic [40]. Hence, such systems often use presenting complaints rather than final diagnosis to provide an early signal of unusual illnesses in a patient population. Some methods also employ clustering of diagnoses to define disease prodromes of interest in order to include all conditions that might conceivably be applied to a patient presenting with relatively early symptoms of an infectious or toxic syndrome. These characteristics are also seen in vaccine safety surveillance. More recent methodologies employ techniques specifically designed to detect spatial and temporal outbreaks in real time [41, 42], although other methods initially developed for other purposes have also been adapted such as quality assurance (i.e., process control chart-based methods, which are discussed in the next section), influenza excess mortality (cyclical regression), and pattern mining (e.g., WSARE: '*What's Strange About Recent Events*') [43].

The nature of the events or conditions of interest in biosurveillance systems makes the generation of alarms depend not so much on the plausibility of whether a biological agent (or something else, e.g., an existing medical condition) is responsible; any large-scale change in disease pattern is deemed a suspicious threat, unless proven otherwise. Unlike an adverse drug event that usually involves one patient at a time; a bioterrorist attack is anticipated to spread to other individuals in other areas in a relatively short period of time, and therefore temporal-spatial correlation is crucial. Furthermore, the exposure of interest is usually unknown or difficult to identify; hence, confounding has not been as crucial an issue as it is in drug safety monitoring, for example. Some methods, however, employ patient-level medical histories to take into account baseline risks and background incidence rates. In biosurveillance systems there is an emphasis on changes in disease patterns not only with respect to time, but also with respect to geographic location. The rapidity of detecting unusual geographic patterns is a paramount issue because of the small window of opportunity in the case of a real bioterrorism threat, which is often not the case in drug safety monitoring.

Chart-Based Methods

Chart-based methods were originally developed for use in statistical process control (SPC) and have long been used in quality control systems in the manufacturing industry. In the SPC context, these methods are used to monitor a process during its run and identify signals tipping off a process that is about to go 'out of control.' Because such methods can anticipate a glitch in the process, allowing the process to be discontinued or adjusted before the occurrence of the actual problem, these chart-based methods are often suited for prospective surveillance. Cumulative sum (CUSUM) and Shewhart comprise two of the most commonly used examples of these methods. CUSUM charts are based on sequential monitoring of cumulative events over a period of time. Several developments and adaptations exist such as the 'Observed minus Expected' (O-E) and the Log-likelihood CUSUM chart methods. CUSUM detects sudden changes in the average value of a parameter of interest and provides estimates of both magnitude and timing of change, with an alarm threshold value being defined a priori. An application of this method related to drug safety surveillance is a retrospective study showing that trends in hospitalization for myocardial infarction were highly concordant with the rise and fall of prescription rates for the cyclo-oxygenase (COX)-2-selective inhibitors rofecoxib and celecoxib [44]. Other applications in the healthcare setting include monitoring of clinical outcomes such as low Apgar scores [45], congenital malformations [46], occupational asthma [47], hospital-acquired infections [48], and mortality in intensive care [49]. Chart-based methods are all grounded on the statistical principle of process variability, and in industrial applications quality control can be translated as aiming to be 'on target with minimum variation.' In medicine and healthcare, however, there are much greater inherent variabilities, case mixes, or differences in risk than in most industrial processes [50]. As such, most of these methods are probably more valuable in detecting trends over time, rather than in detecting sentinel associations per se.

Algorithm-Based Methods

Algorithm-based methods comprise the more 'traditional' data mining methods and involve supervised learning (also called predictive modeling). Supervised learning, which is essentially a coming together of artificial intelligence and statistics, allows algorithms to automatically develop (predictive) models from observations within a system by relating a dependent variable with a set of independent variables, analogous to multiple regression analysis. Although initially widely used for financial applications (e.g., credit scoring, trading), supervised learning—and all its variations—has been a tool of choice to analyze the increasingly complex healthcare data generated by both routine care and

research [51, 52••]. Common applications in the biomedical domain include tumor detection and drug discovery, among other things. There are two types of supervised learning: classification, for categorical dependent variables where the data can be separated into specific 'classes'; and value prediction (also known as regression), for continuous dependent variables. Common classification algorithms include support vector machines (SVM), neural networks, Naïve Bayes classifier, decision trees, discriminant analysis, and k-nearest neighbors (kNN) [51, 52••]. Classification is suitable if the aim is to predict group membership of new records based on their characteristics (independent variables); this is done by identifying the most influential variable and using it to split the data into groups. The process is then iterated using the next most influential variable until all the data are fully characterized. Decision tree and rule induction are the more widely used classification techniques. An example would be the construction of a classification criterion (or decision rule) that discriminates between different groups of patients with and without adverse events based on age, sex, or co-morbid illness. Value prediction, on the other hand, uses both classification and regression to forecast an outcome in the future based on, say, patients' demographic or socioeconomic characteristics. It is important to remember, however, that, as in any data analysis of continuous outcomes, results of value prediction may be influenced by the presence of outliers in the data.

Clustering or database segmentation (unsupervised learning) employs an algorithm that splits a database based on dissimilarities between records [53]. A practical use of segmentation in the context of identifying drug-related sentinel associations would be grouping patients with similar clinical symptoms/manifestations or diagnoses to reduce a large sample of records to a smaller set of specific clusters without losing much information about the entire sample (either by hierarchical clustering or partition clustering). The inherent heterogeneity between clusters also allows hypothesis generation with respect to the nature of variation between subgroups. For example, if a database contained details of different gastrointestinal conditions (e.g., gastrointestinal bleeding) and medications (e.g., alendronate and other bisphosphonates), clustering analysis may have segregated patients with respect to gastrointestinal disease and found bisphosphonates to be one of the main factors in this group. It is then possible to explore whether there is an association, causal or otherwise, between gastrointestinal bleeding and bisphosphonates.

Link analysis comprises methods that identify associations or links between sets of data by using an 'if *x* then *y*' type rule and by assessing behavioral patterns sequencing/intervals of events (i.e., co-occurring events). Link analysis is aimed towards identifying and exploiting common indicators of an event in order to anticipate and offset any adverse consequences. Most applications of link analysis have been in monitoring criminal activity and in counterterrorism (http://csis.org/files/media/csis/pubs/040301_data_mining_report.pdf). The more extensively used types of link analysis involve associations, sequential patterns, and stratification. Associations are simply groups of 'instances' that frequently and consistently occur together. For example, if an outbreak of a rare disease occurs, it might be important to discover whether the affected patients come from the same locality, whether they share the same water source (if the disease is water-borne), or reside in the vicinity of a nuclear power plant (in the case of leukemia and other radiation-induced cancers). Because associations are the simplest link relationships and are often the easiest to discover, they are very useful in generating hypotheses for data mining. Sequential patterns that appear consistently can be used to formulate heuristics (rules of thumb), while stratification divides the potential association strata in such a way that enables formulating an answer to a question. Link analysis has been used in the healthcare domain to understand the relationship between nursing care and medical errors [54]. The same approach may be adapted in the context of pharmacovigilance; for example, link analysis may be used to identify relevant factors such as the effect of renal (or cardiac) dysfunction on the safety profile of antiretroviral drugs, or the influence of concomitant use of oral contraceptives on the safety profile of human papillomavirus vaccines.

Deviation (or anomaly) detection searches for outlier observations or events that deviate from the expected pattern within a particular dataset. Outlier detection frequently employs techniques that enable the visual recognition of patterns hidden in data (e.g., scatter plots or histograms, multidimensional graphs for multivariate data, and time series plots). Such an approach could be used to identify, for example, patients with pharmacologically unpredictable (i.e., idiosyncratic) reactions or those who exhibit clinical manifestations that are different from the usual toxidrome, which could be related to medication and may constitute a signal. Outlier detection has been applied in drug safety surveillance in the context of spontaneous reporting databases. Juhlin et al. showed that masking by outliers has substantial impact on the identification of specific ADRs using disproportionality analysis in VigiBase [55].

Data Mining: Caveats

While mining huge databases may provide a wealth of information, there remain caveats in the interpretation of signals derived from such data mining endeavors. Most of the data used are not primarily intended for recording medication-related adverse events and potential

associations are inferred outside the actual patient–physician encounter that leads to suspicion of such adverse events. Data mining methods that filter out alternative explanations for these associations (by controlling for bias and confounding) attempt to simulate the causality assessment that is actually performed by reporting physicians. Debates and discussions on the merits and challenges of mining secondary use healthcare data are found in the literature, including how the type of database influences the structure and content of the data [2, 6]. Data in medical record databases are recorded in the course of clinical care and, hence, take a healthcare practitioner's view of what is going on with a patient. On the other hand, claims databases document information as a byproduct of fiscal transactions, and therefore provide an auditor's view of healthcare data, and coding of outcomes can be biased by differences (real or perceived) in reimbursement. Data derived from health management organizations or social security systems could be affected by a lack of incentive to record sufficient data to allow proper case classification. Drug use patterns derived from 'real-world' healthcare data are influenced by changes in clinical practice, including changes brought about by preferential prescribing and disease management guidelines, and may lead to underestimation of risks. Finally, before data mining is conducted for signal detection purposes—for whatever type of data source—decision makers need to anticipate the question of what happens if and when a signal is detected; having a robust process of signal management in place will minimize the unpleasant implications of spurious and unsubstantiated signals for public health.

Conclusions

Data mining techniques are used to transform overwhelming volumes of data through the discovery of relationships, patterns, or clustering of records based on similarities among variables in the data. Data mining approaches have been used for a long time in the financial industry, but now find extensive and important applications in the medical domain. The advantages of automated surveillance and quantitative detection of signals in pharmacovigilance are obvious, but there is a legitimate concern that such data mining on a grand scale may generate more signals than can be followed up effectively with currently available resources. The unpleasant implications of spurious and unsubstantiated signals for public health cannot be underestimated and minimizing false positive alarms should be an important target when performing data mining. The lack of a true 'gold standard' remains a challenge in the development and evaluation of data mining methods, particularly in the detection of drug (or vaccine)-related sentinel associations.

Compliance with Ethics Guidelines

Conflict of Interest P.M. Coloma and S. de Bie both declare no conflicts of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
• Of importance
•• Of major importance

1. Smyth P. Data mining: data analysis on a grand scale? Stat Methods Med Res. 2000;9:309–27.
2. Hauben M, Aronson JK. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. Drug Saf. 2009;32:99–110.
3. Ing GM, Olman CL, Oyd JR. Drug-induced (Thalidomide) malformations. Can Med Assoc J. 1962;87:1259–62.
4.•• Coloma PM, Trifiro G, Patadia V, Sturkenboom M. Postmarketing safety surveillance : where does signal detection using electronic healthcare records fit into the big picture? Drug Saf. 2013;36:183–97. *Provides an overview of international initiatives exploring data from electronic healthcare record databases for signal detection and also describes the signal detection methods used in both databases containing spontaneous reports of adverse drug events and databases containing routinely collected healthcare data*.
5. Garcia Rodriguez LA, Perez Gutthann S. Use of the UK general practice research database for pharmacoepidemiology. Br J Clin Pharmacol. 1998;45:419–25.
6. Hennessy S. Use of health care databases in pharmacoepidemiology. Basic Clin Pharmacol Toxicol. 2006;98:311–3.
7. Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects–advantages and disadvantages. Nat Clin Pract Rheumatol. 2007;3:725–32.
8. Edwards IR. Adverse drug reactions: finding the needle in the haystack. BMJ. 1997;315:500.
9. Lieu TA, Kulldorff M, Davis RL, Lewis EM, Weintraub E, et al. Real-time vaccine safety surveillance for the early detection of adverse events. Med Care. 2007;45:S89–95.
10. Sebille V, Bellissant E. Sequential methods and group sequential designs for comparative clinical trials. Fundam Clin Pharmacol. 2003;17:505–16.
11. Freedman LS, Spiegelhalter DJ. Comparison of Bayesian with group sequential methods for monitoring clinical trials. Control Clin Trials. 1989;10:357–67.
12. O'Neill RT. Biostatistical considerations in pharmacovigilance and pharmacoepidemiology: linking quantitative risk assessment in premarket licensure application safety data, post-market alert reports and formal epidemiological studies. Stat Med. 1998;17:1851–8. *discussion 1859–1862*.
13. Coloma PM, Trifiro G, Schuemie MJ, Gini R, Herings R, et al. Electronic healthcare databases for active drug safety surveillance:

is there enough leverage? Pharmacoepidemiol Drug Saf. 2012;21: 611–21.
14. Coloma PM, Avillach P, Salvo F, Schuemie MJ, Ferrajolo C, et al. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. Drug Saf. 2013;36:13–23.
15.•• Schuemie MJ, Coloma PM, Straatman H, Herings RM, Trifiro G, et al. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. Med Care. 2012;50:890–7. *Evaluates the relative performance of different methods for detecting drug-related sentinel associations ('signals') in databases containing electronic healthcare records*.
16.•• Liu M, McPeek Hinz ER, Matheny ME, Denny JC, Schildcrout JS, et al. Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. J Am Med Inform Assoc. 2013;20:420–6. *Correlates relationships between drugs and laboratory tests within an electronic medical records database by evaluating several signal detection methods traditionally used in databases of spontaneous reports of suspected ADRs*.
17.•• Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc. 2013;20:413–9. *Proposes and evaluates an approach to signal detection that explicitly combines data from electronic medical records and spontaneous ADR reports*.
18. Bate A, Lindquist M, Orre R, Edwards IR, Meyboom RH. Data-mining analyses of pharmacovigilance signals in relation to relevant comparison drugs. Eur J Clin Pharmacol. 2002;58:483–90.
19. Bate A, Edwards IR. Data mining techniques in pharmacovigilance. In: Hartzema AG, Tilson HH, Chan KA, editors. Pharmacoepidemiology and therapeutic risk management. Cincinnati: Harvey Whitney; 2008. p. 239–72.
20. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Methods in health service research. An introduction to bayesian methods in health technology assessment. BMJ. 1999;319:508–12.
21. Kadane JB. Bayesian methods for health-related decision making. Stat Med. 2005;24:563–7.
22. Lucas P. Bayesian analysis, pattern analysis, and data mining in health care. Curr Opin Crit Care. 2004;10:399–403.
23. Lewis RJ, Wears RL. An introduction to the Bayesian analysis of clinical trials. Ann Emerg Med. 1993;22:1328–36.
24. Resnic FS, Zou KH, Do DV, Apostolakis G, Ohno-Machado L. Exploration of a bayesian updating methodology to monitor the safety of interventional cardiovascular procedures. Med Decis Making. 2004;24:399–407.
25. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. Biometrics. 2004;60:418–26.
26. Lindquist M. VigiBase, the WHO global ICSR database system: basic facts. Drug Inf J. 2008;42:409–19.
27. Dumouchel W. Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System. Am Stat. 1999;53:177–90.
28. Almenoff JS, DuMouchel W, Kindman LA, 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. 2003;12:517–21.
29. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10:483–6.
30. Egberts AC, Meyboom RH, van Puijenbroek EP. Use of measures of disproportionality in pharmacovigilance: three Dutch examples. Drug Saf. 2002;25:453–8.
31. Heeley E, Wilton LV, Shakir SA. Automated signal generation in prescription-event monitoring. Drug Saf. 2002;25:423–32.
32.• Schuemie MJ. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. Pharmacoepidemiol Drug Saf. 2011;20:292–9. *Presents a sequential set of novel methods for detecting and filtering drug safety signals in electronic healthcare record databases*.
33.• Norén GN, Hopstadius J, Bate A, Star K, Edwards IR. Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Disc. 2010;20:361–87. *Introduces a pattern discovery methodology that summarizes graphically the temporal association between the prescription of a drug and the occurrence of a clinical outcome*.
34.• Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res. 2013;22:39–56. *Explores the application of commonly used disproportionality methods to simulated and real electronic healthcare data*.
35. Bonneterre V, Bicout DJ, de Gaudemaris R. Application of pharmacovigilance methods in occupational health surveillance: comparison of seven disproportionality metrics. Saf Health Work. 2012;3:92–100.
36. Siadaty MS, Knaus WA. Locating previously unknown patterns in data-mining results: a dual data- and knowledge-mining method. BMC Med Inform Decis Making. 2006;6:13.
37. Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health. 2001;1:9.
38. Lober WB, Karras BT, Wagner MM, Overhage JM, Davidson AJ, et al. Roundtable on bioterrorism detection: information system-based surveillance. J Am Med Inform Assoc. 2002;9:105–15.
39. Bravata DM, McDonald KM, Smith WM, Rydzak C, Szeto H, et al. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. Ann Intern Med. 2004;140:910–22.
40. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc. 2004;11: 141–50.
41. Tsui FC, Espino JU, Dato VM, Gesteland PH, Hutman J, et al. Technical description of RODS: a real-time public health surveillance system. J Am Med Inform Assoc. 2003;10:399–408.
42. Reis BY, Kirby C, Hadden LE, Olson K, McMurry AJ, et al. AEGIS: a robust and scalable real-time public health surveillance system. J Am Med Inform Assoc. 2007;14:581–8.
43. Mostashari F, Hartman J. Syndromic surveillance: a local perspective. J Urban Health. 2003;80:i1–7.
44. Brownstein JS, Sordo M, Kohane IS, Mandl KD. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. PLoS One. 2007;2:e840.
45. Sibanda T, Sibanda N. The CUSUM chart method as a tool for continuous monitoring of clinical outcomes using routinely collected data. BMC Med Res Methodol. 2007;7:46.
46. Babcock GD, Talbot TO, Rogerson PA, Forand SP. Use of CUSUM and Shewhart charts to monitor regional trends of birth defect reports in New York State. Birth Defects Res A Clin Mol Teratol. 2005;73:669–78.
47. Hayati F, Maghsoodloo S, Devivo MJ, Carnahan BJ. Control chart for monitoring occupational asthma. J Saf Res. 2006;37:17–26.
48. Morton AP, Whitby M, McLaws ML, Dobson A, McElwain S, et al. The application of statistical process control charts to the detection and monitoring of hospital-acquired infections. J Qual Clin Pract. 2001;21:112–7.
49. Koetsier A, de Keizer NF, de Jonge E, Cook DA, Peek N. Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: a simulation study. Crit Care Med. 2012;40:1799–807.

50. Noyez L. Control charts, Cusum techniques and funnel plots. A review of methods for monitoring performance in healthcare. Interact Cardiovasc Thorac Surg. 2009;9:494–9.
51. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, et al. Machine learning in bioinformatics. Brief Bioinform. 2006;7:86–112.
52.•• Inza I, Calvo B, Armananzas R, Bengoetxea E, Larranaga P, et al. Machine learning: an indispensable tool in bioinformatics. Methods Mol Biol. 2010;593:25–48. *Provides a basic taxonomy of machine learning algorithms and describes the characteristics of main data preprocessing, supervised classification, and clustering techniques*.
53. Wilson AM, Thabane L, Holbrook A. Application of data mining techniques in pharmacovigilance. Br J Clin Pharmacol. 2004;57:127–34.
54. Potter P, Wolf L, Boxerman S, Grayson D, Sledge J, et al. An analysis of nurses' cognitive work: a new perspective for understanding medical errors. In: Henriksen K, Battles JB, Marks ES, Lewin DI, editors. Advances in patient safety: from research to implementation (Volume 1: Research Findings). Rockville: Agency for Healthcare Research and Quality (US); 2005.
55. Juhlin K, Ye X, Star K, Noren GN. Outlier removal to uncover patterns in adverse drug reaction surveillance - a simple unmasking strategy. Pharmacoepidemiol Drug Saf. 2013;22:1119–29.