ELSEVIER

Decision support methods for the detection of adverse events in post-marketing data

M. Hauben[1,2,3,4,5] and A. Bate[5,6]

[1] Pfizer, New York, USA
[2] New York University School of Medicine, New York City, USA
[3] New York Medical College, Valhalla, New York, USA
[4] School of Pharmacy, University of Maryland, USA
[5] Department of Information Systems and Computing, Brunel University, London, UK
[6] The Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring (WHO-UMC), Uppsala, Sweden

Spontaneous reporting is a crucial component of post-marketing drug safety surveillance despite its significant limitations. The size and complexity of some spontaneous reporting system databases represent a challenge for drug safety professionals who traditionally have relied heavily on the scientific and clinical acumen of the prepared mind. Computer algorithms that calculate statistical measures of reporting frequency for huge numbers of drug-event combinations are increasingly used to support pharamcovigilance analysts screening large spontaneous reporting system databases. After an overview of pharmacovigilance and spontaneous reporting systems, we discuss the theory and application of contemporary computer algorithms in regular use, those under development, and the practical considerations involved in the implementation of computer algorithms within a comprehensive and holistic drug safety signal detection program.

Introduction

It is well accepted that safety information about medicinal products will sometimes only come to light after market approval of a drug [1]. Since the 1960s, surveillance systems have been in place to capture such adverse drug reactions (ADRs). ADR signal detection in post-marketing surveillance (PMS) has largely been based on astute observations and analysis of spontaneously reported suspected ADRs by expert clinical reviewers [2,3]. With increasingly large databases that strain the capacity of clinical reviewers, quantitative methods have been increasingly used [4–6]. Recent research has predominantly focused on methods for optimising the highlighting of single drug-single ADR combinations for clinical review, based solely on spontaneous reported data, although historically pharmacovigilance has used multiple methods and data streams, including screening for increases in reporting rates [7,8]. In addition, there are applications for screening of hospital data [9] and also other adverse event monitoring systems for signal detection in primary care

Corresponding author: Hauben, M. (Manfred.hauben@Pfizer.com)

MANFRED HAUBEN MD
Manfred Hauben MD, MPH is currently Senior Director, Risk Management Strategy at Pfizer Inc. and holds faculty positions in family and community medicine and pharmacology at New York Medical College, in the Division of Clinical Pharmacology, Department of Medicine at New York University School of Medicine, in the Department of Pharmaceutical Health Services Research at the University of Maryland school of pharmacy, and at the Department of Information Systems and Computing at Brunel University in West London. He is board certified in preventive medicine and public Health as well as in clinical pharmacology. He has 18 years of experience in drug safety, pharmacoepidemiology, and risk management and has published extensively on data mining and signal detection in pharmacovigilance. He is a member of the USFDA-PhRMA Safety Evaluation Tools (SET) Expert Working Group, the EMEA Eudravigilance expert Working Group, and is team leader for the methodology subgroup of the Council of International Organization of Medical Science (CIOMSVIII) working group on signal detection and management in pharmacovigilance.

ANDREW BATE PHD
Andrew Bate PhD a Masters degree in chemistry from Oxford University, a PhD in clinical pharmacology – his thesis was on the subject of data mining the WHO database – and is a Visiting Professor of Information Systems and Computing at Brunel University, London. He is a member of the CIOMS VIII working group on signal detection and management in pharmacovigilance, an editorial board member of the international journal '*Drug Safety*', an appointed expert adviser to the European Medicines Agency (EMEA), and has been employed at the WHO Collaborating Centre for International Drug Monitoring since 1997 as a Research Manager, and is responsible for research at the institute. He has coordinated the development of methods for the data mining of a 4 million record database of suspected side effects of drugs (spontaneous reports) and other data sets such as electronic patient records, including the use of a Bayesian Confidence Propagation Neural Network (BCPNN). This data-mining tool has been routinely used since 1998 for the early detection of possible new side effects of drugs and has produced internationally high profile findings.

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2008.12.012

[10,11]. Some preliminary research has also been done on highlighting of three way dependencies or interactions (e.g. drug–drug-event associations) representative of more complex safety phenomena [12–14]. In this article we focus on methods currently used predominantly in the analysis of spontaneous reports of possibly causal associations between a single drug and a single AE, but also discuss the methods that have been used for the detection of larger groups of related concepts including interactions and syndromes in these types of data. In addition to surveillance and screening of post-marketing data, formal epidemiological analysis using well-defined questions to illuminate causality and estimate magnitude of drug effects is a standard response element after a signal is detected. These formal analyses include cohort and case control studies and newer methods, such as case crossover studies and are performed on both established datasets and datasets created to address specific research questions (for more details please see Strom [15]). This review will be restricted to surveillance methods.

The application domain: pharmacovigilance

Pharmacovigilance (PhV), has been defined as: ''The science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem'' [16]. It has often been used synonymously with post-marketing surveillance (PMS) or drug safety monitoring. The historic equivalence of 'PhV' with 'PMS' relates to the fact that clinical trials in support of drug applications, with their necessary constraints on size, duration, and patient heterogeneity, cannot reliably capture the full range of ADRs observed in widespread clinical use. Therefore, ADRs that are rare, or occur only after prolonged latency, are often unknown at the time of initial approval. Just as the drug discovery process is continuous, with no rigid boundaries despite the classic segmentation used to depict drug development (e.g. phases I–IV), PhV is, however, becoming more holistic and integrative and commencing earlier in the drug development process.

Signal detection in PhV

The 'front line' of pharmacovigilance consists of 'signal' detection—the expeditious identification of early clues of potential ADRs that may be novel by virtue of their nature, severity and/or frequency.

There is considerable variation in the use of the term 'signal' [17]. The World Health Organisation (WHO) definition, and the most widely cited definition, is ''reported information on a possible causal relationship between an adverse event and drug, the relationship being unknown or incompletely documented previously'' [18].

When a credible signal of a new adverse event is detected, it triggers an evaluation that usually begins with a detailed review of individual case reports of the association that are submitted to spontaneous reporting system (SRS) databases as described below. The initial investigation of a signal may determine that a causal relationship is sufficiently likely to warrant some action (e.g. labeling amendment), that the relationship is most likely non-causal, or that it is unclear but continued monitoring and/or further studies are indicated. Depending on the nature of the event, a formal study (e.g. epidemiological analysis or large simple clinical trial) may be triggered by the detection of a credible signal. Often, however, the action needed on the basis of a signal will be 'no action', other than ongoing follow-up of the signal.

The above scenarios illustrate that decision-making in the setting of residual uncertainty is inevitable in contemporary pharmacovigilance, from initial signal detection to final adjudication of whether an association is causal and the appropriate action. The downstream investigations that are triggered by the detection of a signal involve scientific disciplines and analytical processes that are subjects in their own right, and are beyond the scope of this exposition but suffice it to say that safety reviewers must constantly weave clinical, epidemiological, quantitative, and molecular science and logic at the level of individual cases and aggregate data. In this review we focus on the front-end of the process: strategies for the initial identification of possible emerging safety issues.

As described below, there are computerized data-mining algorithms (DMAs) that calculate a number that reflects whether, and by how much, the frequency of a given drug-adverse event association exceeds a null or control value. Reporting frequency in excess of chance expectation is one of the multiple possible indicators of a previously unrecognized association with significance for patient safety. These numbers, however, viewed in a biological vacuum, should not be equated with a signal, as defined by the WHO (see above), and may not require a formal investigation, depending on the clinical context. We refer to these, therefore, as 'signals of disproportionate reporting' or SDRs [17] and stress that the elevation of an SDR to a credible signal is based on a cognitive clinical review process. In the pharmacovigilance literature, SDRs have also been defined as 'associations' [19].

The problem space of signal detection in PhV

To appreciate better the problem space of signal detection in PhV, we review its components:

- The sample space of ADRs
- The reporting mechanism for submitting ADR reports
- The ADR databases including the terminologies used to encode information
- Methodologies to interrogate the data

Our focus in this paper is component #4, specifically DMAs used to screen large safety databases. Knowledge of the first three elements will facilitate an understanding of the fourth.

The sample space of ADRs

Pharmacovigilance is unique among surveillance systems in the range and complexity of medical phenomena under surveillance. This applies to both clinical phenotype and quantitative frequency/risk of occurrence of ADRs. These factors influence the choice of surveillance methods.

With the increasing number of molecular targets and corresponding drugs, ADRs rival syphilis and miliary tuberculosis as exemplars of 'great imitators' in medicine, in terms of their extremely protean clinical presentations. Some of these clinical presentations challenge the traditional views of ADRs as consisting of allergic reactions, hepatitis, rashes and gastrointestinal disturbances. Kidney stones, biliary stones, pure red cell aplasia, thrombotic thrombocytopenic purpura/haemolytic uremic syndrome, many forms of vasculitis, pneumothorax, tendon rupture, myopia, pyloric stenosis, hiccups, hypothermia, non-cardiogenic pulmonary edema and cardiomyopathy are but a few examples. Some ADRs defy therapeutic/pharmacological expectations—for example, anaphylactic reactions to corticosteroids, which are used to

treat allergic reactions, or hypertensive reactions from drugs given to treat hypertension. The latter two ADRs are examples of 'paradoxical reactions' [20]. This underscores the importance of the prepared mind expecting the unexpected [21].

The quantitative frequency or incidence of ADRs ranges from very rare to common in treated and untreated populations. How rare or common the ADR is in treated versus untreated populations, determines the difficulty in differentiating ADRs from background illness, or the natural history/complications of the disease under treatment, especially given their myriad presentations. This influences the optimum methods for detection and/or evaluation [22].

The reporting mechanism for submitting ADR reports

Every country and/or geographic region (e.g. the European Union) has its own legal and regulatory framework governing spontaneous reporting of adverse drug reactions, but there are commonalities. Except for pharmaceutical companies that are legally bound to report suspected ADRs to health authorities, it is usually a voluntary activity by the source reporter (e.g. health care practitioner, patient). This is the basis for the term 'spontaneous reporting'. Importantly, the reporter does not need proof of causality-any suspicion, however tentative, along with an identifiable patient, drug and event, is sufficient for submitting a spontaneous report.

Since reporting is voluntary and anecdotal, differential influences that have nothing to do with actual causality or risk may result in certain suspected ADRs being preferentially observed, attributed/misattributed to drug and reported/not reported. The data elements in individual reports are also subject to considerable qualitative and quantitative deficits in the form of missing or incorrect information and duplicate reporting [23]. Finally, it is impossible to know which/how many ADRs were never reported and how many patients were exposed to the drug. Therefore, while SRS data can be used to quantify reporting, it cannot be used accurately to quantify the corresponding risk/incidence.

The ADR databases including the terminologies used to encode information

Two important characteristics of SRS databases are size and sparsity. Large trans-national drug monitoring centres, health authorities and pharmaceutical companies with large portfolios maintain continuously growing databases of suspected ADR reports, often numbering in the millions with a large annual inflow of reports. These databases are also sparse, meaning most potential drug-event combinations have never been reported and most that are reported consist of one or two reports. This is compounded by the hyper-granular structure of the adverse reaction dictionaries used to record adverse events terms, where very similar medical concepts may be fragmented across literally distinct dictionary terms [24].

For example, the World Health Organization Uppsala Monitoring Centre (UMC) database contains about 4 000 000 adverse event reports listing 720 000 drug-event combinations (DECs) of which 360 000 have only a single report, 106 000 have two reports, and 80% of events have fewer than 10 reports. With so little information on most DECs, differentiating signal from noise is challenging, both to the human eye and when applying computerized methods (described below) [25]. We now discuss the element that is the focal point of our paper: the methodologies routinely used to explore pharmacovigilance data.

Methodologies to interrogate the data

Reported ADRs may stand out and be selected as possible signals for various reasons, both clinical and quantitative. The clinical criteria and heuristics used in pharmacovigilance have been discussed in detail elsewhere [26–28].

We focus on ADRs that first come to attention only after accumulation of a crucial mass of cases. Determining this crucial mass is the key conundrum in signal detection and where quantitative approaches based on computer-based statistical calculations can help.

Contemporary computer algorithms in pharmacovigilance primarily perform what is commonly called 'disproportionality analysis'. Key to understanding this analysis is the 2×2 contingency table that classifies reports according to the presence/absence of the suspect drug of interest and the presence/absence of the event of interest in reports (for example phenytoin and cerebellar atrophy in Table 1). It summarizes the number of cases in the database that list phenytoin as suspect drug and cerebellar atrophy as the event, the number of reports listing phenytoin with other events, the number of reports of all other drugs listing cerebellar atrophy and the number of reports listing any other drug and any other event. The vast majority of reports will fall into the last category (cell D). Given the sparsity of SRS databases and a focus on rare adverse events in pharmacovigilance, cell A will have the fewest reports. A similar table can be constructed for every possible drug-event combination (drug-event combinations with no reports will have the cell count $A = 0$).

The distribution of the number of reports in the table is informative. Basic quantitative drug safety analysis of any sort often involves comparing the number of joint occurrences of drug and adverse drug event (ADE) of interest to the number expected, on the basis of the play of chance given the unconditional reporting frequency of drugs and events. The more the number of reports exceeds the number expected by chance, the more interesting and, possibly, worthy of further investigation. Basic calculations provide the number of reports that might reasonably be expected in

TABLE 1

Contingency table used in disproportionality analysis.

	Reports listing cerebellar atrophy	Reports for all other events	Total
Reports listing phenytoin	A	B	A + B
Reports for all other drugs	C	D	C + D
Total	A + C	B + D	A + B + C + D

TABLE 2

Common measures of association for 2 × 2 tables used in disproportionality analysis.

Measure of association	Formula	Probabilistic interpretation	Chance expectation
Relative reporting (RR)[1]	$\frac{A(A+B+C+D)}{(A+C)(A+B)}$	$\frac{Pr(ae\|\text{drug})}{Pr(ae)}$	1
Proportional reporting rate ratio (PRR)	$\frac{A(C+D)}{C(A+B)}$	$\frac{Pr(ae\|\text{drug})}{Pr(ae\|\sim\text{drug})}$	1
Reporting odds ratio (ROR)	$\frac{AD}{CB}$	$\frac{Pr(ae\|\text{drug})Pr(\sim ae\|\sim\text{drug})}{Pr(\sim ae\|\text{drug})Pr(ae\|\sim\text{drug})}$	1
Information component (IC)	$\text{Log}_2\frac{A(A+B+C+D)}{(A+C)(A+D)}$	$\frac{\text{Log}_2 Pr(ae\|\text{drug})}{Pr(ae)}$	0

1. The RR, when implemented within an empirical Bayesian framework, is known as the empirical Bayes geometric mean (EBGM). 2. The IC is a logarithmic RR metric that is implemented in a Bayesian framework.

each cell by chance, by which we mean the drug and event are independently distributed in the database, as well as a variety of metrics that measure how far the number exceeds chance expectation (Table 2). Of course, the notion of an expected number of reports is a useful conceptual prop, but, given the enormous limitations of SRS data, in reality it is difficult to say how many reports one should expect.

The number of reports exceeding that expected by chance, according to some arbitrary, even if rational, model, can never prove causality. A number of reports exceeding chance expectation, when considered in isolation, itself does not constitute a signal of suspected causality. There are numerous causes of signals of disproportional reporting (SDRs). First, there will be fluctuations in reporting that are essentially stochastic in nature and that are particularly problematic with sparsely reported associations - in other words unusually large (and small) observed-to-expected ratios (O/Es) may preferentially and transiently occur with associations with very low observed or expected counts. Additionally, the numerous important sources of systematic bias inherent to the data (i.e., the aforementioned confounders, biases, and reporting artifacts) may produce many SDRs. Contemporary data-mining methods cannot effectively address the latter systematic biases and can only mitigate the former stochastic sources of reporting variability. There are two basic approaches to controlling the stochastic variability. One is based on classical or frequentist notions of statistical unexpectedness and the other is based on Bayesian statistics. The dichotomisation of methodologies in our exposition should not be interpreted as a systematic comparison of frequentist versus Bayesian statistics or an endorsement of one approach over the other. This is because the intensity of research, development and implementation devoted to Bayesian methods in pharmacovigilance has dwarfed that devoted to enhanced or more complex implementations of frequentist approaches. Expressed a little differently, the set of commonly used methods, which form a core of our discussion, consist of some simple frequentist approaches and relatively more complex Bayesian approaches.

Classical or frequentist approaches

In this case, classical statistical notions of unexpectedness are used to help improve the signal-to-noise ratio. The common feature of these approaches is that they rely solely on information contained in the specific 2 × 2 table corresponding to the DEC of interest [6,29]. For example, when calculating a PRR for a given 2 × 2 table, the analyst may also specify additional threshold criteria of at least three reports and an associated χ^2 value of >3.85 (corresponding to a *p*-value of ≤0.05) or a *p*-value of the chosen disproportionality metric below a specified threshold. A limitation in such a binary approach (i.e. a separating threshold dividing ADRs into two classes—SDR+ versus SDR−, as discussed in further detail below under 'Practical considerations') is that even with very small observed counts, if the expected count is small, the statistics will fail to screen out such associations, some of which may be false positives. It remains to be seen if the χ^2 threshold can be titrated toward a desired level of sensitivity and specificity. A similar approach may be used with the *p*-value of each statistic. Alternatively, the standard error may be used to determine a credibility interval/lower limit (5% threshold) of the 90% confidence interval of the statistic. Asymptotic expressions for the standard error of all the common disproportionality metrics have been derived, some using the delta method. This reduces the number of associations presented to the analyst and mitigates stochastic fluctuations.

Of course, as described below in the section 'Practical considerations' there is no restriction against using higher thresholds of statistical unexpectedness, or using a ranking versus a binary classification approach. One form of ranking implementation described above is a bivariate plot of the disproportionality metric (e.g. the PRR and the ROR) versus the measure of statistical unexpectedness, which we illustrate in Figure 4 in that section. Analysts would then view the DECs in the upper right hand corner as most quantitatively interesting, since they are both very disproportionate and much less likely to represent stochastic fluctuations, with the least quantitatively interesting DECs in the lower left corner.

The Bayesian approach

Overview

The challenge of sparsity in spontaneous report datasets was one of the impetuses for the development of Bayesian methodologies since, in other arenas, Bayesian approaches have demonstrated superiority to frequentist approaches when the available information is extremely limited, There are currently two major Bayesian techniques used for data mining in pharmacovigilance, the Bayesian Confidence Propagation Neural Network (BCPNN) [4] and the multi-item Gamma-Poisson shrinker (MGPS) [14].

[1] IC and RR formulated in a Bayesian framework in BCPNN and M(GPS), respectively.

Bayesian methods, first adapted to drug safety signal detection by the WHO-UMC [4], may be viewed as a composite of two approaches to calculating an O/E ratio for each drug-event combination. One approach, based on the frequentist paradigm of statistics, views each DEC as representing a realization of a unique process and that the huge numbers of spontaneously reported DECs have unrelated sources of variability. An alternative is to view all of the reported drug-event combinations as realizations of the same random process and just take an overall or grand mean of these O/E ratios, based on marginal reporting frequencies/probabilities—basically a 'null 2 × 2 table'; neither view, nor even a composite view, is absolutely 'correct', hence their combination in a Bayesian approach. This approach appeals to our prior knowledge and plausible belief that given the sparsity of the data, the numerous reporting artifacts and confounders, most ADEs are not being reported unexpectedly frequently when stochastic fluctuations are taken into account, and do not have implications for public safety.

It follows that two basic conceptual steps characterize the Bayesian approaches. The first step is to calculate an 'expected' or null 2 × 2 O/E ratio or table based on overall reporting patterns (in contra-distinction to calculating just an 'expected count' in the classical or frequentist approaches). The expected 2 × 2 table actually encodes an expectation and range of plausible 2 × 2 tables. The second step involves constructing a weighted composite of the null and observed 2 × 2 tables. The null 2 × 2 table reflects our 'prior belief' or 'first guess' about the O/E for any ADE and, in effect, 'shrinks' or pulls high O/Es in individual observed 2 × 2 tables supported by minimal data toward this prior belief. This is the basis for the term 'Bayesian shrinkage'. This grand mean O/E is also referred to as the 'moderating prior', which in fact is not a single value, but reflects a range of plausible values, each with an associated probability leading to a probability distribution of possible O/Es with an associated expectation value. This amount of shrinkage is inversely related to the amount of data on the ADR of interest. In other words, for rarely reported ADRs, the null O/E is very influential on the weighted average, but as reports accumulate this influence diminishes until a crucial mass of cases is achieved and the effect of the moderating prior is then swamped by the local O/E [30]. As with the moderating prior, the composite 2 × 2 table is actually associated with an expectation and range of plausible 2 × 2 tables or O/E ratios. The distribution of plausible O/Es and their associated probabilities comprise what is known as the posterior distribution. Viewed a bit differently, the local 2 × 2 table, or the information on a specific drug-event combination, is being used to update the moderating before produce the posterior distribution of O/Es. This is known as Bayesian updating.

In a sparse dataset, unusually high or low observed/expected ratios will preferentially be reported in sparse areas of the database, for example, for combinations with low expected counts. Many, but not all, of these will represent stochastic fluctuations. By dampening these fluctuations, the signal-to-noise ratio may overall be increased, but possibly at the expense of missing signals. There is a lack of decision theoretic framework basis for quantifying the balance of costs and utilities for such a procedure and, given the heterogeneity of different users needs, we feel that there is unlikely to be such a universal framework. Consequently, various organizations make choices based on their individual experience and some organizations focus solely on the Bayesian method, some the frequentist methods and some use both in parallel. All organizations that use data mining, however, need also to have techniques for qualitative filtering of the data.

The principle Bayesian methods: BCPNN and MGPS

There are currently two major Bayesian methodologies based on 2 × 2 tables: The Bayesian Confidence Propagation Neural Network (BCPNN) and the Multi-item Gamma-Poisson-Shrinker. Fundamentally, the difference between the two approaches is the manner in which the moderating prior is derived. The BCPNN uses a Bayesian approach, while MGPS uses an empirical Bayesian approach. With the Bayesian BCPNN, constraints are placed upon the expected 2 × 2 table to achieve a desired null O/E = 1 and a desired level of shrinkage, whereas the empirical Bayesian MGPS uses the database to determine the null value (which may be one, or greater or less than one) and the corresponding strength of the shrinkage. Each calculates a Bayesian version of the RR or O/E, along with a range of plausible values.

BCPNN

Since 1998, a Bayesian confidence propagation neural network (BCPNN) has been used for screening the WHO ADR database as part of the routine signal detection process [4,13,31,32]. A measure of disproportionality, called the Information Component (IC) (see Table 2), and its credibility interval is calculated for each drug adverse reaction combination in the dataset. While initially the neural network solution was integral to the approach, as it was used to calculate IC values, IC analysis no longer requires a neural net solution and the BCPNN is now exclusively used for more complex pattern recognition. The IC is defined (Table 2) for a specific drug adverse reaction combination as [32,33]:

$$IC = \log_2\left[\frac{\text{Observed count}}{\text{Expected count}}\right]$$

To calculate the IC within Bayesian framework, the BCPNN effectively constructs a null 2 × 2 table for each possible ADR by simultaneously constraining the count in cell 'a' to be 0.5 with all cell counts conforming to the marginal relative frequency expectations of drug and event counts (O/E = 1, IC = $\log_2$O/E = 0). This is accomplished by specifying the hyperparameters of a Dirichlet distribution. The constraint on cell 'a' effectively determines the strength of the shrinkage since it is influential on the shape or variance of the moderating Dirichlet prior.

Recent work [34,35] shows that the mean of the IC is well approximated by the following simple and computationally expedient metric:

$$IC = \log_2\left[\frac{\text{Observed count} + 1/2}{\text{Expected count} + 1/2}\right]$$

Thus, it amounts to an extra batch of data consisting of 0.5 reports for which the drug and event are independent. While the constraint on cell count 'a' of 0.5 is titrated to achieve a desired level of shrinkage in the WHO database, other databases might justify different values. Drug-ADR pairs with positive values for the lower 95% confidence limits for the IC (IC_{025}) are highlighted for clinical review. The measure has been shown to be effective in predicting future listing in the literature [36].

MGPS

The empirical Bayesian MGPS uses the existing data to determine the null 2 × 2 table and, consequently, the amount of shrinkage. This amounts to borrowing information, from all possible 2 × 2 tables to determine the prior probability distribution of O/Es (that collectively represents the null O/E as a random variable with an expectation and variance), and then forming a weighted composite of the null O/E and the 'local' O/E of the individual 2 × 2 tables. The underlying rationale is that 2 × 2 tables with very high/very low O/Es that represent stochastic fluctuations may be occur preferentially among sparsely reported associations so stochastic fluctuations in opposite directions may cancel when the tables are pooled, while effectively increasing the sample size at the same time. Expressed a little differently, the observed 2 × 2 tables are viewed as realizations of an underlying population of 2 × 2 tables, distributed according to mixture of two gamma distributions.

As the data are used to determine the null 2 × 2 table, rather than using a prior belief to determine the null table, the null 2 × 2 table may have a mean O/E that is different from one, which, in turn, determines the point toward which shrinkage occurs, the extent of the shrinkage determined by the spread of the O/E values in the dataset. This then has the property that the point to which shrinkage occurs and the extent of shrinkage will vary from dataset to dataset—so that the extent of shrinkage will be applicable to the dataset of interest. Clearly, however, this comes at a cost of loss of transparency, as this variable shrinkage will vary both between datasets, but also within the database over time. Additionally if the dataset has unexpected properties, leading to a skewed or very asymmetric prior probability distribution, this may influence the shrinkage significantly and will not necessarily be transparent. Another potential drawback is that such an approach may be computationally intensive, resulting in relatively long run times that restrict the ability to use results for exploring new hypotheses [40] in an interactive manner by teams of pharmacovigilance analysts.

The Bayesian disproportionality metrics provided by MGPS are the EBGM, which is the corresponding empirical Bayesian implementation of the RR, and the EB05, which is the 5th percentile of the posterior distribution of plausible RRs.

Examples of data-mining outputs

Below we provide a few graphical examples of the application of IC analyses to the WHO database to both illustrate key principles and familiarize readers with the actual data-mining outputs that are available to analysts.

Figure 1 illustrates the method with the classic historical example of cough and captopril. The association of an antihypertensive drug with a common medical event that certainly does not meet the criteria for a DME, represents one scenario that might challenge early detection without a quantitative screening strategy in place. The figure shows the cumulative evolution of the IC over the life of the database. Initially the IC is 0—the IC value reflecting a prior assumption of independence between drug and AE in the absence of data. The wide confidence intervals reflect the volatility of the IC value to increasing data accumulation. The IC drops due to reporting of the drug and AE—but not the combination. The IC then drops as reporting of the AE and drug occurs, but not concomitantly; therefore the observed count remains at 0, but the expected count increases. When the first observed case of captopril implicated coughing, the IC increases, but the shrinkage means that the IC does not increase as high as with a frequentist estimate and, similarly, the still-wide confidence intervals reflect the still-sparse data on this combination. As more cases of suspected captopril induced coughing are reported, the IC increases to a value of 4 and the confidence intervals shrink. IC values on WHO data are now routinely highlighted for clinical review when the IC_{025} becomes positive.

A more contemporary example is the association between the antiepileptic topiramate and glaucoma, shown in Figure 2. For this combination the IC_{025} became newly positive in the second

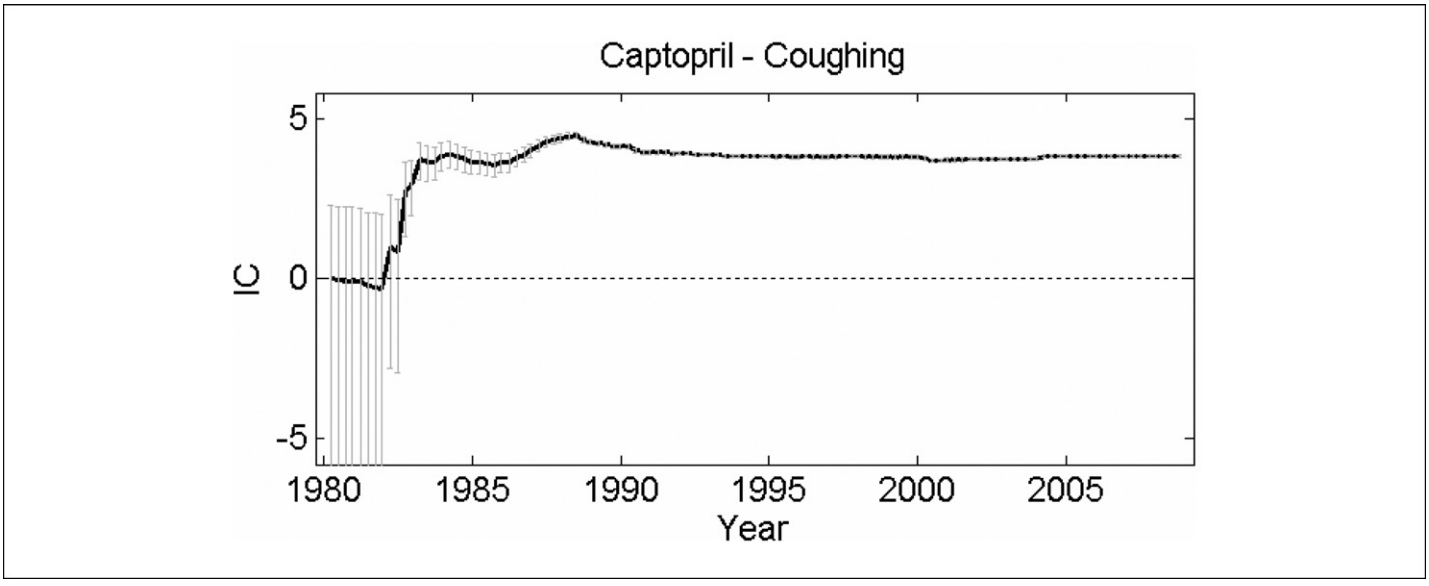

FIGURE 1

Information Component Time Scan of the Association of captopril and Cough in the WHO ADR database. The graph shows how the IC value with upper and lower 95% confidence level has changed as the WHO database has evolved from 1980 onwards. IC values are plotted based on cumulative data.

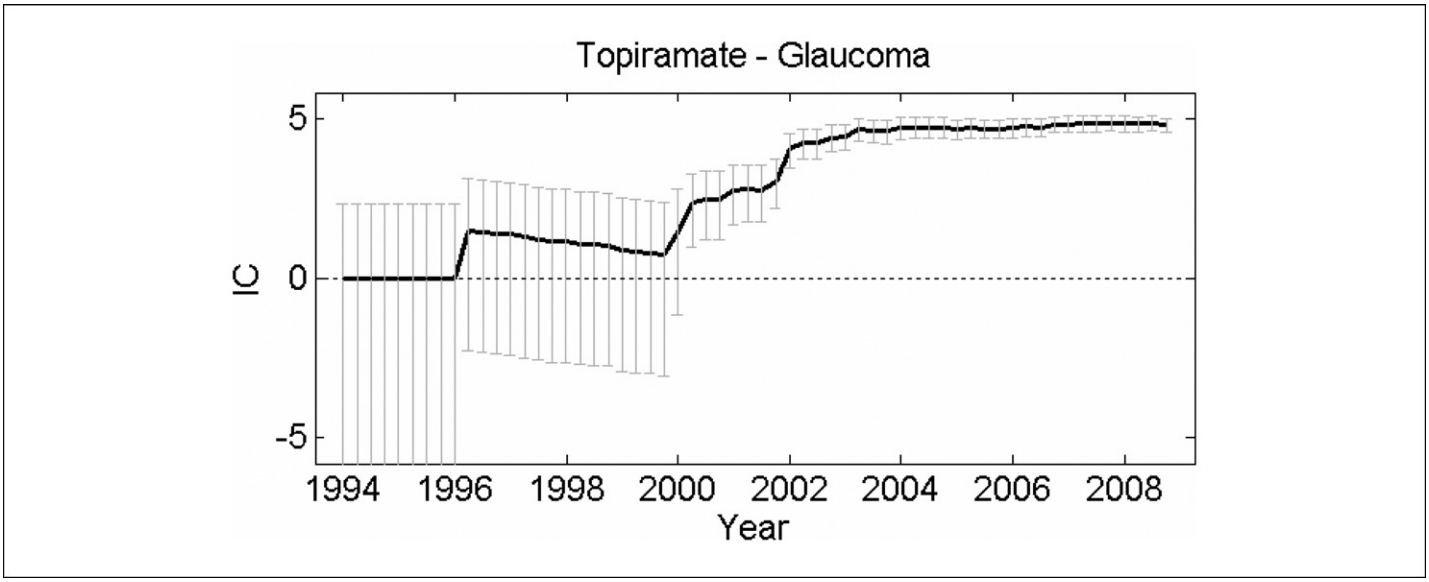

FIGURE 2

The change in IC value over time for the combination of antiepilepic topiramate and glaucoma in WHO data base.

quarter of 2000. This stimulated clinical review of the case series, then in April 2001 the signal was disseminated to the national centres that make up the WHO Programme for International Drug Monitoring. Knowledge accumulated on the signal, the first literature case report was published in July of the same year, and the association became established in the autumn of that same year, as exemplified by the FDA issuing a 'Dear Healthcare Professional letter' in October 2001.

We draw on another classic historical episode in pharmacovigilance to show how an analyst may use these techniques to explore the drug selectivity of a given adverse event within a pharmacological/therapeutic class (see Figure 3). Practolol was a beta-blocker that was introduced into clinical practice in 1964 in the UK. Initial reports of practolol were characterized by the apparently non-serious nature of the adverse events. After seven years in clinical use, the first report of sclerosing peritonitis was submitted and cumulative experience indicated the long temporal latency that may be observed with this event that explains the reporting of cases long after the drug was withdrawn in 1975 [37]. Sclerosing peritonitis is an unusual clinical disorder, in which the peritoneal cavity is the site of a dense fibrotic reaction that can encase the small bowel, resulting in intestinal obstruction; again, another example of the wide variety of ADRs encountered in PhV. Complete encasement results in the so-called 'abdominal cocoon'.

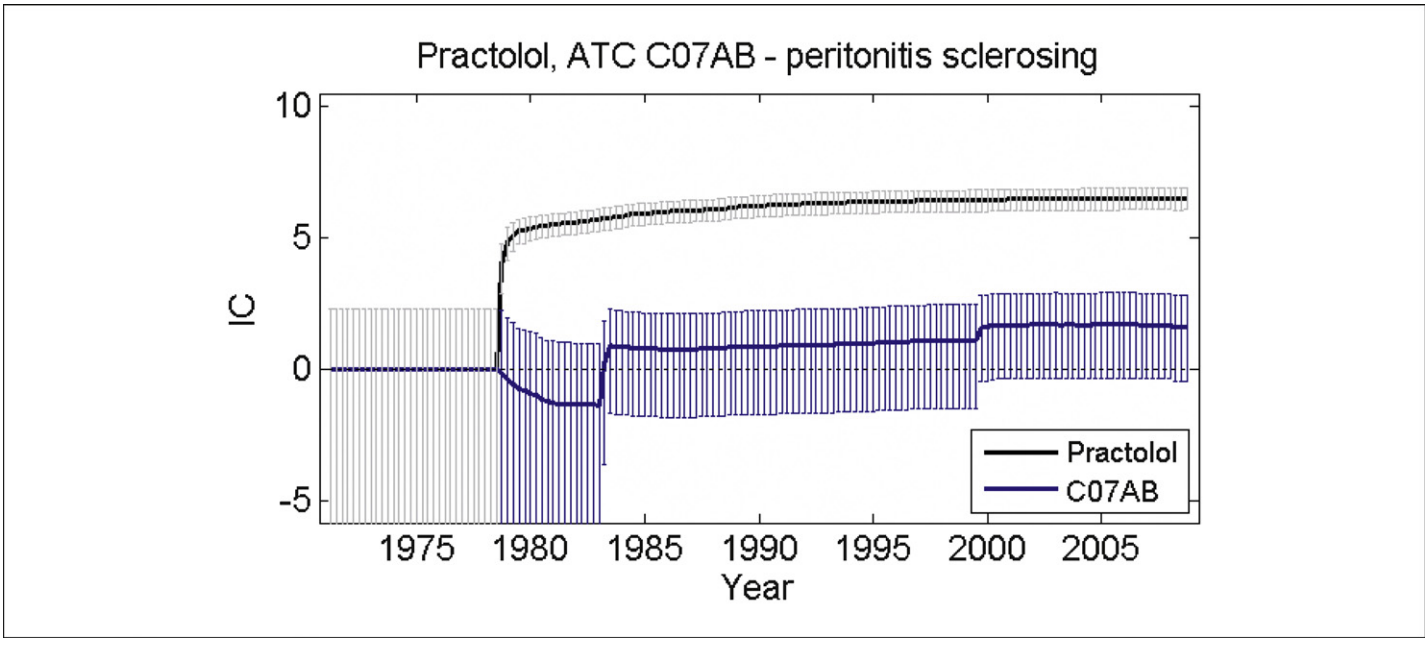

FIGURE 3

IC Time scan of practolol and selective beta-blockers as a group (WHO Anatomical Therapeutic Chemical (ATC) classification group C07AB) and sclerosing peritonitis in the WHO database.

While the drug-associated disorder is quite distinctive, other potential causes of sclerosing peritonitis include: continuous peritoneal dialysis; ventriculoperitoneal shunts and various other infectious, neoplastic and autoimmune disorders. It has also been reported in the absence of any identifiable antecedents [38,39].

Figure 3 shows the change of IC for practolol with the term sclerosing peritonitis. Superimposed is the group of selective beta-blockers (excluding practolol), defined by Anatomical Therapeutic Chemical (ATC) group, that is all beta blockers classified as C07AB. Clearly, the group is very different from the practolol graph and this is indeed a well-established drug-selective side effect. While spontaneous reports cannot be used to determine drug-specific associations, this type of comparison can clearly help in the generation of such an hypothesis, to be tested using other methods and datasets. The time scan is not a substitute for causality assessment and cannot be used to exclude causality in reports of sclerosing peritonitis involving other beta blockers, but can provide one piece of the puzzle that is ultimately assembled into a coherent hypothesis.

Method testing

The validation of these tools, or any signal detection procedure in pharmacovigilance for that matter, is not an easy task for numerous reasons [41]. The basic idea behind testing and validating these tools may also be encapsulated with a specific 2×2 table:

	True positive	**True negative**
Test positive	A (Correct prediction)	B (False positive)
Test negative	C (False Negative)	D (Correct prediction)

With any procedure, the ultimate downstream objective is to maximize the number of correct classifications (detecting new causal relationships/not highlighting relationships that are manifestly non-causal) and minimize the number of incorrect classifications (false positives and missing causal relationships).

One of the particularly contentious elements of validation exercises involves defining and identifying what constitutes a 'true positive' and 'true negative association'. For example, some have argued for focusing validation on performance in the detection of associations for which causality is guaranteed. We maintain that a more flexible approach that recognizes the importance of detecting associations that are possibly or probably real, even if not guaranteed with metaphysical certitude, is appropriate for real-world pharmacovigilance. This is because decisions must frequently be made in the setting of residual uncertainty and where the consequences of different errors are not identical.

For a signal detection system to be successful, it must highlight issues that will go on to be well established, while they are still emerging issues. It is not necessarily true that methods adept at focusing attention on now well-established drug safety issues, would have been able to highlight such issues when an apparent association was unknown or controversial; as the quantity and quality of pharmacovigilance data is very different for well-established side effects (particularly those that are publicized extensively). This dependence on time adds to the challenge of assessing the usefulness of the tools. Another problem is attempting to determine the number of true negatives, that is: things not highlighted by the method and considered true negatives, as many such issues may not even be reported! Nevertheless, several evaluation studies have been performed focusing on four specific testing elements:

1. Specific examples either shown retrospectively or prospectively of now well-established issues that could have been, or were highlighted early with DMAs, for example [4,13].
2. Assessment of concordance of the measures
3. Systematic retrospective testing of combinations to estimate the predictive value of DMAs by comparison to some external reference material (e.g. [36,42]) and finally,
4. Testing on theoretical test sets constructed specifically for the evaluation tasks (e.g. [43,44]). There are multiple nuances and sources of variability in data-mining procedures, outputs and performance assessment [45], some of which have received only limited attention in the data-mining literature [46].

The proper role of data mining, whichever software is selected, is within a comprehensive suite utilizing multiple strategies, tools and data streams, and how, expeditiously, to triage and evaluate signals originating from any source. The reality is that judicious implementation (e.g. titrating thresholds of disproportionality, statistical unexpectedness and/or minimum case counts), based on the level of sensitivity and specificity appropriate for the task at hand, it is possible to achieve comparable performance with any method, particularly if they are being used as binary classifiers.

Furthermore, for purposes of exploratory data analysis of this sort other performance metrics are valid considerations, such as computational burden [40]. Some approaches, such as MGPS, are computationally intensive. Some have questioned the added value of such intensive additional computational steps. Simple Bayesian approaches [35] or enhancements to frequentist techniques [47] have been suggested as useful alternatives Computational expediency may also present advantages in real-world pharmacovigilance scenarios [40].

Practical considerations

DMAs are important additions to the pharmacovigilance toolbox. However with DMAs that have an extensive mathematical veneer, it is especially easy to become desensitized to the rate-limiting effects of SRS data. The reality is that while these tools have enhanced the signal detection activities of a broad range of organizations, and, therefore, have legitimate indications, they also have side effects that need to be recognized, such as the generation of findings that will often divert resources investigating associations that prove to be spurious, and the fact that they may miss relevant associations, absolutely or relatively in terms of timing relative to conventional methods [31,36,48–50].

Deploying a DMA requires the analyst to make a variety of selections of various degrees of arbitrariness from a large space of available choices that define the configuration of an individual data-mining analysis. Some of these choices influence the numerical outputs and others influence the interpretation and/or response to a given set of outputs. We will discuss two of these choices to give the reader a taste of some of the nuances involved in real-world data mining in pharmacovigilance. One is whether the DMA is used as a binary versus ranking classifier, which we discuss now. Another is the issue of whether the analysis should include covariate stratification methods. The latter deals with the

fundamental issues of confounding and effect modification, discussed in detail below, under the heading: 'The need for more complex methods'.

Measures of disproportionality and/or statistical unexpectedness can be used as a thresholding tool to separate combinations into two groups: those requiring further consideration (i.e. those combinations exceeding specified threshold(s), and those that do not (e.g. those combinations at or below threshold(s)). Also, the values can be used to rank the combinations: the principle being that, all other factors being equal, combinations at the top of the list, or furthest from the origin in a 2-D plane of disproportionality, and statistical unexpectedness, are more likely to represent emerging signals and that review should start at the top of the list and work down.

Figure 4 provides an example received from the Swedish Medical Products Agency (MPA), which has used PRRs for signal detection.

The two approaches are not mutually exclusive and one can define a specific threshold, but use numerical ranking to triage associations exceeding threshold(s). In reality, however, all other factors are rarely, or never, equal in the complex domain of pharmacovigilance, and triage decisions are typically cognitive processes that blend the aforementioned numerical information with scientific knowledge and judgment. The limitations of spontaneous reports mean that caution is needed not to place inappropriate focus on the ranking order, but instead see it, as with thresholds, as one of multiple pragmatic approaches to focus on clinical review on issues most likely to represent emerging drug safety issues.

While thresholds have been proposed for each DMA, these are dataset-specific and have been chosen on the basis of empirical testing and some notion of a target range of sensitivity and specificity, which can be highly situation-dependent. Similarly while there seems to be agreement that this process of ranking works there is no, or very limited, discussion in the literature of when a user having worked their way down a list, can disregard the remaining drug AE combinations in the safe assumption that emerging signals will not be missed.

Given the numerous nuances and limitations of datasets, methods, and performance assessment most use of data mining in PhV is done as part of a holistic approach to signal detection utilizing a comprehensive suite of methods and data streams, both clinical and quantitative. Figure 5, adapted from Lindquist [19] is an illustration of how one major drug safety organization, the WHO-UMC, utilizes data mining.

Note that triage steps are accommodated. Many organizations utilize additional triage criteria, which, while not standardized or validated, are based on sound public health and decision-making principles. Table 3 displays the triage criteria used at WHO [31]. The concept of such triage criteria was first delineated by Venulet, who referred to them as discerning parameters [51].

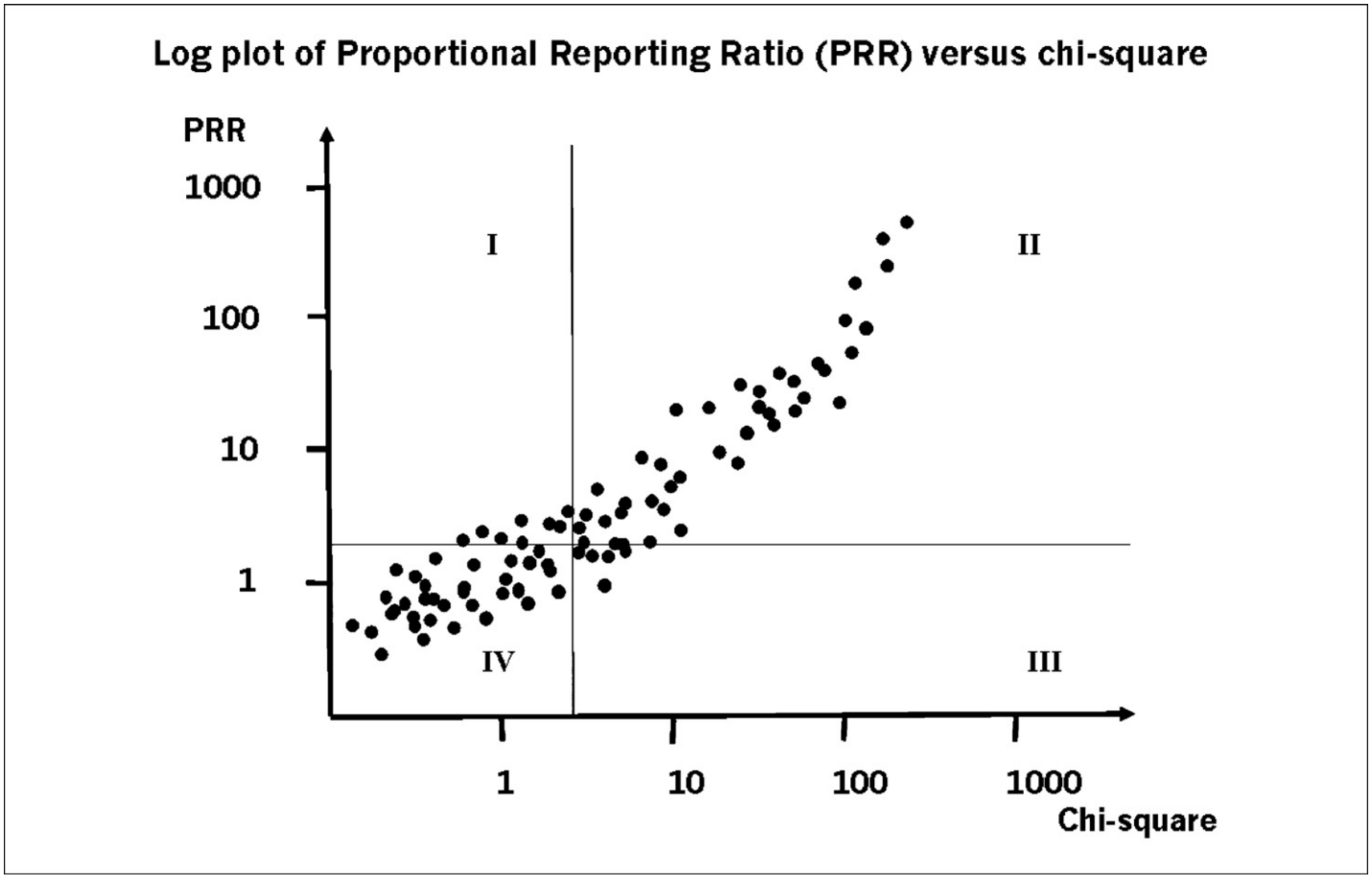

FIGURE 4

Bivariate plot of PRR versus χ^2.

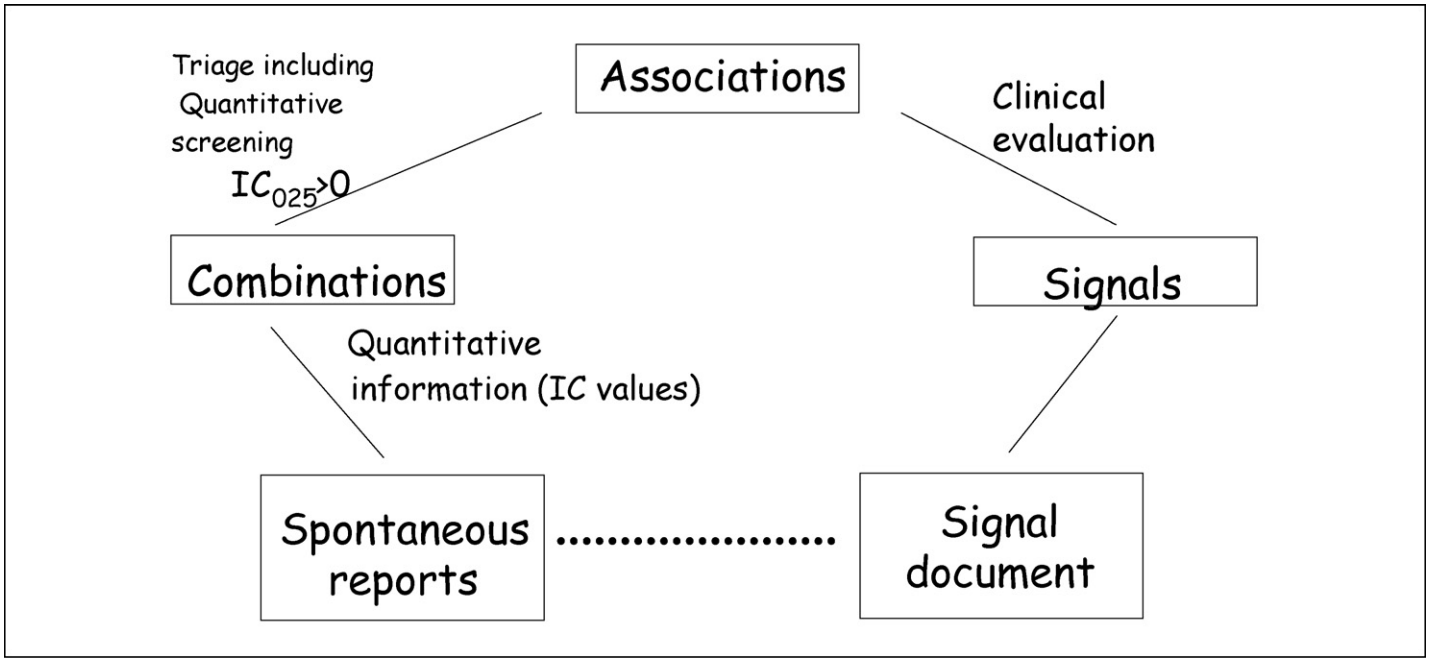

FIGURE 5

The Signal detection process at the Uppsala Monitoring Centre.

Importantly, quantitative approaches are typically discussed in the context of initial detection of signals but they may play an important role in the initial process of evaluating a signal initially detected by other methods applied to SRS data, sometimes called signal strengthening (or weakening) or signal refinement. This can be understood in terms of pre- and post-test probabilities. If, for example, the clinical information associated with a potential signal is very compelling and strongly suggestive of drug causation or an alternative aetiology, then the added information provided by the corresponding SDR is marginal at best. On the contrary, it is often the case that the clinical information is ambiguous and does not point strongly in favour of drug causation versus an alternative aetiology then the presence or absence of an SDR may be the first step pointing in one direction or another. The former situation is akin to a very high or very low pre-test probability, while the latter is akin to an intermediate pre-test probability. In the former case, a positive or genitive screening test adds little new information, while, in the latter, it does.

TABLE 3

Triage criteria used for screening the WHO database.

• Use of different selection criteria to filter out the combinations of likely greatest interest
• Predefined algorithms focusing on -Unknown/unexpected reaction -Disproportionality -Rapidly increasing disproportionality (IC-IC old > 2) -New drug -Serious reaction • WHO Critical Term, Outcome death -Reports involving many countries -Positive rechallenge -Special interest reaction • For example, agranulocytosis, Stevens Johnson syndrome

***Need for more complex methods for pattern recognition*

The range of complexity of the phenomena under surveillance, the structure of the data, and the intrinsic limitations of 2×2 table-based methods suggest that more complex techniques may improve our ability to identify relevant reporting associations in SRS data.

While 2-D associations account for the bulk of phenomena encountered in day-to-day pharmacovigilance, there are more complex higher-dimensional phenomena important for patient welfare. Associations may involve multiple interacting drugs (e.g. $drug_1$-$drug_2$ events) or drug-induced syndromes, in which a constellation of signs and symptoms (e.g. drug-$event_1$–$event_2$–$event_3$) exist. Not only are these phenomena important to detect, but for drug-induced syndromes, once identified, it may be useful to define the full range of clinical phenotypes and to distinguish distinct, but clinically overlapping, syndromes. For example, neuroleptic malignant syndrome and serotonin syndrome are distinct entities with overlapping clinical phenotypes, involving neuromuscular and autonomic features. Another example is drug-induced embryopathy. Even for 2-D associations, relationships between the drug, the event and additional covariates may ultimately contribute to a greater understanding of potential risk factors or high-risk subgroups.

The reduction of dimensionality to 2×2 tables, while practically useful, as shown above, necessarily results in a loss of information that potentially reduces applicability to more complex safety issues and further probing of simpler associations. A current research challenge is to exploit more fully the information on individual drug reporting, rather than merely lumping all drugs, other than the one under immediate consideration, into a single group 'other drugs' in a 2×2 table. The same applies to the AE terms reported. In other words, if you could 'unpack' cells B, C, and D in the 2×2 table you would be reminded that these single categories lump together huge numbers of drugs as 'other drugs'

and numerous events as 'other events'. Drugs can be very heterogenous from one another, as can events, and both will have their own relationships with each drugs and events in the dataset that may be important to understanding safety phenomena, such as for example drug–drug interactions and bystander effects, in which a drug may be associated in the 2×2 table because it is frequently co-prescribed with another drug known to have that side effect, and drug-induced syndromes.

To date, a few basic techniques have been applied to reveal more complex relationships. Covariate stratification is used to attempt to control the effects of confounding factors. Extensions of disproportionality analysis to higher dimensions and multiple logistic regression has been applied mostly to drug–drug interactions. Finally, unsupervised pattern recognition, has been applied to a limited extent to the detection and characterization of drug-induced syndromes. We discuss each in turn, but stress that the potential of more sophisticated methods to facilitate knowledge discovery in this domain does not eliminate the important role that clinical pharmacological knowledge continues to play in the detection and understanding of more complex safety phenomena, especially drug–drug interactions [52–54].

The impact of other variables on drug-event combinations

Among the other information that is invisible in a 2×2 table are data on variables that may be confounding factors (also known as 'lurking variables'), or effect modifiers, that may be the key to understanding even apparently 'simple' 2-D SDRs. Some such associations can be relatively easy to observe in certain circumstances, such as confounding by age, gender, year of report and so on. The number of potential confounding factors and effect modifiers, however, both recorded and unrecorded, presents difficulties in that they can result in spurious or masked associations [55]. Furthermore, the interplay of multiple variables can potentially reveal complex drug–drug interactions and drug-induced syndromes.

This is a simple example of a more general phenomenon. In general, particular patterns of association between observed and unobserved variables can lead to essentially arbitrary measures of association involving the observed variables. These measures can contradict the true unknown underlying causal model that generated the data. For example, in addition to drug–drug interaction detection, other co-reporting of pairs of drugs needs to be highlighted to prevent the aforementioned 'innocent bystander' being inappropriately associated with an apparent adverse drug reaction, in fact caused by a co-prescribed and reported drug [56]. Screening out for confounders can be done, but adjustment by too many variables can lead to the missing of signals in the application of data mining [55].

Confounding can, in principle, be relatively easily handled by stratification, although its practical implementation in PV data mining is far from intuitive and is fraught with difficulties [57]. For example, measures of disproportionality can be adjusted for the effect of a confounder using a Mantel-Haenszel adjustment to adjust the expected count for the impact of a third variable [5,30]. Clearly, such adjustments are not appropriate in the presence of effective modifiers [30] and alternative methods are needed. Screening for stratum-specific effects will also add value [13]. The large numbers of drugs in the database means that Mantel-Haenszel approaches are not well-suited to addressing confounding by drug (with the large number of strata) [30], logistic regression is a more appropriate approach that could be used to address confounding by drug, although there is relatively limited work on the application of logistic regression in post-marketing surveillance (as discussed below).

Higher-dimensional disproportionality analysis

A three way reporting disproportionality exists if the probability of a randomly selected report listing all three elements (e.g. drug1–drug2-event) is greater than might be expected from the general reporting of the three elements [13]. An 'expected' reporting frequency is calculated representing the number of reports expected given that the two drugs and the event are independently distributed in the database. In other words, if the probability of observing two specific drugs and an event in a randomly selected report is higher than the product of the probabilities of observing each one in a randomly selected report, one could say that this is an unusual three-way occurrence [13].

It is possible, however, to observe such an unusual occurrence because of strong two-way dependencies [14]. Therefore, a measure of disproportionality can also be defined with an expected count pair based on pair wise dependencies, such that the probability of a randomly selected report listing the most strongly dependent pairs among the former triplet (e.g. drug–drug, drug1-event or drug2-event). So some approaches calculate the $[O_{\text{three way}}/E_{\text{two-way associations}}]$ [14,35,58].

The limited success of measures of disproportionality has, at least partly, been due to the methods' focus on a multiplicative model for calculating an expected count; recent research has shown that an additive model can be more effective for spontaneous report screening [35,58].

Nevertheless, drug–drug interaction data mining in spontaneous reports may well be useful in signal detection. Spontaneous report screening has already been shown to have value in highlighting known drug–drug interactions that continue to be frequently reported [59], despite the warnings of severe established interactions, emphasising ongoing patient safety issues.

Multiple logistic regression

One potential approach to a fuller understanding of the complex interdependencies in SRS data is multiple logistic regression [12] that 'unpacks' the 2×2 table by controlling for co-medications. In effect, it creates a composite predictor variable of all potentially relevant covariates (e.g. all co-medications). The predictive weight of each individual covariate is determined by seeing how much of the variance is explained by all other covariates. The residual variance that remains unexplained by the other covariates, therefore, represents the independent contribution, or weight, of that element of the composite predictor variable. Until quite recently, the computational challenge presented by such regressions with upwards of 10 000 drugs as covariates was a significant barrier along with a significant potential for overfitting. Therefore, logistic regression application to spontaneous reports was restricted to specific questions [12] rather than large-scale screening. Now, however, several programs exist that can carry out linear and logistic regressions with millions of covariates, one method of

which is the BBR developed by Genkin *et al.* [60], which has been applied to spontaneous reports, preliminary results of which suggest shrinkage regression is promising as a surveillance tool, but is likely to be a complement to, rather than replacement of, the bivariate measures of disproportionality discussed earlier.

Unsupervised pattern recognition

Unsupervised pattern recognition methods may be applicable to the detection and delineation of complex drug-induced syndromes. It is a well-known problem in spontaneous ADR reporting that not all adverse drug reactions that are suspected will be reported [61,62]. Rarely, even when a case is reported, will all relevant data, such as the dosage administered, be recorded. Additionally, there are problems of either incorrect diagnosis, or certain symptoms not being recognized. Assuming that all the symptoms occurred, which is often not the case, the choice of term when recording the symptoms will exhibit inter-reporter variability and intra-reporter variability. This results in suspected ADR case reports where there is a large amount of missing data [23]. Follow-up reports may often give more case details, and further information on later symptoms and the outcome of the suspected ADR.

In terms of the adverse reaction terms listed, some may be incorrectly diagnosed, some incorrectly coded and some may be missed altogether. When looking for syndromes, the consequence is that few, if any, case reports will have all symptoms of a syndrome listed. It is clearly of interest, however, to detect overall clusters of related symptoms from this incomplete reporting. Similarly, there will be other large clusters of characteristics that are never all reported together involving many different types of variables.

It is impossible to define conditions that precisely describe the properties of such patterns of interest in general terms, for example the number of members within each pattern, the specific variables that will be involved, and even how many patterns of interest will exist in a particular dataset can all vary. While the descriptions of specific patterns of interest might allow them to be detected, being able to discover patterns of interest with as few preconceptions as possible, generating questions that might not otherwise have been considered is a key problem of interest.

Unsupervised learning using neural networks has been traditionally applied to find relationships in data, on the basis if learning from training data and test data, rather than providing decisions on how the neural network should learn or preconceptions on relations between variables. Such methods are also computationally efficient when searching for relations between many variables. Applications of neural networks include handwriting recognition [63], prediction of credit risk bankruptcy [64], ozone concentration [65] and even tornados [66]. A neural network method was, therefore, potentially useful for the problem of unsupervised pattern recognition in post-marketing surveillance.

The IC disproportionality method described above has been extended to find patterns amongst several variables, the IC representing weights in a recurrent Bayesian confidence propagation neural network (BCPNN). The recurrent BCPNN as a tool for unsupervised pattern recognition has been tested on theoretical data and shown effective in finding known syndromes in all haloperidol-reported data in the WHO database [67].

One example is clustering of the different adverse events listed on similar reports. This can represent several patterns of interest including symptoms that constitute a syndrome. As described above in an ADR database, the sparse nature of the data means that rarely, if ever, will all constituent symptoms of a syndrome be listed on any single case report. The individual ADR terms that make up a syndrome will not even necessarily show strong associations (positive scores of measure of disproportionality) with the drug causing the syndrome. The symptoms will occur sometimes with the drug in small groups of terms and have strong associations to other, more commonly, drug related symptoms in the syndrome. Therefore, searching for co-reporting of all symptoms

Cij	i	A1202	A0116	A0725	A0154	A0092	A0791	A0163	A0091	A0093	A0224	A0151	A0210	A0043	A0280	A0576	A0156	A0507
j	Ci/Cj	723	585	517	357	348	270	217	174	174	145	143	125	108	108	92	66	60
A1202	723	-	109	171	23	29	126	17	20	11	39	24	27	22	8	43	6	3
A0116	585	109	-	121	67	43	88	26	20	13	16	40	24	26	19	17	8	6
A0725	517	171	121	-	33	38	109	18	25	14	47	30	43	35	9	38	3	4
A0154	357	23	67	33	-	24	9	8	6	11	11	12	8	20	8	3	5	1
A0092	348	29	43	38	24	-	25	39	7	9	14	10	8	11	5	10	16	2
A0791	270	126	88	109	9	25	-	7	13	5	25	14	19	11	5	47	5	5
A0163	217	17	26	18	8	39	7	-	6	6	5	10	5	5	4	2	3	2
A0091	174	20	20	25	6	7	13	6	-	19	9	2	5	6	2	1	5	6
A0093	174	11	13	14	11	9	5	6	19	-	3	8	8	1	3	1	1	7
A0224	145	39	16	47	11	14	25	5	9	3	-	6	29	18	2	7	1	1
A0151	143	24	40	30	12	10	14	10	2	8	6	-	1	5	4	3	2	5
A0210	125	27	24	43	8	8	19	5	5	8	29	1	-	4	2	6	3	1
A0043	108	22	26	35	20	11	11	5	6	1	18	5	4	-	5	3	2	1
A0280	108	8	19	9	8	5	5	4	2	3	2	4	2	5	-	1	1	2
A0576	92	43	17	38	3	10	47	2	1	1	7	3	6	3	1	-	2	1
A0156	66	6	8	3	5	16	5	3	5	1	1	2	3	2	1	2	-	2
A0507	60	3	6	4	1	2	5	2	6	7	1	5	1	1	2	1	2	-

FIGURE 6

A cluster of ADR terms detected by an analysis of haloperidol data in the WHO database using a recurrent Bayesian confidence propagation neural network (rBCPNN). Column and row headings are codes representing specific ADR terms. Second row and column represent overall reporting of a specific ADR term for haloperidol. All other numbers represent the total reporting of the pair of ADR terms based on the column and row. White filled box represent a pair of ADR terms with a positive IC value; blue boxes negative IC values.

has limited use and more sophisticated methods are needed to find such relationships. A recurrent Bayesian Confidence Propagation Neural Network (BCPNN) has been applied to the WHO database of suspected ADRs [67]. This method is able to highlight clusters of ADR terms reported for specific drugs, such as the following cluster of ADR terms highlighted within reporting of haloperidol suspected ADRs (Figure 6).

In a feed-forward neural network, input data enter the network by setting the level of activation of nodes in an input layer, and then, *via* weighted connections, influence the activation levels of an output layer of nodes, to give results. The weight of the connections in a Bayesian Confidence Propagation Neural network is the IC value [4]. In this recurrent neural network, however, there is one network layer where the activation of each node is effected by the state of all the other nodes; the greater the weight the (IC value) between two nodes, the greater the influence the activity of each node has on the other. The activation levels of the individual nodes are initially set by an external stimulus and then, over time, the activation of each node changes, on the basis of the activation level of all other nodes. The states of all nodes are iteratively recalculated until the states of all nodes stabilize. Initially, the network has a certain ''energy'' associated with it, the network then searches for a lower energy level. When an energy minimum is found, the activity of all nodes stabilizes. Nodes that are active when the energy minimum is reached are the members of the output pattern.

The Column and rows in Figure 6 list the same ADR codes that refer to specific ADR terms. The numbers in the body of the table are the numbers of suspected haloperidol ADRs, where the pair of ADR terms in the row and column is co-listed. Each white square in the figure represents a pair of ADR terms between which there is a positive IC value, the blue squares a negative IC value.

The highlighted ADRs in this pattern were: NMS, hypertonia, fever, tremor, confusion, increased creatine phosphokinase, agitation, coma, convulsions, tachycardia, stupor, hypertension, increased sweating, dysphagia, leukocytosis, urinary incontinence and apnoea. Only 1 ADR term code A0116 (hypertonia) had a positive IC with all other terms in the pattern; also this list does not simply correspond to the most reported ADRs (nor highest IC value terms) for haloperidol. All ADRs are symptoms associated with NMS in standard literature sources, with the exception of dysphagia, for which published case reports exist of a possible link to NMS.

Clustering of similar case reports

Similar case reports should be considered together in case-by-case analysis for several reasons. Firstly, such reports might be linked to some underlying cause and, therefore, review of the separate reports might strengthen the probability of detecting a signal. Secondly, such reports might be duplicate copies of the same ADR incident and, if not actively considered as duplicates, might give a misleading strong impression of a signal. Duplicate detection is a well-established problem in spontaneous report screening [68], even more so since the advent of electronic reporting, whereby copies and variants of an original report can more easily occur. The only published algorithm on duplicate detection for screening for similar reports based on information, in addition to drugs and adverse events listed, is based on a hit miss model and is used for detecting similar cluster of case reports in the WHO database [23]. The algorithm is developed from the Copas and Hilton method proposed for record linkage [69]. In principle, an overall similarity score is established for every possible pair of case reports in the spontaneous report dataset. This overall score is the sum of the score calculated for each individual record field, including drugs listed, country of origin and age and gender of patient. Overall high scores are indicative of informatively similar spontaneous reports and trigger clinical review. As well as detecting duplicates, the method has also proved useful in determining other clusters of similar reports, such as series of reports received from the same dentist on the same day – which, while describing separate suspected incidents, clearly cannot be considered independent reports in the same way as two reports received in different time periods from different countries. This duplicate detection algorithm is now in routine use on the WHO database.

Assumed independence of all entered spontaneous reports is a current weakness of the routinely used DMAs in PV, and while the exact magnitude of duplicate detection is not known there is an acceptance that there are examples [70] that illustrate its potential to impact signal detection capability adversely. Consequently, such weighting of reports in the currently used DMAs may provide major performance improvement in signal detection.

Key research challenges in the use of computer algorithms in post-marketing surveillance

The vast majority of spontaneous reports have been coded using hierarchical terminologies. It is well accepted within the field that these hierarchical terminologies are not optimally constructed to support signal detection [24], whether qualitative or quantitative. Increasing efforts are being put into methodological development of the terminologies themselves and the methods themselves to improve signal detection. Two specific initiatives are more sophisticated semantic reasoning [71] and also tools based on a statistical framework for borrowing of information from semantically similar ADR terms [72].

While post-marketing signal detection predominantly focuses on the analysis of data collected after a drug is launched in the form of spontaneous reports, there is an increasing interest in analysing other healthcare data, such as the re-analysis of randomized clinical trial (RCT) data, particularly if pooled together, in order to glean more from the data when it is analyzed from another perspective; some examples of data mining of clinical data are included in references [73,74]. Methods for highlighting possible associations in RCTs could include the implementation of disproportionality measures as presented here, if possible adapted to consider the occurrence of adverse events in placebo groups; or completely different measures. Also the optimum balance of clinical and quantitative surveillance in clinical trials is still very much an open question as the quality and completeness of clinical trial data, relative to SRS databases, is much higher, facilitating clinical causality assessments at the individual case level, and because preserving the blinding may both complicate and improve the potential value of the prospective application of quantitative approaches in 'real-time'. Clearly screening of RCTs will not replace the need for signal detection on observational data, because of the carefully restricted drug use in RCTs. Some data mining of prescription databases has occurred [75]. Similarly, there

is increasing interest in the data mining of electronic patient records [33,76]. We anticipate post-marketing surveillance of adverse effects of drugs and research will increasingly involve combinations of the above datasets, as well as spontaneous reports.

Decision support methods in the identification of ADRs—a holistic approach

Here we need to discuss the partial role quantitative screening of spontaneous reports plays in the discovery of novel safety issues. There are two justifications for its focused use in a well-defined and restricted role. First is that most organizations use quantitative screening as a supplement rather than a substitute to qualitative signal detection strategies. For organizations with a comprehensive suite of pre-existing rigorous signal detection strategies, this use as a supplement obviously restricts the specific contribution of quantitative methods of signal detection. Secondly, just as important a justification is an understanding of pharmacovigilance processes as a continuum from exploratory analysis that generates ideas (i.e. signal detection or hypothesis generation), to confirmatory analysis of these ideas or hypotheses. The process is a continuum, and as such different points in the process share common or overlapping elements of supporting logic and data. But naturally different aspects need to be emphasized at one or the other end of the continuum. There is a relative and judicious premium on openness (sensitivity) to new ideas at the exploratory front-end of signal detection. However once we have a target signal and wish to expeditiously execute an analysis more akin to a confirmatory analysis, we place a higher premium on methods that are more specific, including hypothesis testing studies. It reflects Tukey's metaphor of exploratory data analysis as detective work and confirmatory analysis as the work of a judge or jury. The detective seeks patterns or clues and the data judge determines if these patterns and clues can be trusted [77]. Determining which issues are more likely to represent emerging ADRs using all the available evidence on spontaneous reports [28], before considering more detailed studies, is somewhere in the middle of this continuum, and can be seen as in some ways as adapting and applying the Austin-Bradford Hill criteria for adjudicating causality in epidemiology, to the sphere of signal detection in pharmacovigilance [78]. The effective use of quantitative screening algorithms is therefore just one important option in an overall process of 'good signal detection practice'. Effective strategies for signal assessment, strengthening, follow-up, and management are, while beyond the scope of this article, all equally important to the provision and maintenance of a trustworthy and valuable process.

Conclusions

There are now a variety of tools and computer algorithms to help screen large safety databases. Each can, in effect, compress the data into a high grade ore. Methods to improve the signal-to-noise ratio, whether by classical or Bayesian approaches, are far from perfect, primarily because of the nature of spontaneous reports, and remove signals with noise, necessitating their use as supplemental tools, rather than as stand-alone procedures. While the elegance of the Bayesian approaches is undeniable, their theoretical benefits have not been shown to give large practical benefits in screening of spontaneous reports for many organizations and may have some drawbacks. Judicious implementation of all the methods gives comparable results and far greater variation in performance is seen owing to heterogeneity in implementation choices, such as threshold selection/titration and the triage logic and procedures for investigation of signals. Some ADRs will be most easily detected by quantitative filters, some by qualitative filtering based on the nature of the information listed on the reports [79]. It is an open question how far sophisticated statistical tools can lead to substantial improved performance for single drug-signal AE screening, given the imperfect nature of the datasets they are implemented on, particularly given the increased cost of lack of transparency. Nevertheless, the majority of data mining in PV has neglected the screening of high-risk groups and other more complex patterns, from which many more useful findings could be expected, and we anticipate more sophisticated techniques will play a crucial role.

References

1 Edwards, I.R. and Aronson, J.K. (2000) Adverse drug reactions: definitions, diagnosis, and management. *Lancet* 356, 1255–1259
2 Meyboom, R.H. *et al.* (1997) Principles of signal detection in pharmacovigilance. *Drug Saf.* 16, 355–365
3 Hartmann, K. *et al.* (1999) Postmarketing safety information: how useful are spontaneous reports? *Pharmacoepidemiol. Drug Saf.* 8 (Suppl. 1), S65–S71
4 Bate, A. *et al.* (1998) A Bayesian neural network method for adverse drug reaction signal generation. *Eur. J. Clin. Pharmacol.* 54, 315–321
5 DuMouchel, W. (1999) Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am. Stat.* 53, 177–190
6 Evans, S.J. *et al.* (2001) Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol. Drug Saf.* 10, 483–486
7 Moussa, M.A. (1978) Statistical problems in monitoring adverse drug reactions. *Methods Inf. Med.* 17, 106–112
8 Norwood, P.K. and Sampson, A.R. (1988) A statistical methodology for postmarketing surveillance of adverse drug reaction reports. *Stat. Med.* 7, 1023–1030
9 Shapiro, S. (2000) Case control surveillance. In *Pharmacoepidemiology* (Strom, B.L., ed.), pp. 15–28, Churchill Livingstone
10 Coulter, D.M. (2000) The New Zealand intensive medicines monitoring programme in pro-active safety surveillance. *Pharmacoepidemiol. Drug Safety* 9, 273–280
11 Shakir, S. and Wilton, L. (2000) Drug safety research unit and pharmacoepidemiology. *Int. J. Pharm. Med.* 14, 1–2
12 van Puijenbroek, E.P. *et al.* (2000) Detecting drug–drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs. *Eur. J. Clin. Pharmacol.* 56, 733–738
13 Orre, R. *et al.* (2000) Bayesian neural networks with confidence estimations applied to data mining. *Comput. Stat. Data Anal.* 34, 473–493
14 DuMouchel, W. and Pregibon, D. (2001) Empirical Bayes screening for multi-item associations. In *Proceedings of the Seventh ACM SIGKDD International conference on Knowledge Discovery and Data Mining* pp. 67–76
15 Strom, B.L. (1994) When should one perform pharmacoepidemiology studies? In *Pharmacoepidemiology* (Strom, B.L., ed.), pp. 57–65, Churchill Livingstone
16 The importance of pharmacovigilance, 2002, WHO 48
17 Hauben, M. and Reich, L. (2005) Communication of findings in pharmacovigilance: use of the term 'signal' and the need for precision in its use. *Eur. J. Clin. Pharmacol.* 61, 479–480
18 Edwards, I.R. and Biriell, C. (1994) Harmonisation in pharmacovigilance. *Drug Saf.* 10, 93–102
19 Lindquist, M. *et al.* (1999) From association to alert—a revised approach to international signal analysis. *Pharmacoepidemiol. Drug Saf.* 8, S15–S25
20 Hauben, M. and Aronson, J.K. (2006) Paradoxical reactions: under-recognized adverse effects of drugs. *Drug Saf.* 29, 970

21 Trontell, A. (2004) Expecting the unexpected—drug safety, pharmacovigilance, and the prepared mind. *N. Engl. J. Med.* 351, 1385–1387
22 Aronson, J.K. and Ferner, R. (2005) Clarification of terminology in drug safety. *Drug Saf.* 28, 851–870
23 Noren, G.N. *et al.* (2007) Duplicate detection in adverse drug reaction surveillance. *Data Mining Knowl. Discov.* 14, 305–328
24 Brown, E.G. (2002) Effects of coding dictionary on signal generation: a consideration of use of MedDRA compared with WHO-ART. *Drug Saf.* 25, 445–452
25 Edwards, I.R. and Olsson, S. (2003) The WHO International Drug Monitoring Programme —vision and goals of the Uppsala Monitoring Centre. In *Side Effects of Drugs, Annual 26* (Aronson, J.K., ed.), pp. 548–557, Elsevier Science B.V.
26 Hauben, M. and Aronson, J.K. (2007) Gold standards in pharmacovigilance: the use of definitive anecdotal reports of adverse drug reactions as pure gold and high-grade ore. *Drug Saf.* 30, 645–655
27 Edwards, I.R. *et al.* (1990) Quality criteria for early signals of possible adverse drug reactions. *Lancet* 336, 156–158
28 Meyboom, R.H. *et al.* (2002) Signal selection and follow-up in pharmacovigilance. *Drug Saf.* 25, 459–465
29 van Puijenbroek, E.P. *et al.* (2002) A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol. Drug Saf.* 11, 3–10
30 Noren, G.N. *et al.* (2006) Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Stat. Med.* 25, 3740–3757
31 Stahl, M. *et al.* (2004) Introducing triage logic as a new strategy for the detection of signals in the WHO Drug Monitoring Database. *Pharmacoepidemiol. Drug Saf.* 13, 355–363
32 Bate, A. *et al.* (2002) A data mining approach for signal detection and analysis. *Drug Saf.* 25, 393–397
33 Bate, A. (2007) Bayesian confidence propagation neural network. *Drug Saf.* 30, 623–625
34 Norén, G.N. (2007) *Statistical Methods for Knowledge Discovery in Adverse Drug Reaction Surveillance*. Mathematical Statistics Stockholm University
35 Noren, G.N. *et al.* (2008) A statistical methodology for drug–drug interaction surveillance. *Stat. Med.* 27, 3057–3070
36 Lindquist, M. *et al.* (2000) A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. *Drug Saf.* 23, 533–542
37 Mann, R.D. (2006) An instructive example of a long-latency adverse drug reaction—sclerosing peritonitis due to practolol. *Pharmacoepidemiol. Drug Saf.* 16, 1211–1216
38 Xu, P. *et al.* (2007) Idiopathic sclerosing encapsulating peritonitis (or abdominal cocoon): a report of 5 cases. *World J. Gastroenterol.* 13, 3649–3651
39 Foo, K.T. *et al.* (1978) Unusual small intestinal obstruction in adolescent girls: the abdominal cocoon. *Br. J. Surg.* 65, 427–430
40 Hauben, M. *et al.* (2007) Data mining in pharmacovigilance: computational cost as a neglected performance parameter. *Int. J. Pharm. Med.* 21, 319–323
41 Hauben, M., and Bate, A. (2007) Data mining in drug safety: side effects of drugs essay. In *Side Effects of Drugs* (Aronson, J.K., ed.), pp. xxxiii-xlvi, Elsevier
42 Hauben, M. and Reich, L. (2004) Safety related drug-labelling changes: findings from two data mining algorithms. *Drug Saf.* 27, 735–744
43 Rolka, H. *et al.* (2005) Using simulation to assess the sensitivity and specificity of a signal detection tool for multidimensional public health surveillance data. *Stat. Med.* 24, 551–562
44 Roux, E. *et al.* (2005) Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. *IEEE Trans. Inf. Technol. Biomed.* 9, 518–527
45 Bate, A. (2003) *The Use of a Bayesian Confidence Propagation Neural Network in Pharmacovigilance*. Department of Pharmacology and Clinical Neuroscience Umeå University
46 Hauben, M. *et al.* (2007) Illusions of objectivity and a recommendation for reporting data mining results. *Eur. J. Clin. Pharmacol.* 63, 517–521
47 Davis, R.L. *et al.* (2005) Active surveillance of vaccine safety. A system to detect early signs of adverse events. *Epidemiology* 16, 336–341
48 Hauben, M. and Hochberg, A.M. (2008) The importance of reporting negative findings in data mining: the example of Exenatide and Pancreatitis. *Pharm. Med.* 22, 215–219
49 Hauben, M. and Reich, L. (2005) Potential utility of data-mining algorithms for early detection of potentially fatal/disabling adverse drug reactions: a retrospective evaluation. *J. Clin. Pharmacol.* 45, 378–384
50 Lehman, H.P. *et al.* (2007) An evaluation of computer-aided disproportionality analysis for post-marketing signal detection. *Clin. Pharmacol. Ther.* 82, 173–180
51 Venulet, J. (1988) Possible strategies for early recognition of potential drug safety problems. *Adverse Drug React. Acute Poisoning Rev.* 7, 39–47
52 Horn, J.R. and Hansten, P.D. (1993) Comment: pitfalls in reporting drug interactions. *Ann. Pharmacother.* 27, 1545–1546
53 Hauben, M. (2001) Comments on hypotension associated with intravenous haloperidol and imipenem. *J. Clin. Psychopharmacol.* 21, 345–347
54 Hauben, M. (2002) Comment: phenytoin/isradipine interaction causing severe neurologic toxicity. *Ann. Pharmacother., 2002* 36, 1974–1975
55 Hopstadius, J. *et al.* (2008) Impact of stratification on adverse drug reaction surveillance. *Drug Saf., 2008* 31, 1035–1048
56 Purcell, P. and Barty, S. (2002) Statistical techniques for signal generation: the Australian experience. *Drug Saf.* 25, 415–421
57 Bate, A. *et al.* (2003) Violation of homogeneity: a methodologic issue in the use of data mining tools—the authors' reply. *Drug Saf.* 26, 364–366
58 Thakrar, B.T. *et al.* (2007) Detecting signals of drug–drug interactions in a spontaneous reports database. *Br. J. Clin. Pharmacol.* 64, 489–495
59 Strandell, J. *et al.* (2008) Drug–drug interactions—a preventable patient safety issue? *Br. J. Clin. Pharmacol.* 65, 144–146
60 Genkin, A. *et al.* (2007) Large-scale Bayesian logistic regression for text categorization. *Technometrics* 49, 291–304
61 Begaud, B. *et al.* (2002) Rates of spontaneous reporting of adverse drug reactions in France. *JAMA* 288, 1588
62 Backstrom, M. *et al.* (2004) Under-reporting of serious adverse drug reactions in Sweden. *Pharmacoepidemiol. Drug Saf.* 13, 483–487
63 Oh, I.S. and Suen, C.Y. (2002) A class-modular feedforward neural network for handwriting recognition. *Pattern Recognit.* 35, 229–244
64 Atiya, A.F. (2001) Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Trans. Neural Networks* 12, 929–935
65 Yi, J.S. and Prybutok, V.R. (1996) A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environ. Pollut.* 92, 349–357
66 Marzban, C. and Stumpf, G.J. (1996) A neural network for tornado prediction based on Doppler radar- derived attributes. *J. Appl. Meteorol.* 35, 617–626
67 Orre, R. *et al.* (2005) A Bayesian recurrent neural network approach for finding dependencies in large incomplete data sets. *Int. J. Neural Syst.* 15, 207–222
68 Edwards, I.R. (1997) Adverse drug reactions: finding the needle in the haystack. *BMJ* 315, 500
69 Copas, J.B. and Hilton, F.J. (1990) Record linkage: statistical models for matching computer records. *J. R. Stat. Soc. Ser. A Stat. Soc.* 153, 287–320
70 Hauben, M. *et al.* (2007) 'Extreme duplication' in the US FDA adverse events reporting system database. *Drug Saf.* 30, 551–554
71 Bousquet, C. *et al.* (2005) Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. *Int. J. Med. Inform.* 74, 563–571
72 Berry, S.M. and Berry, D.A. (2004) Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics* 60, 418–426
73 Cerrito, P. (2001) Application of data mining for examining polypharmacy and adverse effects in cardiology patients. *Cardiovasc. Toxicol.* 1, 177–179
74 Harrison, J.H., Jr (2008) Introduction to the mining of clinical data. *Clin. Lab. Med.* 28, 1–7
75 Bytzer, P. and Hallas, J. (2000) Drug-induced symptoms of functional dyspepsia and nausea. A symmetry analysus if one million prescriptions. *Aliment Pharmacol. Ther.* 14, 1479–1484
76 Bate, A. *et al.* (2004) Knowledge finding in IMS Disease Analyser Mediplus UK database-effective data mining in longitudinal patient safety data. *Drug Saf.* 27, 917–918
77 Tukey, J.W. (1969) Analzying data: sanctification or detective work? *Am. Psyschologist* 24, 83–91
78 Shakir, S.A. and Layton, D. (2002) Causal association in pharmacovigilance and pharmacoepidemiology thoughts on the application of the Austin Bradford-Hill criteria. *Drug Saf.* 25, 467–471
79 Aronson, J.K. and Hauben, M. (2006) Anecdotes that provide definitive evidence. *BMJ* 333, 1267–1269