LEADING ARTICLE

Drug Safety 2005; 28 (11): 981-1007
0114-5916/05/0011-0981/$34.95/0

Perspectives on the Use of Data Mining in Pharmacovigilance

June Almenoff,[1] *Joseph M. Tonning*,[2] *A. Lawrence Gould*,[3] *Ana Szarfman*,[2] *Manfred Hauben*,[4,5,6] *Rita Ouellet-Hellstrom*,[2] *Robert Ball*,[2] *Ken Hornbuckle*,[7] *Louisa Walsh*,[8] *Chuen Yee*,[9] *Susan T. Sacks*,[10] *Nancy Yuen*,[1] *Vaishali Patadia* *,[11] *Michael Blum*,[12] *Mike Johnston* **,[2] *Charles Gerrits* ***,[13] *Harry Seifert*[1] and *Karol LaCroix*[1]

1 GlaxoSmithKline, Research Triangle Park, North Carolina, USA
2 US Food & Drug Administration, Rockville, Maryland, USA
3 Merck Research Laboratories, West Point, Pennsylvania, USA
4 Pfizer Inc., New York, New York, USA
5 Department of Medicine, NYU School of Medicine, New York, New York, USA
6 Departments of Pharmacology and Community and Preventive Medicine, New York Medical College, Valhalla, New York, USA
7 Eli Lilly and Company, Indianapolis, Indiana, USA
8 AstraZeneca LP, Wilmington, Delaware, USA
9 Johnson & Johnson Pharmaceutical Research & Development L.L.C., Titusville, New Jersey, USA
10 Hoffmann-La Roche Inc., Nutley, New Jersey, USA
11 Allergan Inc., Irvine, California, USA
12 Wyeth Research, Collegeville, Pennsylvania, USA
13 Schering-Plough Research Institute, Springfield, New Jersey, USA

Abstract

In the last 5 years, regulatory agencies and drug monitoring centres have been developing computerised data-mining methods to better identify reporting relationships in spontaneous reporting databases that could signal possible adverse drug reactions. At present, there are no guidelines or standards for the use of these methods in routine pharmacovigilance. In 2003, a group of statisticians, pharmacoepidemiologists and pharmacovigilance professionals from the pharmaceutical industry and the US FDA formed the Pharmaceutical Research and Manufacturers of America-FDA Collaborative Working Group on Safety Evaluation Tools to review best practices for the use of these methods.

In this paper, we provide an overview of: (i) the statistical and operational attributes of several currently used methods and their strengths and limitations; (ii) information about the characteristics of various postmarketing safety databases with which these tools can be deployed; (iii) analytical considerations for using safety data-mining methods and interpreting the results; and (iv) points to consider in integration of safety data mining with traditional pharmacovigilance methods. Perspectives from both the FDA and the industry are provided.

* Currently with Amylin Pharmaceuticals.
** Retired from the US FDA.
*** Currently with Takeda Global Research and Development.

Data mining is a potentially useful adjunct to traditional pharmacovigilance methods. The results of data mining should be viewed as hypothesis generating and should be evaluated in the context of other relevant data. The availability of a publicly accessible global safety database, which is updated on a frequent basis, would further enhance detection and communication about safety issues.

The term 'data mining' refers to the use of computerised algorithms to discover hidden patterns of associations or unexpected occurrences (i.e. 'signals') in large databases. These signals can then be evaluated for intervention as appropriate. Information gained from data-mining analyses can generate hypotheses that can be validated by other means.

Large postmarketing drug safety databases are the key data source currently used for drug safety data mining. Analysing these data is challenging because these voluntary reporting systems are subject to the problems of under-reporting, various reporting biases and incomplete, unverified data. The number of drug safety databases is also growing rapidly, with some databases containing millions of records. The application of computerised algorithms offers the opportunity to analyse these large databases in a timely and consistent manner. This paper will discuss the role of data mining in pharmacovigilance.

1. History and Mission of the Working Group

In the last 5 years, regulatory agencies and drug monitoring centres have been developing computerised data-mining methods to better identify reporting relationships in spontaneous reporting databases that could signal possible adverse drug reactions. Some pharmaceutical manufacturers are now using these methods via several commercial applications that have become available. However, at present there are no guidelines or standards for the use of these methods in routine pharmacovigilance.

In 2003, we formed a collaborative working group of statisticians, pharmacoepidemiologists and pharmacovigilance professionals from both the US pharmaceutical industry and the US FDA. Individuals from the industry serving on the Working Group are not official representatives of their organisations. The mission and goals of the Pharmaceutical Research and Manufacturers of America-FDA Collaborative Working Group on Safety Evaluation Tools are:

- to develop a consensus view of best practices to optimise the use of data mining in pharmacovigilance and risk management;
- to better understand the databases used for data mining, including data quality issues, and the optimal configurations and specifications for various uses;
- to better understand the possibility of assessing the performance characteristics of various data-mining methods in the drug safety arena for which no true and established gold standards exist;
- to understand the strengths and limitations of these methods, particularly as they affect the interpretation of results;
- to create opportunities for the FDA and industry to develop a common language, to share systematic approaches to the detection and assessment of safety signals from postmarketing adverse event (AE) data and to improve communication regarding data-mining issues;
- to communicate this information to industry and regulatory colleagues.

The use of data mining in pharmacovigilance is a complex topic and the organisations represented on the Working Group are at different stages of use and acceptance of these methods. Data mining in pharmacovigilance is also an evolving science and there was often lack of agreement among group members regarding preferred methodologies, signal definitions and even whether some of the references cited had adequate data to support the claims that were made. For these reasons, developing a consensus view of best practices was not always possible. Hence, this paper will present a spectrum of views on the uses of these techniques and how they fit into the pharmacovigilance 'toolbox'.

2. The Role of Data Mining in Pharmacovigilance

The role of data-mining methodologies in pharmacovigilance is evolving. Evaluating the value and utility of these methods to the pharmaceutical industry and regulators is a work in progress. The Working Group believes that *potential* values of safety data mining include the following.

- Systematic, automated and practical means of screening large datasets.
- Better utilisation of the large safety databases maintained by the FDA, the WHO and other organisations.
- Improved efficiency by focusing pharmacovigilance efforts on key reporting associations.
- Positive contributions to public health by identifying potential safety issues more quickly and/or more accurately than traditional pharmacovigilance methods.
- Better decision support for the pharmaceutical industry and regulators because of broader insight/knowledge of drug safety.

It is important to state at the outset of this discussion that all of the data-mining methods discussed in this paper identify *observed reporting relationships* between drugs and events in large safety databases. These reporting relationships are based solely on the frequency with which drugs and events are reported and thus cannot prove or refute causal relationships between drugs and events. Reporting relationships identified by data-mining methods must be viewed as *hypotheses* regarding *possible* causal relationships between the drugs and events of interest, when observed in the appropriate clinical contexts. Subsequent detailed clinical case reviews and other investigations, as appropriate, are necessary to explore hypotheses generated from data mining.

Data mining has the potential to clarify the many complex interdependent factors (e.g. concomitant drugs and/or diseases) that can play a role in the development of AEs in a clinical setting. Traditional methods may not always be able to detect these complex relationships. Drug exposure data and background rates of AEs of interest are often difficult to obtain systematically;[1-3] thus, it is often difficult for a safety evaluator to put counts of reported events in context. For some serious events, the reporting of one or two such events should prompt a review, regardless of context. For non-serious events and for serious events known to occur in the patient population for a variety of reasons (including exposure to drugs), it is less straightforward to define a threshold for the number of reports that should necessitate a review. For a pharmacovigilance department with many drugs to monitor, a comparative measure of reporting frequency (as provided by data mining) may be seen as an improvement over crude frequency counts and may aid in identifying potential safety issues and prioritising work. Data mining may also add value by detecting disproportionalities involving multiple drugs or multiple events that would be too difficult to detect by traditional methods.

Potential limitations of data mining include those inherent to postmarketing safety databases (e.g. under-reporting, reporting biases) that no signal detection method is likely to overcome. There are published examples of known safety issues that are not retrospectively identified by data-mining methods using predefined thresholds; this is not surprising since not all safety issues emerge from spontaneous reports.[2,4,5] There are concerns that, in some situations, data mining may generate more signals than can be followed up effectively with available resources. In this case, focus might be directed to signals with the greatest public health impact and seriousness. There is also concern about the lack of systematic, objective validation of the methods, a problem that also exists for traditional pharmacovigilance methods. Unfortunately, efforts to validate data-mining methods (and traditional methods) are complicated by the absence of a gold standard for identifying true drug toxicities, although various imperfect reference standards may be used to obtain useful insights on the performance of any method (see section 4.2.2). For this and other reasons, it has not yet been practical to evaluate data-mining methods or traditional methods using performance criteria generally accepted for screening and diagnostic tests.

The Working Group believes that data mining has a place in the pharmacovigilance toolbox but acknowledges that more work is needed before that place is fully defined. Systematic evaluation using traditional and data-mining methods with large

databases will be needed to determine if the promise of the methods actually pays off in practice. Hence, the intent of this paper is to review the current use of these methods for quantitative signal detection, to briefly highlight uses other than signal detection, to share the insights and experiences of Working Group members and to stimulate further discussion and investigation into the utility of data mining in pharmacovigilance.

3. The Need for Consistent Terminology

There is a lack of consistency of terminology in the quantitative signal detection literature. The WHO defines a signal as "reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously. Usually, more than a single report is required to generate a signal, depending on the seriousness of the event and the quality of the information."[6,7] More recently, the report of the Council for International Organizations of Medical Sciences (CIOMS) VI project offered the following definition for signal: "a report or reports of an event with an unknown causal relationship to treatment that is recognised as worthy of further exploration and continued surveillance".[8]

Since these definitions do not specify the *type* of information that constitutes a signal, it is reasonable to view a signal as any information, qualitative or quantitative, that prompts further investigation of the relationship between a drug and an event.

In the context of data mining, some authors use the terms 'association' and 'signal' interchangeably. The term 'signal' is often defined in terms of the quantitative association alone. Others distinguish a signal as an association that has additional supportive clinical information.[1,2,7,9]

The Working Group believes that the consistent use of terminology related to data mining would facilitate communications among pharmacovigilance practitioners. We encourage those who generate, report, publish and/or present data-mining analyses to provide clear, unambiguous definitions of such terms, as these definitions are critical to understanding and evaluating the results. The definitions for terms used in this paper are given in table I.

Table I. Definitions of terms used in this paper

Term	Definition
Drug-event pair	Refers to the co-reporting of a drug and an event in a case report
Association	A relationship between a drug and an event, irrespective of the strength of the relationship. The presence of a *reporting* association between a drug and an event is *purely statistical* and by no means implies a direct, or even an indirect, causal relationship
Signal	A relationship between a drug and event that is strong enough, using a predefined threshold or criteria set by an analyst, to warrant further evaluation
Signal 'score'	A number reflecting the 'strength' of a reporting association, i.e. by how much the observed frequency differs from 'expected'. 'Expected' can be defined in various ways, depending on the criteria that are set for the analysis. There are several methods for computing a signal score
Quantitative signal detection	Refers to computational or statistical methods used to identify drug-event pairs (or higher-order combinations of drugs and events) that occur with disproportionately high frequency in large safety databases
Reporting proportion	The reporting proportion for a specific time period is defined as the number of adverse event reports containing both the target drug and the target event divided by the total number of adverse event reports for the target drug over the same period of time (see also section 4.1)
Reporting ratio	The reporting ratio corresponding to the target drug and the target event over a defined time period is equal to the reporting proportion for the target drug and target event divided by the marginal reporting proportion for the target event. The marginal reporting proportion is equal to the total number of reports for the target event divided by the total number of reports in the database. The marginal reporting proportion for the target event may be computed using all of the reports or using only those reports that do not mention the target drug (see also section 4.1)
Safety data mining/ disproportionality analysis	The application of computer-assisted computational and statistical methods to large safety databases for the purpose of systematically identifying drug-event pairs reported at disproportionately high frequencies, relative to what a statistical independence model would predict. 'Safety data mining' and 'disproportionality analysis' are used interchangeably in this paper

4. Overview of Data-Mining Methods used for Quantitative Signal Detection

4.1 Statistical Principles

Many descriptions of data-mining methods, as applied to pharmacovigilance, are available.[2,10-13] The most commonly used methods with the greatest published experience are the proportional reporting ratio (PRR)[9,14] and the reporting odds ratio (ROR),[15-17] as well as Bayesian[18,19] and empirical Bayesian[20,21] methods that account for the variability associated with small report counts. All of these methods identify statistical *associations* between drugs and events in the reports contained in the spontaneous reporting databases. These associations are based solely on the frequency with which drugs and events are reported together and thus must be viewed as *hypotheses* regarding possible causal relationships between the drugs and events of interest, recognising that there are many possible reasons other than a direct causal relationship for the observed association.

The quantitative evaluation of the relative frequency of reports in spontaneous reporting databases that mention both a particular target drug and a particular target AE is based primarily on the entries in table II.

Thus, of a total of T reports in the database, M mention the target drug, N mention the target AE, A mention both the target drug and the target AE, C mention the target AE but not the target drug, and D mention neither the target drug nor the target AE. The reporting ratio (RR) [equation 1]

$$RR = \frac{A}{Expected(A)} = \frac{A}{MN/T} \quad \text{(Eq. 1)}$$

measures relatively how much more or less often reports in the database actually mention the target drug and target AE than would be expected if the mention of either was statistically independent of whether the other was mentioned or not. Under this assumption of independence, MN/T reports would be expected to mention the target drug and target AE.

The expected value of A can be expressed in other ways, giving rise to statistics similar to the RR.

Table II. Number of reports mentioning a target drug and target adverse event (AE)

No. of reports	Target AE	All other AEs	All AEs
Target drug	A	B	M
All other drugs	C	D	T – M
All drugs	N	T – N	T

The PRR is obtained when Expected(A) = MC/(T − M). The expected likelihood that a report mentioning the target drug also mentions the target AE [i.e. Expected(A)/M = C/(T − M)] for the PRR is based on the target AE reporting proportion among reports not mentioning the target drug. In contrast, the expected likelihood for the RR [Expected(A)/M = N/T] is based on all reports, including those mentioning the target drug.

If many reports mention the target drug and many reports mention the target AE, then there will not be much uncertainty associated with Expected(A). However, if the target drug has not been on the market long (i.e. M is small) or if the target AE is rare (i.e. N is small), then there may be considerable uncertainty about Expected(A) that should be accounted for when any of the statistics are interpreted. Methods have been described for doing so, based on Bayesian[18] and empirical Bayesian[20,21] principles. Software is available to carry out the calculations. Gould[10] provides a detailed comparison of the two approaches.

4.2 Performance Characteristics

4.2.1 Overview

Pharmacovigilance professionals and institutions contemplating the use of data-mining methods to supplement their traditional[22] or manual methods of signal detection should consider a number of factors, including which method to use, which database(s) to use and the choice of variable configurations that can be specified for each data-mining run. Research on data-mining algorithms is dynamic, with new methods under development. This report focuses on the most commonly cited disproportionality methods: PRR, ROR, Bayesian methods and empirical Bayesian methods. Other methods that have been or are being developed based on various statistical algorithms/techniques include probability filtering ('PROFILE'),[23] fuzzy

logic,[24,25] sequential probability ratio testing[26,27] and a tree-based scan statistic.[28]

4.2.2 Challenges in Assessing Performance

Many methodological issues complicate a systematic and comprehensive assessment of the performance of the methods discussed in this paper. These issues include: the variety and volume of data populating spontaneous reporting databases; variations in database environments/architectures; the lack of standards for adjudicating causality and expectedness; the lack of a gold standard with which to calculate traditional screening or diagnostic metrics (i.e. predictive values, sensitivity and specificity); and the lack of clear guidelines on desirable performance characteristics in pharmacovigilance.

A major difficulty arises in trying to evaluate how well any pharmacovigilance method identifies toxicities that truly are causally associated with drugs. The language of diagnostic evaluation is intuitively appealing for this purpose. The truth table (table III) provides explicit definitions of terms used in diagnostic evaluation.

In the context of signal detection, a 'false-positive' finding would be a disproportionately high frequency of drug-event reports that is shown, by other means, to represent an artifactual relationship. Similarly, a 'false-negative' finding would be a drug-event association that does reflect a causal relationship but is not disproportionately reported or is reported less frequently than expected, based on all other drug-event associations in the database.

Unfortunately, the lack of an objective measure of 'truth' (a gold standard) or evidence of a true causal relationship makes the evaluation and validation of all signal detection methods in terms of sensitivity, specificity and predictive value difficult. Given that there are no perfect gold standards, some authors have attempted to validate these methods using imperfect gold standards ranging from selected events in product labelling,[9,19] to selected published information from epidemiological studies and/or reports of positive rechallenge,[29] to labelled events updated by information from large clinical trials.[2]

Table III. Truth table for assessing signalling

Signal	True causality	
	Yes	No
Yes	a	b
No	c	d

Negative predictive value = true negative signal/negative signa = d/(c + d).

Positive predictive value = true positive signal/positive signal = a/(a + b).

Sensitivity = true positive signal/true causal = a/(a + c).

Specificity = true negative signal/true noncausal = d/(b + d).

One disproportionality measure cannot be judged better (or worse) than another because it yields a 'signal' (i.e. exceeds its arbitrary critical value) more often in the absence of a well accepted gold standard. Sensitivity can always be increased at the expense of specificity and *vice versa*. More experience is necessary over a broad spectrum of potential drug-event relationships, algorithms and threshold metrics in 'real life' pharmacovigilance settings to provide a better idea of the diagnostic potential of disproportionality measures. However, the internal validity of these methods is suggested by their ability to reliably detect relationships that are already known. This is reassuring, because failure to recognise these relationships would suggest the possibility of a high false-negative rate and would seriously compromise the value of exploring spontaneous reporting databases for early detection of potential toxicity issues.

Brief summaries follow in section 4.2.3 of published efforts to validate or describe the utility of the most commonly used methods. The Working Group believes that much work remains to be done in this area.

4.2.3 Published Evaluations of Performance

Proportional Reporting Ratio

Evans et al.[9] used the Adverse Drug Reactions On-line Information Tracking (ADROIT) database, the postmarketing safety database of the Medicines and Healthcare products Regulatory Agency (MHRA, formerly the MCA) in the UK. They evaluated 481 'signals' for 15 drugs in the database that were identified using the PRR and found that 339 (70%) were recognised adverse reactions (per labelling), 62 (13%) were events considered likely to be related to the underlying disease and 80 (17%) required further evaluation. Of the 80 events requiring evaluation, 22 warranted a detailed review (leading

to requests for labelling changes for three events), 22 were to be kept under continuing review and no further action was planned for the remaining 36 events. The MHRA determines which signals identified by the PRR need further follow-up in terms of four factors (called SNIP criteria): Strength of signal; whether the signal is New or not; the clinical Importance; and the potential for Preventive measures. More recently, the MHRA has piloted a scoring system to assess which signals require detailed evaluation.[30] In this system, signal strength (PRR value) is one of three factors used to compute an 'evidence score' that is plotted against a 'public health score' to assess the potential importance of the signal. Preliminary evidence suggests that this innovation may be useful for systematising the evaluation process and aiding scientific discussion.[31,32]

Reporting Odds Ratios

The ROR (A/B ÷ C/D or AD/BC, see table II) has been described in the pharmacovigilance literature as an additional analytical approach for disproportionality analysis of spontaneous data.[16,17] The ROR, like the traditional odds ratio, is an estimate of the incidence rate ratio; it estimates the odds of the AE in those exposed to a particular drug divided by the odds of the AE occurring in those not exposed to drug. The ROR is not affected by general underreporting for a specific drug or a specific event.[16] Rothman et al.[33] have proposed that the ROR may, in theory, be a less biased methodology than other disproportionality methods in that a series of spontaneous reports can be viewed as cases and controls; the 'cases' are those experiencing a specific AE, the 'controls' are those that do not experience the AE (in a spontaneous reporting database that would be those with 'other AEs') and the 'exposure' is exposure to the specific drug under study. However, others believe that in practice, both the PRR and the ROR yield similar results and that there is no benefit in using the ROR instead of the PRR.[17,34]

Bayesian Confidence Propagation Neural Network

The Uppsala Monitoring Centre (UMC) uses the Bayesian Confidence Propagation Neural Network (BCPNN) software to identify drug-event pairs that stand out statistically from the background of all reports in the database. Members of the UMC international expert assessment panel then decide which of these constitute potential drug-AE relationships that should undergo more detailed evaluation. Bate et al.[18] described how the system could have detected the relationship between captopril and cough before it was widely reported in the literature and provided an example of false-positive signal avoidance. Lindquist et al.[7] checked case reports for critical terms published in *Reactions Weekly* from January to June 1998 against the WHO database for the same time period. They found that 12 of 43 pairs appearing as 'first reports' in *Reactions Weekly* fulfilled the criteria of association in the WHO database using the BCPNN system at the same time as, or before, appearing in the publication. Lindquist et al.[19] also described a retrospective evaluation where the 'gold standard' was whether or not the signalled association was described in the reference literature (*Martindale's Extra Pharmacopoeia* and the *Physicians' Desk Reference*) at a given point in time or whether the association was confirmed or strengthened over a specified period. In this evaluation, the BCPNN detected signals in the WHO database with a 'positive predictive value' of 44% and a 'negative predictive value' of 85%. The BCPNN identified six of the ten signals produced by the former system used at the UMC, four of the six being detected earlier than with the former system.

Recently, the UMC has introduced triage logic to further filter the large number of associations that are generated by the BCPNN and sent to reviewers for evaluation.[35] The filters are applied to the combinations database produced by the BCPNN scan to reduce the number of combinations highlighted for assessment and to help focus on the areas of greatest importance. The filters currently in use highlight rapid increases in reporting, serious reactions with new drugs and reactions of special interest such as those very likely to be drug related. After using these filtering strategies for some time, the UMC plans to evaluate how successful they have been in finding 'potential signals' and to examine whether early detection of important signals has been enhanced.

Empirical Bayes (Gamma Poisson Shrinker, Multi-Item Gamma Poisson Shrinker)

Several retrospective studies in which empirical Bayesian methods detected early signals of AEs

have been published. An analysis utilising multi-item gamma Poisson shrinker (MGPS) showed a number of signals for rhabdomyolysis and renal failure for cerivastatin several years before this drug was removed from the market.[2] MGPS also showed important signals of various adverse drug events in paediatric and adult patients.[1,3] In a previous study, the gamma Poisson shrinker (GPS, a precursor to MGPS) was applied to 30 drug-event combinations declared as signals by FDA epidemiologists using traditional methods applied to the Spontaneous Reporting System (SRS) [now known as the Adverse Event Reporting System (AERS)] database.[2] The GPS method signalled all 30 of these selected drug-event combinations, with 20 signalled by GPS in the data collected 1–5 years before index cases were detected by traditional methods, nine signalled by GPS the same year and one signaled by GPS a year after the data were detected by traditional methods.

The GPS was also used to analyse the differences in time in detecting 160 drug-event combinations involving 85 drugs. These 160 drug-event combinations were coded as signals between 1985 and 1996 by FDA safety evaluators and collected in the FDA Monitoring Adverse Reports Tracking (MART) system. These 160 drug-event signals were selected for data-mining analysis because the drug-event pair names in the MART matched the drug-event pair names in the SRS. GPS signaled 97 drug-event combinations in the SRS data collected 1–4 years before they were entered as signals in the MART system, with 36 signaled by the GPS the same year and 27 signaled by the GPS 1–3 years later. One-half of the 27 signals detected later by GPS included designated medical events (such as severe liver events, Stevens-Johnson syndrome, aplastic anaemia and anaphylaxis) that are easier to characterise with fewer reports.[2,36]

These studies illustrate the potential for the GPS/MGPS to detect signals of drug-event combinations that have been declared to be signals by traditional methods.

Comment on the Comparative Performance of Data-Mining Methods

There are no published, large-scale, systematic comparisons of data-mining methods currently used for pharmacovigilance and the published performance characteristics vary, depending on criteria selected for signal detection. Although the precise statistical approaches of the methods differ, they all involve some assessment of disproportionality and would therefore be expected to provide overlapping results. Some investigators have shown concordance among methods when the number of reports exceeds four.[17] It has also been observed that Bayesian and empirical Bayesian methods generate fewer signals than the PRR when commonly cited thresholds are used. This is to be expected since both Bayesian and empirical Bayesian methods make adjustments for the increased variability associated with small actual and expected report counts. It should be noted that the number of drug-event pairs signaled by any of the available methods depends in large part on selection of empirical signal thresholds that involve subjective judgements. By itself, the number of signals flagged by a particular method is an inappropriate criterion for comparing the performance of data-mining algorithms. As discussed previously, all methods incur tradeoffs between sensitivity and specificity, especially when varying criteria for eliciting signals are used.[2-4,37-41] When examining the literature on performance analyses of drug safety data mining, it should be borne in mind that sensitivity and specificity are highly dependent upon the definition of signal thresholds used, the minimum number of reports required before a signal is declared, the number of relevant event codes analysed, the type of data configurations utilised (reports from manufacturers versus reports from all other sources, etc.) and many other factors. There is no basis at present for recommending any of these methods and signal thresholds as superior for all users and situations.

5. Postmarketing Safety Databases used for Quantitative Signal Detection

Databases utilised in drug safety data mining include large postmarketing databases maintained by regulators, pharmaceutical manufacturers and various consortia, each with their own reporting criteria, coding dictionaries and data entry rules. Extracting meaningful data from these databases is often challenging because voluntary reporting systems are subject to the problems of under-reporting, various reporting biases and incomplete, unverified

data. Some of these databases are quite large, containing hundreds of thousands and even millions of AE reports. Despite their inherent limitations, the size and scope of these databases make them appealing for pharmacovigilance.

The Working Group acknowledges that this situation of differing databases is far from ideal, as results may vary between databases. Ideally, there would be a single 'canonical' database available to industry and regulators in real time; such a database would contain worldwide data on all products, have no duplicate reports and employ consistent conventions for drug naming, event coding and data entry. The reports submitted to such a database would be complete and include treatment indication, past medical history and co-medications.

Until such a 'canonical' database exists, essentially three databases are available to pharmaceutical manufacturers for signal detection activities: (i) their own internal safety database(s); (ii) the FDA's public-release safety databases (SRS/AERS and the Vaccine Adverse Event Reporting System [VAERS]); and (iii) the database of the WHO International Drug Monitoring Programme. These databases are described and discussed in sections 5.1–5.4. Regulators typically rely on their own agency databases for signal detection activities.

5.1 US FDA Adverse Event Reporting System (AERS) and Spontaneous Reporting System (SRS)

AERS is the FDA's postmarketing safety database and is used herein to refer to the combined datasets of SRS (1968 to October 1997) and AERS (November 1997 to present). The public-release version of AERS is available for purchase from the National Technical Information Service (NTIS) on a quarterly basis [1] and, as of December 2004, from the FDA's website beginning with the January 2004 quarterly data (http://www.fda.gov/cder/aers/default.htm). AERS is a surveillance system that relies on voluntary reporting of AEs to the FDA by healthcare professionals and consumers, as well as required reporting by pharmaceutical manufacturers. AERS includes spontaneous reports from US sources, serious and unlabelled spontaneous reports from non-US sources and serious, unlabelled and attributable postmarketing clinical trial reports from all sources. As of December 2004, AERS contained approximately 2.6 million reports. The size and diversity of this database are its primary advantages.

At present, there are several important limitations in using the public-release version of AERS data. Although historically there has been a lag time of 9–12 months for release of data through the NTIS, the FDA anticipates that this interval will shorten.

Another limitation of the database is the presence of duplicate and multiple reports for some cases. NTIS data contain all reports received by the FDA in AERS. Multiple reports of the same case are generated from updates by the manufacturers to previously submitted original reports. Potential duplicate reports of the same case are generated from reports by multiple manufacturers and 'direct' reports received from healthcare providers and consumers via the FDA's MedWatch programme. The multiple reports are linked by the manufacturer's control number in the internal FDA database only, leaving the public-release database with potential duplicates and multiple reports for a case. Although there are no plans at present to remove duplicate or multiple reports from the public-release version of AERS, commercial vendors provide versions of AERS that have been 'cleaned' by consolidating multiple and duplicate reports. However, there may be differences between datasets provided by various vendors because of the use of different duplicate detection and removal algorithms.

The AERS database also lacks standardisation of drug names. The FDA attempts to link the reported verbatim drug name to an 'active ingredient'. Correction or standardisation of names of combination products, different formulations, herbal products, foreign drug products and spelling errors is necessary to consolidate spellings and product names to

1 Access to the entire AERS database (public-release version) is available from commercial vendors on a subscription basis; these vendors offer the service of collecting all of the updates issued by the NTIS into one repository. Each vendor uses its own rules and algorithms to 'clean' the database by standardising drug names, accommodating changes to coding dictionaries and removing duplicate reports. Selection of a vendor may require an evaluation period to ensure that the methods used to format and clean the public data are acceptable to users.

improve the fidelity of data mining. The FDA is participating in efforts to develop a global coding dictionary with ICH M5 (Data Elements and Standards for Drug Dictionaries) [http://www.ich.org]. Commercial vendors provide access to datasets in which drug names have been organised according to their respective methods.

Reports from outside the US that are present in AERS are likely to be serious, unlabelled events, whereas both labelled and unlabelled events, regardless of seriousness, are present in AERS for reports from within the US. Therefore, it is possible that signals could be generated for drugs with a high proportion of foreign to domestic reports because most of the foreign reports received for these drugs are only for serious, unlabelled events. This can be addressed by stratifying on reports by foreign/domestic submissions, using separate analyses or analysing company databases.

The types of case reports that are being entered into AERS are changing over time. Since data-mining algorithms derive the frequency of 'expected' drug-event pairs used as the denominator from the total AERS database, understanding the impact of changes in the composition of AERS is vital. The electronic submission of reports and the availability of waivers for submission of non-serious, expected events are two examples of changes that have influenced the content of AERS. Since 1998, non-serious, expected reports for drugs marketed for ≥3 years have not been entered into AERS if they were submitted on paper. However, electronic submissions may be directly imported into AERS, although at this time most are still undergoing the FDA quality control process before being entered into AERS. As electronic submissions increase, this 'shift' in the content of the AERS database has the potential to modify the database in a favourable way so that all reports, regardless of their labelledness, will be entered, thus creating a more complete and coherent dataset. The effect of these changes on the results of data-mining analyses is unknown.

5.1.1 Working Group Recommendations Regarding AERS

The members of the Working Group, although understanding the financial constraints of the FDA, believe that improvements to the public-release version of AERS would enhance the utility of the database. Toward this end, the Working Group has made several recommendations concerning the NTIS AERS product, some of which have already been addressed by the FDA.

- Decrease the lag time between report receipt by the FDA and the public release of AERS.
- Publish the entry specifications and coding conventions to enhance understanding of the data.
- Make available as many data fields (including narratives) as possible without infringing on patient privacy.

5.2 US FDA Vaccine Adverse Event Reporting System Database

In the US, surveillance of AEs after vaccination is undertaken by the government using VAERS, which the FDA and the Centers for Disease Control and Prevention (CDC) jointly manage. VAERS is the national system for surveillance of AEs after vaccination. It was initiated by the 1986 National Childhood Vaccine Injury Act and was established in 1990. The uses of VAERS include detecting novel AEs, monitoring the frequency and severity of known AEs, identifying possible risk factors, and vaccine lot surveillance.

VAERS is substantially smaller than AERS, receiving 10 000–15 000 reports per year on top of approximately 160 000 existing reports. AE data in these reports are coded using the Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART) dictionary. Some reports are submitted directly to VAERS and are therefore not in manufacturers' databases. Because vaccines are more likely to be given to children than adults, and to healthy rather than ill people, the VAERS database contains a predominance of reports involving children.

5.2.1 Systems of Administration and Reporting

The system for administering vaccines and reporting AEs is different than the system used for drugs. Drugs are primarily administered by licenced practitioners by prescription in a healthcare system focused on treating illness. While vaccines are sometimes individually prescribed by practitioners, they are often given based on public health guidelines as part of a systematic disease prevention programme, which might include physician's offices,

public health clinics, and the military. Similarly, AEs are reported from each part of this system at different rates that might influence the disproportionality calculated by various algorithms. For example, the recent smallpox vaccination campaign was limited to the military and certain public health workers. Both the military and the CDC had safety surveillance systems in place that went beyond VAERS, although all reports of AEs were submitted to VAERS. Interpretation of data-mining analyses that compared the smallpox vaccine with other adult vaccines would need to take this differing reporting mechanism, and the possibility of higher reporting rates, into consideration.

5.3 WHO Safety Database

The WHO safety database is a large, global database with >3.4 million individual case reports spanning >30 years (1968 to present). AE reports are contributed by national centres participating in the WHO International Drug Monitoring Programme.[42]

There are currently 78 member countries that submit domestic AE reports to the WHO database, ideally on a quarterly basis. A significant proportion of the WHO database comprised reports from the AERS (US) database. The top five contributors by number of reports received since joining the programme are the US (1 314 525), UK (391 868), Germany (160 648), Australia (146 116) and Canada (136 192). Only domestic cases from the US are entered. Differences in reporting requirements between countries should be considered in an analysis of this database. Some of the differences between countries relate to whether reports are voluntary or mandatory, or whether consumer reports are accepted. Furthermore, reporting rates and profiles may also be influenced by differences in medical practice and societal factors.

The report sources include healthcare professionals, consumers and marketing authorisation holders. Most consumer reports are from the US. Duplicate cases are identified by a systematic check for the same case ID and by analysis of case series. AEs are coded using the WHO-Adverse Reaction Terminology (WHO-ART) coding dictionary and drugs are coded using the WHO Drug Dictionary, which offers indexing and retrieval of drugs by the hierarchical Anatomical Therapeutic Chemical (ATC) classification.

Strengths of the WHO database include the capability to evaluate drugs by generic or trade name, the capability to identify between-country differences and the capability to identify well documented reports via a quality grading system. As with AERS, limitations of the WHO database include a limited systematic process for identification of duplicates, many empty data fields and the unavailability of case narratives.

Access to the WHO database is available by subscription either directly from the WHO or through commercial vendors. As with other databases, selection of a vendor may require an evaluation period to ensure that the vendor's methods for formatting the data are acceptable to users.

5.4 European Medicines Agency EudraVigilance Database

The EMEA created and maintains a pharmacovigilance database management system and data processing network known as EudraVigilance. EudraVigilance was created for the electronic exchange and processing of AE reports involving medicinal products authorised in the European Economic Area (EEA). It offers remote access to registered partners and their administrative and scientific users in the European Commission, the EMEA, Competent Authorities in the EEA and pharmaceutical companies via a secure connection over the internet. EudraVigilance contains both a clinical trial (EVCTM) and a post-authorisation module (EVPM). The EVPM was established in December 2001 to support reporting requirements for spontaneous reports and adverse reaction reports originating from organised data collection systems (e.g. registries and post-authorisation safety studies). As of July 2005, 118 791 Individual Case Safety Reports (ICSRs) corresponding to 70 901 cases were reported from outside the EEA to the EVPM and 60 326 ICSRs corresponding to 35 649 cases were reported to the EVPM from within the EEA. The latter reporting activity originated from 47 market MAHs and 15 member states. Retrospective electronic population of EVPM with legacy data is underway. EudraVigilance includes data analytical capabilities and quantitative signal detection func-

tionality based on PRRs and RORs. At present, pharmaceutical companies have restricted access to EudraVigilance in that each can only view AE reports that they have submitted to EMEA.[43]

5.5 Company Safety Databases

Although it is technically feasible to use data mining with in-house company safety databases, there are a number of caveats to consider. Although there are no precise guidelines, the database should be of sufficient size and diversity to serve as a suitable 'background' for evaluating disproportionate reporting. Among the potential limitations of company databases are a relative lack of diversity of events or drugs, which leads to a greater likelihood of masking (see section 6.6).[10] One way to measure diversity in a safety database is to examine the number and distribution of reports in the database by therapeutic area or drug product. It may be prudent to also compare the results of analyses using the proprietary database with those obtained using AERS or WHO for several 'well characterised' products. However, interpretation of the clinical impact of such 'diversity' or lack thereof is complicated by the often cited lack of gold standards.

Each institution's proprietary database may have strengths that can be exploited, for example companies with a global dataset, rather than one weighted toward US cases (as AERS is) may find this useful. If the company's database started earlier than 1968, when SRS/AERS started, it may be possible to explore relative reporting frequencies for older drugs. There may be more data elements with consistent data quality and coding, which may allow for further exploration of relative reporting frequencies among demographic subsets. Importantly, because the data are not subject to the delays associated with the public databases, company databases may allow for earlier detection of safety signals, particularly for new products.

6. Analytical Considerations

6.1 Overview

The essential first step in undertaking an exploration of a spontaneous reporting database with data mining is to specify the purpose of the analysis. Depending on the questions specified, technical/analytical options that might be considered include:

- whether to include all drug-event pairs in the analysis or only those pairs where the role of the drug of interest was considered 'suspect';
- whether to base the calculations on counts of drug-event pairs or counts of reports;
- whether to perform the analysis using specific AE terms or groups of related AE terms that are aggregated under a 'higher-level term' with hierarchical AE dictionaries such as the Medical Dictionary for Regulatory Activities (MedDRA);
- whether to stratify the calculation of expected counts (see section 4.1) and, if so, by which variables.

These and other considerations are discussed briefly in the remainder of section 6.

6.2 Role of the Drug (Suspect Only versus Suspect and Concomitant)

Spontaneous AE reports originate from individuals who suspect they have experienced, observed or heard about an adverse drug reaction. A typical report will cite a drug(s) and an event(s) that the reporter believes are related. The reporter may mention other medications, but the reason for the report is the belief that an event is related to a particular drug or drugs. In order to capture this distinction, safety databases such as AERS and proprietary databases maintained by pharmaceutical manufacturers typically classify each drug cited in a report as 'suspect' or 'concomitant'.

The drug and event information in a safety database can be thought of as a very large two-way table composed of many cells. The number in any given cell is the number of reports containing the drug and the event that define that cell. Obviously, the value in a particular cell will be different depending on the choice to only count reports where a drug is coded as suspect (S only) versus all reports containing the drug irrespective of coding as suspect or concomitant (S + C).

The experience of some Working Group members, using empirical Bayesian methods, is that computed relative RRs are slightly higher with 'S only' compared with 'S + C' and that a greater number of drug-event pairs meet or exceed a numerical 'signal'

threshold with 'S only' compared with 'S + C'. However, there are instances where the reverse is true. The Working Group is not aware of any data to date that demonstrates that the differences are clinically important. It would be reasonable to expect that other methods would produce similar results.

There is no reason to believe that either strategy is superior with respect to identifying or not identifying reporting relationships that may turn out to reflect causal relationships. As users gain familiarity with the performance of these methods, they may need to adjust data-mining strategies accordingly.

6.3 Counts of Drug-Event Pairs versus Counts of Reports

Another factor to consider in the implementation of these algorithms is whether the statistical parameters should be calculated with respect to the total number of drug-event pairs or the total number of reports in a given database. Calculations appear to be based on numbers of drug-event pairs in the paper by Evans et al.[9] describing PRR and in the paper by DuMouchel[20] on empirical Bayesian methods. Bate et al.[18] and DuMouchel and Pregibon[21] base their calculations for Bayesian and empirical Bayesian methods, respectively, on numbers of reports. Any of these methods can be executed either way. Statisticians in the Working Group note that either approach is acceptable, although counting reports probably provides a more intuitively appealing estimate of sample size, since reports often contain multiple events and multiple drugs that are not independent of each other. Presently, there are no data demonstrating that the choice of denominator is clinically important.

6.4 Combining Drug and Event Terms

The validity and utility of combining (pooling, 'lumping', collapsing) drug and/or event terms in the setting of safety data mining is not well studied. Combining drugs of the same class or medically related AE terms may allow earlier detection of safety issues by increasing the power of the analysis through larger numbers. This is of particular interest for databases that encode AE data using highly granular dictionaries such as MedDRA. Combining drug or event terms must be done carefully because, as the following example shows, relative RR values can decrease when terms are combined.

Table IV provides the number of reports mentioning the target drug and either of two synonyms for the target AE.

The RR for the target drug using the first synonym is (equation 2):

$$RR_1 = A_1 \div \frac{(M \times N_1)}{T} = A_1 \div E(A_1) \quad \text{(Eq. 2)}$$

The RR for the target drug when the counts for the two synonyms are combined (assuming, of course, that no report ever mentions both synonyms) is (equation 3):

$$RR_c = (A_1 + A_2) \div M \times \frac{(N_1 + N_2)}{T} \quad \text{(Eq. 3)}$$

It is easy to show that the RR value when the synonyms are combined is greater than the value using only the first synonym ($RR_c > RR_1$) if and only if (equation 4):

$$\frac{A_2}{N_2} > \frac{A_1}{N_1} \quad \text{(Eq. 4)}$$

that is, if and only if the target drug is mentioned more often among the reports that mention the second target AE synonym than among the reports that mention the first synonym.

Another consequence of this demonstration is that the increase of RR with the combination over the reporting with the first synonym implies a decrease of RR with the combination related to the RR with the second synonym. This effect on RRs is not necessarily a bad thing. A few reports of a rare event can lead to a very large, but very imprecise, RR value. A high RR value is not meaningful by itself. It takes whatever meaning it might have only when considered in the right context. Combining synonyms will decrease the value obtained for some of

Table IV. Number of reports mentioning a target drug and either of two terms for a target event, assuming no report mentions both adverse event (AE) terms

No. of reports	Target AE 1	Target AE 2	Other AEs	Total
Target drug	A_1	A_2	$M - A_1 - A_2$	M
Total	N_1	N_2	$T - N_1 - N_2$	T

the synonyms, but the ratio based on the combination will become more precise and perhaps more medically relevant.

Combined terms should be highly similar or synonymous to minimise the risk of distorting a result (e.g. QT prolonged and corrected QT [QT_c] prolonged). The driving considerations must be the medical meaning and coding practices, not the statistical consequences. For example, the specific term 'torsade de pointes' should not be combined with the general term 'arrhythmia', because torsade de pointes is a highly specific type of arrhythmia and has different pathophysiological implications than most other arrhythmias. The same point can be made for combining an event such as torsade de pointes with other specific arrhythmia terms that are more likely than torsade to result from non-drug causes.

When planning an analysis with combined terms, the terms should be specified *a priori* and with careful consideration to the medical/scientific meaning of the combination and historical coding practices in the database. Repeated searching for a combination that 'works' may increase the false-positive rate because of multiplicity considerations.

6.5 Stratification

Stratification is a statistical procedure for mitigating the effects of confounding by adjusting for associations between a drug and a variable and an event and the same variable. For example, suppose that drug A is frequently prescribed for men aged >60 years and event B is common in men aged >60 years. Disproportionality analysis might detect a strong association between drug A and event B when the true associations are between the drug and men aged >60 years and between the event and men aged >60 years. In this example, stratification of the computation of expected counts by age and sex removes the effect of confounding. Another commonly used stratification variable is year of report. Stratification by year of report reduces the chance of detecting spurious associations because of temporal factors that may influence the reporting of specific drugs and/or specific events. Some members of the Working Group have also found it useful, when concerned about effects from publicity that stimulates consumer reporting, to stratify on report source (i.e. consumer, healthcare provider). However, stratification will not adjust for over-reporting of a specific drug-event pair. One should be aware that many factors can stimulate reporting and these factors may extend across report sources.[44]

The effect of stratification can be illustrated with an example suggesting that sensible stratification generally should be used. Suppose that one has counts as in table V.

The value of the stratified RR is (equation 5):

$$RR_{str} = A \div \left(M_1 \times \frac{N_1}{T_1} + M_2 \times \frac{N_2}{T_2}\right) \quad \text{(Eq. 5)}$$

The value of the RR ignoring stratification is (equation 6):

$$RR_{uns} = \frac{AT}{NM} \quad \text{(Eq. 6)}$$

The difference between the unstratified and stratified ratios is proportional to (equation 7):

$$\left(\frac{M_1}{T_1} \quad \frac{M_2}{T_2}\right) \times \left(\frac{N_1}{T_1} \quad \frac{N_2}{T_2}\right) \quad \text{(Eq. 7)}$$

The stratified ratio is not necessarily greater than the unstratified ratio, nor is it necessarily less. If the target drug and the target AE are both mentioned more frequently in stratum 1 than in stratum 2, then the stratified ratio will be less than the unstratified ratio regardless of the values of A_1 and A_2. The stratified ratio also will be less than the unstratified ratio if the target drug and the target AE are both

Table V. Number of reports mentioning a target drug and a target event in each of two distinct subgroups (strata) of the patients providing reports

No. of reports	Stratum 1		Stratum 2		Combined	
	target AE	total	target AE	total	target AE	total
Target drug	A_1	M_1	A_2	M_2	$A = A_1 + A_2$	$M = M_1 + M_2$
Total	N_1	T_1	N_2	T_2	$N = N_1 + N_2$	$T = T_1 + T_2$

AE = adverse event.

mentioned less frequently in stratum 1 than in stratum 2. If the target drug is mentioned more frequently in stratum 1 than in stratum 2, but the target AE is mentioned less frequently in stratum 1 than in stratum 2 (or *vice versa*), then the stratified ratio will exceed the unstratified ratio.

If the within-stratum RR is actually the same in both strata, then the stratified ratio will equal the common within-stratum value. However, the unstratified ratio will usually differ from the common within-stratum value. Consequently, since unstratified estimates may present a distorted picture of reporting relationships, especially when RRs differ little among strata, it seems generally advisable to stratify sensibly.

Many potential stratification factors can affect the values of disproportionality measures based on data from spontaneous reporting datasets. The dataset may contain values for some of these, but will not contain values for many others because the information was not captured on the report form or is not recorded in an easily recoverable form. No analysis can stratify by all of the recognised factors, let alone the unrecognised ones.[45] The fact that the value of the RR (or any disproportionality measure) could be increased or decreased by stratification should be borne in mind during any analysis. Disproportionality measures should be computed for important subsets of the patients when there is reason to believe that potential toxicity risks may be particularly elevated in some, but not all, of these subsets (e.g. for elderly, but not non-elderly, patients).

The general consensus of the Working Group is that routine use of stratification for computing expected counts is a reasonable approach. Software programmes should be designed to provide alerts when unusual strata or data distributions exist.

6.6 Masking of Drug-Event Relationships by Experience with Related Drugs

The terms 'masking' and 'cloaking' have been used to describe the effects that experience with related drugs may have on the observed reporting relationships between a drug and various AEs. Masking is possible in any database but because the pharmacovigilance databases held by pharmaceutical companies are generally smaller and less diverse than regulatory databases, the former may be more vulnerable to these effects.

Table VI. Number of reports mentioning either of two drugs and a target event. The rows are not mutually exclusive because a report could mention both drug A and drug B

No. of reports	Target AE	All AEs
Drug A	A	M_A
Drug B	B	M_B
All drugs	N	T

AE = adverse event.

For example, if drug A is an angiotensin-2 antagonist and drug B is an ACE inhibitor that has been on the market longer than drug A, then the information accumulated about drug B may affect relative RR values for drug A.[10] Let us suppose that the reports can be summarised as shown in table VI.

The ratio of the RRs for drug A with and without reports mentioning drug B (which we assume never mention drug A) is (equation 8):[10]

$$RR_A^{(\text{excl B})} \div RR_A^{(\text{incl B})} = 1 + \frac{M_B}{N - B}\left(B/M_B - N/T\right) \quad \text{(Eq. 8)}$$

Clearly, if the reporting proportion for the target AE on drug B (B/M_B) is greater than the overall reporting proportion for the target AE (N/T), then the RR for drug A based on all of the reports will be less than the RR for drug A calculated after removing all of the reports mentioning drug B. Conversely, if the reporting proportion for the target AE on drug B is less than the overall reporting proportion, then the RR for drug A based on all of the reports will be greater than the ratio after removing the reports mentioning drug B. Because of this, and because the value of N/T may largely be determined by reports involving drugs other than drug A or drug B, no blanket recommendation can be given about whether the RR should be calculated including or excluding drug B. When most of the target AEs can be identified with drugs like drug A and drug B, then it may be advisable to compute the RRs both ways.

6.7 Signal 'Absorption' ('Innocent Bystander' Phenomenon)

Signal 'absorption', also known as the 'innocent bystander' phenomenon, occurs when a drug that is

commonly co-prescribed with another drug appears to be associated with an event that is actually associated with the other drug. This is a problem in polypharmacy scenarios. Currently, this phenomenon is identified by case review and is difficult to quantify. Regression techniques may be used to untangle the relative contributions of individual drugs to the high relative RR.[46]

The following example illustrates how the 'innocent bystander' phenomenon can arise. Suppose among a total of T reports in the database, M_B mention drug B and, of these, M_{AB} mention drug A and drug B. Let π_A denote the true likelihood that an AE is mentioned among the reports mentioning drug A (equation 9),

$$\pi_A = \text{Prob}(AE \mid A) \quad \text{(Eq. 9)}$$

and let (equation 10)

$$\pi_B = \text{Prob}(AE \mid B) \quad \text{(Eq. 10)}$$

denote the true likelihood that the AE is mentioned among the reports mentioning drug B. Suppose that $\pi_A > \pi_B$, and that the true probability that a report mentions the AE if it mentions drug A is π_A regardless of whether drug B is mentioned or not, that is (equation 11),

$$\pi_{AB} = \text{Prob}(AE \mid A \text{ and } B) = \text{Prob}(AE \mid A) = \pi_A \quad \text{(Eq. 11)}$$

Let p_B denote the fraction of the reports mentioning drug B that are observed also to mention the AE. The quantity p_B is what one observes if only information about the mention of drug B and the AE in the reports is used. Since some of the reports that mention drug B also mention drug A, the observed reporting proportion for drug B, p_B, will not exceed the true reporting proportion, π_B because (equation 12):

$$\begin{aligned} p_B &= \text{Prob}(AE \mid A \text{ and } B)\ \text{Prob}(A \text{ and } B \mid B) + \\ &\text{Prob}(AE \mid B \text{ and not } A)\ \text{Prob}(B \text{ and not } A \mid B) \\ &= \pi_B + (M_{AB}/M_B)(\pi_A - \pi_B) > \pi_B \text{ when } \pi_A > \pi_B \end{aligned} \quad \text{(Eq. 12)}$$

If the AE is mentioned in N of the T reports, then the RR for the combination of the AE and drug B will be the reporting proportion divided by the overall reporting proportion, N/T. The true RR for drug A is (equation 13)

$$RR_A^{True} = \pi_A \div \frac{N}{T} \quad \text{(Eq. 13)}$$

and the true RR for drug B is (equation 14)

$$RR_B^{True} = \pi_B \div \frac{N}{T} \quad \text{(Eq. 14)}$$

If $\pi_A > \pi_B$, then the *observed* RR for drug B is (equation 15)

$$RR_B^{Obs} = RR_B^{True} + \frac{M_{AB}}{M_B}\left(RR_A^{True} - RR_B^{True}\right) \quad \text{(Eq. 15)}$$

If s $\pi_A \leq \pi_B$, then because we assume (equation 16)

$$\pi_{AB} = \pi_A,\ RR_B^{Obs} = RR_B^{True} \quad \text{(Eq. 16)}$$

In other words, if the true RR for drug A exceeds that for drug B, then the observed RR for drug B will be greater than the true ratio and the degree to which it is increased will depend on how many of the reports mention drug A as well as drug B. Otherwise, the RR for drug B will not be affected. In particular, this means that the RR for drug A will not be affected by the presence of drug B even though the converse is not true, at least under the assumptions used for the argument. The fact that the RR for drug B might be inflated by the effect of drug A does not mean that the true RR for drug B necessarily reflects no association; i.e. drug B might not be an 'innocent bystander'.

7. Issues in Interpreting Data-Mining Outputs

7.1 Overview

On the surface, interpreting the results of a disproportionality analysis is straightforward. Regardless of the method used, the numeric result, coupled with a defined 'threshold', indicates whether or not a drug-event pair of interest has been reported more frequently than 'expected' considering the background (usually the entire database) to which it was compared. The magnitude of the numeric result describes the degree of disproportionality, often referred to as the magnitude or 'strength' of the

'score'. Most methods also return some measure of confidence in the result (e.g. confidence limits or p-values).

Beyond this straightforward interpretation of reporting frequency, much caution is needed. One must understand the strengths and limitations of the method, the configuration details and the database in order to begin to understand the results. Even more importantly, one must understand the product, the event, the treatment of disease, complications of the underlying disease(s), the known pharmacotoxicology of the product and the external reporting environment in order to place the result in context. Apparent associations can occur for many reasons other than causal relationships between drugs and events. As stated previously, associations identified via data mining must be viewed as *hypotheses* regarding *possible* causal relationships between the drugs and events of interest. Causality can be established, if at all, only by careful medical follow-up of the clues about possible associations that are provided by quantitative signal detection methods. In some instances, epidemiological investigation of the issue may be valuable. It is also critically important to remember that the absence of a reporting relationship in a spontaneous reporting database does not rule out the existence of a safety problem and cannot be used to refute a signal detected by other means. Data-mining algorithms assist but do not replace the acuity of knowledgeable medical reviewers.[1,2]

The remainder of this section will discuss some of the issues related to databases, products, and the external environment that one must consider when interpreting the results of quantitative signaling methods. Most, if not all, of these issues are relevant to any method used to evaluate observational data.

7.2 Adverse Event Coding

As with any analysis of AE data, knowledge of the coding dictionary and the conventions used to code event terms is key. Most pharmaceutical manufacturers use MedDRA, which was introduced in the FDA's safety database AERS in November 1997 (replacing COSTART), although some of the new MedDRA terms were introduced in the 1995 version of COSTART. At its introduction, MedDRA had ten times more preferred term (PT) codes than COSTART and was designed with a hierarchical structure to allow the inclusion of more specific terms. Consequently, direct comparison of COSTART codes and MedDRA codes can be problematic.

Considerations related to event coding and quantitative signal detection include the following. (The WHO database uses a different dictionary, WHO-ART, to which many of the same considerations apply.)

- Although potentially important safety events cannot always be anticipated, prospectively grouping AE terms and developing case definitions whenever possible could be beneficial. Prospective grouping might be particularly important for syndromes involving multiple body systems such as serotonin syndrome and drug withdrawal.
- Generating results at the high-level term (HLT) level is generally not helpful as MedDRA HLTs often contain non-homogeneous medical concepts. Non-homogeneous groupings can contain disparate medical concepts, such as both high and low blood pressure PTs under the same HLT or can be groupings of very important and specific terms with less important and less specific terms under the same HLT. Examples of potentially misleading results at the HLT level include:[47]

(i) a high relative RR for the HLT 'ventricular arrhythmia', which is an event that is often of high clinical significance, may raise concern when the majority of reports are for the PT 'extrasystoles', which is an event that is often of minor clinical significance;

(ii) a low relative RR for the HLT 'febrile disorders' may not raise concern, but hidden under this grouping may be a high score for the PT 'neuroleptic malignant syndrome', which is a clinically significant event.

- A single medical concept may be represented by more than one PT and related medical concepts may be distributed in different system organ classes (SOCs). As an example, consider the PT 'hyperkalaemia' under the SOC 'metabolism and nutrition disorders' and the PT 'blood potassium increased' under the SOC 'investigations'. Advantages of this granularity are improved likelihood of capturing the actual event and reduced likelihood of misclassification resulting from the lack of a coding match. It is important to let

signal detection methods identify all disproportionately reported drug-event pairs and then to further investigate all related terms.

- Signals may be generated for an event for which a new MedDRA term was recently introduced. These situations can result in very high relative RRs because the expected value is very low because of the recent inclusion of the term in the database.
- Although both AERS and internal company databases use MedDRA, each organisation has its own coding rules that allow for consistency in data retrieval and data analysis. Different coding rules can profoundly affect signal detection characteristics (also see section 6.4).

7.3 Product Age (Time on Market)

When a drug first receives market authorisation, there is generally a steep increase in spontaneous reporting of AEs that plateaus after a number of years and eventually declines. The chance of a given event ever being reported increases as more data are accrued. There is evidence that as a drug matures, higher proportions of the reported AEs include known reactions and disease-related events.[48] Based on this evidence it is intuitively plausible, but not proven, that the number of new signals detected is likely to reach a peak over time, with a subsequent decline. However, a new dosage regimen or indication for a mature product, or its introduction into a new market, may result in a new pattern of reporting. Therefore, one should be aware of the lifecycle status of the drug in question, as well as the years of introduction to new markets and significant changes in its use.

It is also important to note that the number of spontaneous reports to AERS has increased dramatically since the start of the MedWatch programme in 1993. Although SRS/AERS began in 1968, Working Group members have noted that one-half of the reports in AERS were reported after 1997, with approximately 90% reported since 1990.

7.4 Targeted Surveillance and Stimulated Reporting

AE reporting for a given product may be influenced by many factors, including the initiation or intensity of targeted surveillance activities, selective prescribing (channeling) by physicians and publicity resulting from regulatory activities, litigation or highly publicised studies. Awareness of targeted surveillance and stimulated reporting situations is important when using spontaneous reporting databases for signal detection, regardless of the methods used.

7.4.1 Targeted Surveillance

Targeted surveillance activities include postmarketing epidemiological studies, product registries and surveillance requirements imposed by risk-management programmes. An example of targeted surveillance is the encouraged reporting of inadvertent exposure during pregnancy to two pregnancy category X drugs over several years following their approval[49] (category X is a designation in US product labelling that denotes the potential for fetal harm and contraindicates use during pregnancy). In situations such as this, relative RRs generated via data mining for adverse pregnancy outcomes or complications of pregnancy would need to be viewed in light of the targeted surveillance.

7.4.2 Selective Prescribing (Channeling)

Physicians' prescribing decisions are influenced by a number of factors. A patient's disease characteristics, including severity and prognosis, can influence prescribing, creating the potential for confounded drug-effect associations.[50-55] Prescribing also may be influenced by third-party payer formulary restrictions and by a patient's level of insurance coverage, again creating the potential for confounded drug-event associations.[56,57]

7.4.3 Stimulated Reporting

Publicity resulting from advertising, litigation or regulatory actions (e.g. 'Dear healthcare provider' letters and product withdrawals) may result in increased reporting and can generate higher-than-expected relative RRs.[56,58] Relative RRs should be examined over time in hopes of detecting these influences, although there are no definitive criteria for using data-mining techniques to reliably identify such effects.

8. Vaccines

8.1 Importance of Vaccine Safety

The ethical principle 'first do no harm' (*primum non nocere*) is the basis for the imperative for continuous evaluation of the safety of pharmaceutical products. Several features of vaccination add to this universal principle. Vaccinees are generally healthy and a large number of people are vaccinated, compared with drugs that are generally given to targeted groups of ill individuals. Paediatric vaccinations are often universally recommended or mandated by law, and children are a vulnerable population that needs special protection. Delivering the benefits of vaccination depends on maintaining the public's confidence in vaccine safety with both monitoring for previously unknown adverse effects or increases in known effects and careful analysis of hypothesised vaccine adverse effects.

There are established methods for ensuring the safety of vaccines postlicensure[59] that are beyond the scope of this section to review. Data-mining methods are a relatively new addition to these approaches, so there is a need to carefully consider how vaccines might require special consideration when applying these methods. Although the operational aspects of applying data-mining methods to vaccine AE databases and drug AE databases are identical, interpretation of the outputs of these methods might vary because of intrinsic differences between drugs and vaccines, as well as differences between the vaccine and drug AE databases.

8.2 Product Differences

Data mining in vaccine safety databases is likely to have different characteristics than data mining in drug safety databases because of intrinsic differences between drugs and vaccines. Because of their large preregistration trials, vaccines may be less likely than drugs to have common novel adverse effects emerge in the marketplace. However, the broad populations, with widely varying health profiles and co-morbidities, to which vaccines are administered postmarketing may increase the potential risks of developing rare AEs when compared with drugs, which are often administered as therapy for a single (or relatively narrow) set of conditions. The prophylactic use of most vaccines, versus the therapeutic use of most drugs, imparts a different benefit-risk profile. The pathophysiology of vaccine adverse effects is not as well defined as most drug adverse effects (e.g. hepatotoxicity after paracetamol [acetaminophen]), making it more difficult to use basic toxicological information and clinical judgement to interpret data mining results.

8.3 Concomitantly Administered Vaccines

There are fewer marketed vaccines than drugs, but many vaccines are given in specific combinations, especially during childhood. While technically it is possible to evaluate specific combinations of vaccines and AEs using data-mining methods, the fact that some vaccines are rarely administered or reported to VAERS alone may make it more difficult to distinguish AE associations for individual vaccines than for individual drugs. For example, intussusception is accepted as being caused by rotavirus vaccine, but rotavirus vaccine was usually administered simultaneously with diphtheria, tetanus and acellular pertussis (DTaP) vaccine, leading to DTaP being falsely signaled as being associated with intussusception (see section 6.7 regarding 'signal absorption'). Additional information from traditional methods of safety surveillance is needed to resolve such issues.

In data-mining analyses of vaccine safety data, we are attempting to identify associations between vaccines and AE coding terms. We know that vaccines are administered according to patient age (e.g. children receive 7-valent pneumococcal vaccines, whereas adults generally do not) and that the spectrum of AEs that occur in children is different than in adults (e.g. sudden infant death syndrome [SIDS] is limited by definition to infants, adults develop lung cancer). These patterns will influence the vaccine-event pairs that are reported to VAERS. Similarly, vaccines may be administered disproportionately by sex (e.g. more women may receive hepatitis B vaccine because of their status as healthcare workers) and disease patterns may differ in men and women (e.g. women experience autoimmune conditions more often than men). It is important that estimates of disproportionality be calculated based on a comparison in groups that have a similar likelihood of receiving similar vaccines and experiencing

similar AEs. This approach helps to prevent vaccine-AE pairs from being signaled because of differences in the underlying populations, rather than true differences in reporting of the AE from similar populations (e.g. DTaP-SIDS PRR misleadingly elevated because the comparison products are given to adults who do not routinely receive DTaP or die from SIDS). In the analysis, one can control for such confounding either by partitioning the data into like groups (e.g. only adults) or by stratification.

8.4 Validation of Data-Mining Methods

Although there are no 'gold standards' for the detection of vaccine-AE associations that can be used to precisely calculate sensitivity and specificity of a particular method, several surrogate measures of adverse effects have been proposed. In addition to product labelling, the Institute of Medicine (IOM) has conducted systematic reviews of vaccine AEs since the late 1980s and provided a list of AEs that they determined to be caused or not caused by vaccination.[60] In the absence of a gold standard, the IOM reviews might provide useful surrogate vaccine-AE pairs on which to retrospectively gauge the performance of various data-mining methods. However, the large number of vaccine-AE pairs for which a determination of causality has not been made and the continual improvement of knowledge about vaccine adverse effects limit our ability to precisely define sensitivity and specificity of these methods. New vaccines are continually introduced, older vaccines are used in new ways (e.g. smallpox vaccine to counter bioterrorism) and reporting patterns change over time; therefore, validation of the usefulness of these methods will ultimately depend on prospective application and successful early detection of an important new signal or signals.

9. Integrating Quantitative Signal Detection and Traditional Pharmacovigilance Methods

9.1 Overview

The process of signal detection in the postmarketing environment is both qualitative (e.g. clinical and scientific judgement) and quantitative. Traditional methods of signal detection and evaluation involve literature searching and case-by-case analysis, as well as crude frequency counts and calculation of reporting rates. The newer quantitative methods involving data-mining techniques are reviewed in section 4. Institutions considering the use of these newer methods should consider how to integrate them with traditional pharmacovigilance methods.[61]

Traditional methods alone are generally satisfactory when the volume of data is manageable. When the number of reports exceeds traditional signal-evaluating resources, combining traditional and data-mining methods may be considered. The choice of whether or not to employ data-mining methods should be evaluated by each institution since the added value of these methods is likely to be highly situation dependent. Among the many factors to consider are: (i) the rigor of existing signal detection practices/protocols based on clinical and scientific judgement; (ii) timelines; (iii) internal domain expertise of drugs and databases; (iv) availability and validation status of newer signal-detection methods; (v) availability and quality of comparative databases; and (vi) the uncertainties that remain about the predictive performance of these methods and databases through time.

As shown in figure 1, the process of signal detection can be initiated by selecting one or more traditional and/or data-mining methods. For example, one can begin the process by using the PRR method and complementing it with case-by-case analysis for signal detection. If the choice is made to use one or more data-mining methods, they should be used as a supplement to traditional methods. It should be noted that traditional methods can reveal safety signals that are otherwise not detected by data-mining methods and thus data mining should not be relied upon as a substitute for traditional methods, particularly with rare events or designated medical events.

If a signal is detected, it is important to evaluate the signal by conducting cumulative case review, literature review, assessment of preclinical and pharmacological data and, if appropriate, pharmacoepidemiological and clinical studies to assess causality. If a signal is not detected or is detected but not verified, then one needs to monitor and periodically repeat the process of safety signal detection or refute and 'close out' the signal if appropriate. Regardless

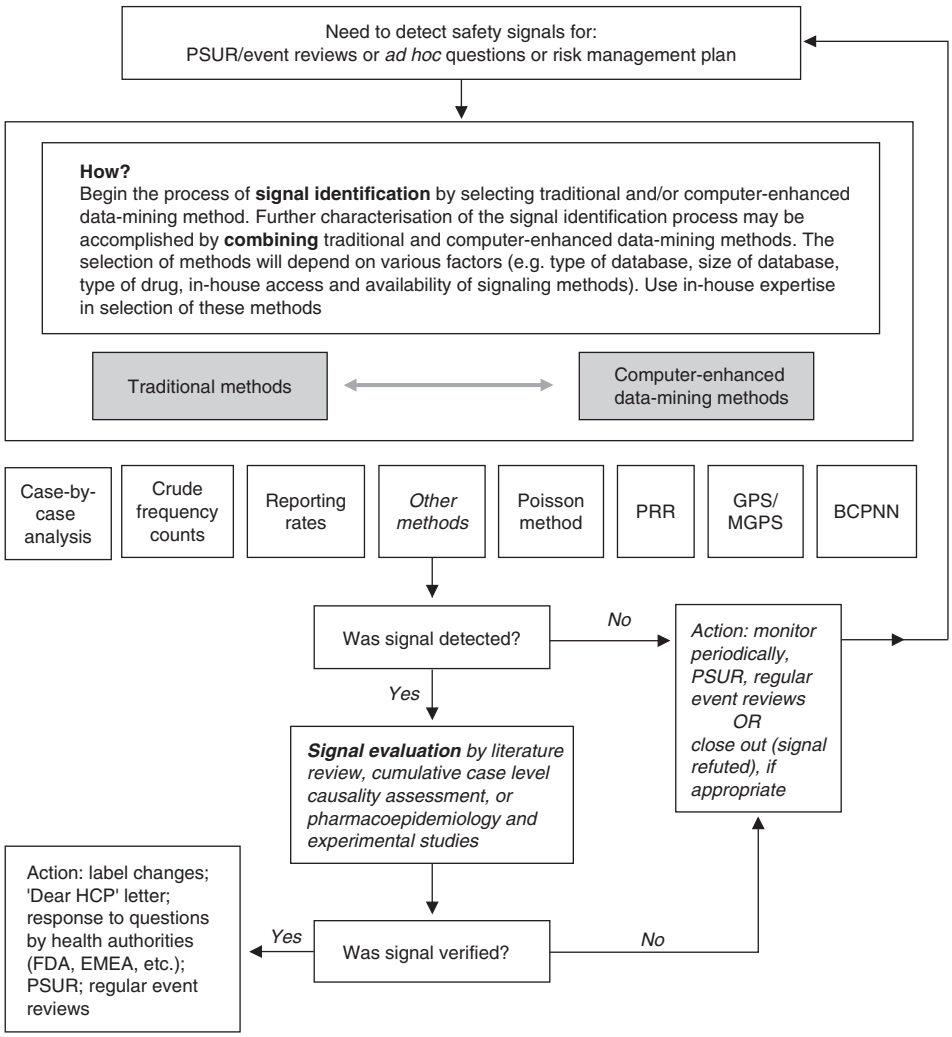

Fig. 1. Integrating computer-enhanced data-mining methods and traditional pharmacovigilance methods in process for signal detection. **BCPNN** = Bayesian confidence propagation neural network; **EMEA** = European Medicines Agency; **GPS** = gamma Poisson shrinker; **HCP** = healthcare provider; **MGPS** = multi-item gamma Poisson shrinker; **PRR** = proportional reporting ratio; **PSUR** = periodic safety update report.

of the method(s) used, the frequency of conducting proactive signal detection is highly dependent on the product and the potential safety issues involved.

In summary, disproportionality methods are not intended to be used in isolation. When these methods are appropriately incorporated into a comprehensive pharmacovigilance programme, clinical judgement and domain expertise should significantly mitigate the impact of false-positive and false-negative signals.

9.2 US FDA Perspective

Traditionally, a signal is generated by a question from the reviewing division, by a publication or by a safety reviewer's judgement based on the number and/or seriousness of reports of an AE for a particular drug. To confirm the observation, safety evaluators use a variety of approaches. Initially, the reviewers retrieve 'raw' numbers of reported cases to provide some perspective on the number of times an

event has been reported for a specific drug. A review of the medical literature may also be done. A 'hands on' review of each report is necessary to eliminate duplicate reports. A crude reporting rate can then be calculated by counting the number of reports of the AE in individual patients exposed to the drug and then dividing by the estimated number of prescriptions for the drug. The reporting rate (which should not be confused with an incidence rate) may be compared with an expected rate in the general population, but often such expected rates are difficult to ascertain.[3]

The empirical Bayesian data mining algorithm was initially implemented in February 1998.[1] In 2002, the FDA entered into a formal Cooperative Research and Development Agreement (CRADA) with a private advanced computer technology firm to collaborate in the development of a data-mining software application (MGPS) for use by safety evaluators, epidemiologists and medical officers at the FDA. When piloting of this system began in March 2003, various issues, including the following, were raised.

- Validity of approach: it was initially thought by some evaluators accustomed to a case-by-case review approach that applying empirical Bayesian methods to a database containing spontaneously submitted reports would not provide an accurate representation of a drug's potential association for AEs. Although some scepticism has diminished among these evaluators, the absolute interpretation of these results continues to pose challenges. These challenges in interpretation served as an important impetus in the formation of this Working Group.
- Lack of guidelines for interpretation: some safety evaluators and epidemiologists stated that they had difficulty in interpreting data-mining outputs because there were no standard guidelines for interpretation. For example, the definition of a signal may be dependent on several factors, including the AE(s) in question, the indication of a drug and the data being analysed (e.g. fatal outcomes). The signal threshold for a drug indicated for a serious disease with few, if any, treatment options may be higher than the threshold for a drug indicated for non-serious conditions with many treatment options. Thus, thresholds for action may be variable.[2,62]
- Added value of data mining: when considering the addition of a data-mining component to an already existing postmarketing surveillance group, questions naturally arose about whether the benefits associated with data mining outweigh its costs (e.g. economic, impact on public health). Indeed, the benefits of data mining can be difficult to quantify in any objective way. For example, the use of data mining is presumed to make postmarketing safety surveillance more efficient. As previously discussed, it is difficult to establish positive or negative predictive values with data mining. It is also difficult to define prospectively what a success with data mining would be and the quantity and quality of evidence needed to formulate a decision on whether data mining should be incorporated into an organisation's pharmacovigilance practices. Nonetheless, these methods do have some advantages over conventional clinical and epidemiological techniques. Because they are computer based, many analyses that would be difficult or impossible to do by standard methods can be carried out conveniently. This includes subsetting of the data, stratification, examination of the evolution of a signal over time and efficiently drilling down to individual reports.
- Added use of data mining: as part of its regulatory responsibility to monitor the potential toxicity of all marketed products, the FDA periodically examines drugs with similar chemical structures but different indications (e.g. α_1-blockers), drugs with different structures but the same indications (e.g. analgesics) and drugs with (or without) a specific, well established toxicity discovered by traditional methods (e.g. hepatotoxicity) to assure that no drug presents a previously unsuspected risk of important AEs noticeably higher than that presented by other drugs. The FDA is currently evaluating a screening approach that compares the confidence interval for the empirical Bayesian estimate of the RR for a suspect compound with the confidence intervals for multiple other 'control' drugs over successive intervals of time. Although the comparisons are formally equivalent to hypothesis tests, in fact the

results are not (and cannot be) interpreted as comparisons between treatments for reasons that are articulated in section 7.1. Instead, they are intended to provide a way to identify compounds for which further follow-up is needed to elucidate the reasons for the apparently elevated RRs. Experience so far is limited, but this approach may have merit as a regulatory screening tool, recognising the inherent problems in comparing RRs.

In 1999, researchers at the FDA/Center for Biologics Evaluation and Research retrospectively applied the empirical Bayesian method to determine when intussusception (a documented adverse reaction to rotavirus vaccine) showed association with rotavirus vaccine in the VAERS database.[63] They found that the empirical Bayesian method was able to detect the signal when only four cases of intussusception had been reported, which suggested the potential usefulness of this method to enhance vaccine safety. Evaluation of the empirical Bayesian and PRR methods in VAERS showed that both methods could contribute to vaccine safety data mining.[64] Application of the PRR method for surveillance of AEs after typhoid vaccines contributed to the detection of atypical allergic reactions after typhim Vi vaccine[65] and photophobia after smallpox vaccine.[66] The CDC also applied PRR methods in their evaluation of Bell's palsy after influenza vaccine,[67] although this was more controversial. In this study, the signal for Bell's palsy was generated independently of the VAERS data and the investigators used the increased PRR for Bell's palsy after influenza vaccines in VAERS, among other lines of evidence, to support the need for further evaluation. In an accompanying editorial, Shapiro[68] criticised this use of a data-mining method because, in his opinion, traditional clinical and epidemiological evaluation of the underlying case reports revealed sufficient limitations to undermine the conclusion.[69] The integration of traditional safety surveillance and data-mining methods for vaccine safety is an area that requires refinement, and delineating key concepts in applying data-mining methods to vaccine AE databases is an important step in this process.

9.2.1 Overall Lessons Learned

Data mining of surveillance systems may assist in identifying possible signals, but additional review and scientific investigations are always required to validate the signal and establish or rule out a causal relationship between a product and an AE. The absence of elevated relative RRs does not rule out a safety problem. Electronic pharmacovigilance systems assist but do not replace the acuity of knowledgeable safety evaluators and medical reviewers.

9.3 Industry Perspective from Pharmaceutical Research and Manufacturers of America Working Group Members

There are currently no regulatory or scientific requirements to use data mining for signal detection nor is there a single recommended approach to signal detection by regulatory authorities and pharmaceutical companies. However, the FDA recently issued guidances describing good practices for pharmacovigilance and pharmacoepidemiological assessment that discuss, among other things, potential roles for data mining in evaluating drug safety based on spontaneous postmarketing reports.[62] Before implementing any of these data-mining methods, a company should take a critical look at their current pharmacovigilance practices to determine what complementary methods might be needed.

If a decision is made to employ data-mining methods, it is very important to educate all members of a drug safety organisation, as well as others outside of drug safety, as to the strengths and limitations of the methods and of the spontaneous reporting databases themselves. People tend to want to draw very broad conclusions from outputs of data-mining analyses. It is important to emphasise that these methods are intended for screening databases, generating hypotheses and helping set priorities for review of reported AEs.

It is also advisable to develop transparent processes for data-mining activities that are consistent with company standard operating procedures (SOPs) and used consistently no matter what methods are implemented. At present, Working Group members are not aware of local or international regulations that cover data-mining processes. As data-mining methods evolve, SOPs may need to be updated. From a legal perspective, signal detection and follow-up of signals are likely to be discovera-

ble in litigation. It is therefore advisable to emphasise their preliminary and non-conclusive status, as well as to follow prudent document management guidelines once final decisions are made regarding signal validity.

Pharmacovigilance practitioners should periodically evaluate the effectiveness of their current procedures and carefully consider whether any additional methods, such as those discussed in this paper, could enhance their pharmacovigilance practices. Potential users should be encouraged to perform their own evaluations, not only to identify potential areas for improvement but also to contribute to further understanding of these methods, thereby promoting optimum use and minimising misuse or misapplication.

10. Other Uses of Data Mining

10.1 Comparing 'Signal Scores' Across Products

It is tempting to compare signal scores at some level and it is easy to construct various statistics for this purpose. However, differences between RRs do not imply differences in risk because spontaneous reporting databases are biased in ways that cannot be measured or controlled. It is not legitimate to infer that differences between scores imply differences between treatments without carefully considering the mechanisms that generate reports, including the known and unknown biases.

10.2 Evaluation of Potential Drug Interactions and Medical Syndromes

The principles of disproportionality may be applied to the detection of drug-drug interactions. Two approaches to analysing the effect of specific drug combinations on a predefined adverse effect of interest have been described. Both methods were tested using well known examples of drug-drug interactions. One approach involves the use of logistic regression modelling to evaluate statistical interactions amongst various therapies.[70] A second approach involves executing disproportionality analysis on therapy combinations of interest, where each combination is analysed as a unique drug variable.[71] In the latter approach, two potentially interacting drugs are analysed by creating three drug category variables: one category corresponds to all reports that mention both of the potentially interacting drugs and the other two categories correspond, respectively, to reports that mention either the first drug or the second, but not both. The use of multidimensional data-mining strategies that simultaneously screen for frequent 'drug-event-event' combinations may provide a fruitful approach to identifying syndromes with multiple AEs that are associated with the use of a drug.[72]

10.3 Evaluation of Demographic and Treatment-Related Factors

Some members of the Working Group have found it useful to conduct disproportionality analyses on subsets of a database, where the subset is defined by variables such as age, sex, report source or year of report. Such partitioning of the database may facilitate analyses involving specific populations of interest (e.g. females, paediatric patients).[1,73,74] For example, if an analysis is to be done on females, a subset of the database is created that only contains cases describing female patients and the data-mining algorithm is run on this data subset. Database subsets can also be used to examine the evolution of relative RRs over time. Subsets of the database are created based on report date and reporting statistics (e.g. RR or Empirical Bayes Geometric Mean) are computed for discrete or cumulative periods of time, typically year or quarter, depending on the age of the drug.[2,18] It may also be possible to compute signal scores for a particular drug according to different doses or dose ranges (if the data are available) by configuring a single drug name variable as multiple drug name variables, each of which represents a unique dose or dose range. Repeated analyses with multiple subsets generally should be avoided.

10.4 Restriction or Customisation of Database Backgrounds

It is possible to perform disproportionality analyses using customised or restricted 'backgrounds' rather than the entire database. One use of this technique is possibly to enhance the detection of signals specific to a particular drug within a drug

class. For example, if a drug class is generally associated with a particular event, a disproportionality algorithm could be run for all drugs in the class using only those drugs as the background. Another use is to assess the relative reporting of a drug-event pair with respect to a population of interest.

Changing the background alters the expected counts and therefore the relative RRs for events of interest. These approaches have not been studied in a systematic way; thus, there is no information on minimum background size or predictive utility. The Working Group agrees with the recommendation of Gogolak[75] that such analyses should be done in parallel with analyses that use the entire background dataset.

11. Summary and Recommendations

It was the goal of the Working Group to provide insights into both the potential utilities and limitations of data-mining methodologies and their role in pharmacovigilance. The following is a summary of key points and recommendations.

- Quantitative signal detection is a potentially useful adjunct to traditional pharmacovigilance practices.
- Data-mining methods analyse *relative reporting* of AEs. Hence, they cannot replace careful medical and scientific evaluation and should be considered as potential supplements to, not substitutes for, traditional signal-detection strategies.
- Data-mining results are hypothesis generating. They should be evaluated only in the context of other relevant data.
- The Working Group is not aware of any regulatory requirements for the use of data-mining methods in pharmacovigilance. Although there is evidence that data mining may be useful, the evidence is not sufficient to fully judge the value of data mining in pharmacovigilance. Time and experience will reveal its value and utility. Individual institutions may wish to evaluate the available methods and determine if data mining could provide value to their pharmacovigilance efforts.
- Statistical refinements that improve the capabilities of data-mining methods (e.g. adjusting for polytherapy and signal-masking effects) should further enhance their usefulness.
- The importance of vaccine safety as well as differences between vaccine and drug safety surveillance warrant special attention to data mining in vaccine AE databases. Continued efforts should be made to determine the best data-mining practices to enhance vaccine safety surveillance.
- A universal 'canonical' database, containing up-to-date information, for use in monitoring drug safety is highly desirable.

Acknowledgements

The Working Group acknowledges the participation and help of Rosanne Ososki, Lesley-Anne Furlong, Cheryl Watton, Dionigi Maladorno and Min Chu Chen. The Working Group also thanks Miles Braun and Paul Seligman for their support of this collaboration and for their helpful reviews of the manuscript.

We acknowledge PhRMA for funding technical support for preparation of this manuscript. A number of the authors are employed by pharmaceutical companies, as described in their respective affiliations.

References

1. O'Neill RT, Szarfman A. Some FDA perspectives on data mining for pediatric safety assessment. Curr Ther Res Clin Exp 2001; 62 (9): 650-63
2. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf 2002; 25 (6): 381-92
3. Szarfman A, Tonning JM, Doraiswamy PM. Pharmacovigilance in the 21st century: new systematic tools for an old problem. Pharmacotherapy 2004; 24 (9): 1099-104
4. Hauben M. Early postmarketing drug safety surveillance: data mining points to consider. Ann Pharmacother 2004; 38: 1625-30
5. Wolkenstein P, Latarget J, Roujeau J, et al. Randomised comparison of thalidomide versus placebo in toxic epidermal necrolysis. Lancet 1998; 352: 1586-9
6. Edwards IR, Biriell C. Harmonization in pharmacovigilance. Drug Saf 1994; 10 (2): 93-102
7. Lindquist M, Edwards IR, Bate A, et al. From association to alert: a revised approach to international signal analysis. Pharmacoepidemiol Drug Saf 1999; 8: S15-25
8. Report of CIOMS Working Group VI. Management of safety information from clinical trials. Geneva: Council for International Organization of Medical Sciences (CIOMS), 2005
9. Evans SJW, 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
10. Gould AL. Practical pharmacovigilance analysis strategies. Pharmacoepidemiol Drug Saf 2003; 12: 559-74
11. Hauben M. A brief primer on automated signal detection. Ann Pharmacother 2003; 37: 1117-23
12. Hauben M, Zhou X. Quantitative methods in pharmacovigilance: focus on signal detection. Drug Saf 2003; 26 (3): 159-86

13. Wilson AM, Thabane L, Holbrook A. Application of data mining techniques in pharmacovigilance. Br J Clin Pharmacol 2003; 57 (2): 127-34
14. Evans SJ. Pharmacovigilance: a science or fielding emergencies? Stat Med 2000; 19 (23): 3199-209
15. Egberts AC, Meyboom RH, van Puijenbroek EP. Use of measures of disproportionality in pharmacovigilance: three Dutch examples. Drug Saf 2002; 25 (6): 453-8
16. van der Heijden PGM, van Puijenbroek EP, van Buuren S, et al. On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of under-reporting on odds ratios. Stat Med 2002; 21: 2027-44
17. van Puijenbroek EP, Bate A, Leufkens HGM, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002; 11: 3-10
18. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998; 54: 315-21
19. Lindquist M, Stahl M, Bate A, et al. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. Drug Saf 2000; 23 (6): 533-42
20. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat 1999; 53 (3): 177-202
21. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. In: Conference on Knowledge Discovery in Data. Proceedings of the seventh ACM SigKDD International Conference on Knowledge Discovery and Data Mining. 2001 Aug 26–29; San Francisco (CA). New York: ACM Press, 2001: 67-76
22. Council for International Organizations of Medical Sciences (CIOMS). Guidelines for Preparing Core Clinical-Safety Information on Drugs. 2nd ed. Report of CIOMS Working Groups III and V. Geneva: Council for International Orgnizations of Medical Sciences (CIOMS), 1999: 27-33
23. Purcell P, Barty S. Statistical techniques for signal generation: the Australian experience. Drug Saf 2002; 25 (6): 415-21
24. Mamedov MA, Saunders GW. Fuzzy set analysis of Australian drug safety data. Proceedings of HIC 2002: Tenth National Health Informatics Conference; 2002 Dec 4–6, Melbourne
25. Mamedov MA, Saunders GW, Yearwood J. A fuzzy derivative approach to classification of outcomes from the ADRAC database. International Transactions in Operational Research 2004; 11 (2): 169-79
26. Spiegelhalter D, Grigg O, Kinsman R, et al. Risk adjusted sequential probability ratio tests: application to Bristol, Shipman and adult cardiac surgery. Int J Qual Health Care 2003; 15: 7-13
27. Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. Stat Methods Med Res 2003; 12 (2): 147-70
28. Kulldorf M, Fang Z, Walsh SJ. A tree-based scan statistic for database disease surveillance. Biometrics 2003; 59: 323-31
29. Hauben M, Reich L. Drug-induced pancreatitis: lessons in data mining [letter]. Br J Clin Pharmacol 2004; 58 (5): 560-2
30. Waller P, Heeley E, Moseley J. Impact analysis of signals detected from spontaneous adverse drug reaction reporting data. Drug Saf 2005; 28 (10): 843-50
31. Waller PC, Heeley EL, Moseley JNS. Impact analysis of signals detected from spontaneous adverse reaction reporting data [abstract]. Pharmacoepidemiol Drug Saf 2004; 13: S323
32. Heeley E, Waller P, Moseley J. Testing and implementing signal impact analysis in a regulatory setting: results of a pilot study. Drug Saf 2005; 28 (10): 901-6
33. Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. Pharmacoepidemiol Drug Saf 2004; 13: 519-23
34. Waller P, van Puijenbroek E, Egberts A, et al. The reporting odds ratio versus the proportional reporting ratio: 'deuce' [letter]. Pharmacoepidemiol Drug Saf 2004; 13: 525-6
35. Stahl M, Lindquist M, Edwards IR, et al. Introducing triage logic as a new strategy for the detection of signals in the WHO drug monitoring database. Pharmacoepidemiol Drug Saf 2004; 13: 355-63
36. Begaud B, Moride Y, Tubert-Bitter P, et al. False-positive in spontaneous reporting: should we worry about them? Br J Clin Pharmacol 1994; 38 (5): 401-4
37. Hauben M. Application of an empiric Bayesian data mining algorithm to reports of pancreatitis associated with atypical antipsychotics. Pharmacotherapy 2004; 24 (9): 1122-9
38. Hauben M. Trimethoprim-induced hyperkalaemia-lessons in data mining [letter]. Br J Clin Pharmacol 2004; 58 (3): 338-9
39. Hauben M, Reich L. Safety related drug-labelling changes: findings from two data mining algorithms. Drug Saf 2004; 27 (10): 735-44
40. Levine JG, Tonning JM, Szarfman A. Reply: the evaluation of data mining methods for the simultaneous and systematic detection of safety signals in large databases. Lessons to be learned [letter]. Br J Clin Pharmacol. In press
41. Hauben M, Reich L. Response to letter by Levine et al. [letter]. Br J Clin Pharmacol. In press
42. Lindquist M. The WHO adverse reaction database: basic facts [online]. Available from URL: http://www.who-umc.org/pdfs/WHO%20Adverse%20Reaction%20Database%20basic%20facts.pdf [Accessed 2004 Sep 14]
43. European Medicines Agency. EudraVigilance [online]. Available from URL: http://www.eudravigilance.org/highres.htm [Accessed 2005 Sep 17]
44. Cosentino M, Leoni O, Michielotto D, et al. Increased reporting of adverse reactions to ACE inhibitors associated with limitations to drug reimbursement for angiotensin-II antagonists. Eur J Clin Pharmacol 2001; 57: 509-12
45. Bate A, Edwards RI, Lindquist M, et al. The authors' reply [letter]. Drug Saf 2003; 26 (5): 364-6
46. Szarfman A, DuMouchel W, Fram D, et al. Lactic acidosis: unraveling the individual toxicities of drugs used in HIV and diabetes polytherapy by hierarchical Bayesian logistic regression data mining [abstract]. 11th Annual FDA Science Forum, 2005 Apr 27-28 [online]. Available from URL: http://www.cfsan.fda.gov/~frf/forum05/H-30.htm [Accessed 2005 Sep 14]
47. Brown EG. Effects of coding dictionary on signal generation: a consideration of use of MedDRA compared with WHO-ART. Drug Saf 2002; 25 (6): 445-52
48. Haramburu F, Begaud B, Moride Y. Temporal trends in spontaneous reporting of unlabelled adverse drug reactions. Br J Clin Pharmacol 1997; 44: 299-301
49. Manson JM, Freyssinges C, Ducrocq MB, et al. Postmarketing surveillance of lovastatin and simvastatin exposure during pregnancy. Reprod Toxicol 1996; 10 (6): 439-46
50. Blais L, Ernst P, Suissa S. Confounding by indication and channeling over time: the risks of beta 2-agonists. Am J Epidemiol 1996; 15 (12): 1161-9
51. Blais L, Ernst P, Suissa S. The authors' reply [letter]. Am J Epidemiol 1997; 146 (10): 886-7
52. Leufkens HG. Pharmacoepidemiology and gastroenterology: a close couple. Scand J Gastroenterol Suppl 2000; 232: 105-8
53. Leufkens H, Urquhart J. Variability in patterns of drug usage. J Pharm Pharmacol 1994; 46 Suppl. 1: 433-7
54. Meijer WEE, Heerdink ER, Pepplinkhuizen LP, et al. Prescribing patterns in patients using new antidepressants. Br J Clin Pharmacol 2001; 51: 181-3

55. Pearce N, Beasley R, Crane J, et al. Confounding by indication and channeling over time: the risks of beta-2 agonists [letter]. Am J Epidemiol 1997; 146 (10): 885-6

56. deBruin ML, van Puijenbroek EP, Egberts ACG, et al. Non-sedating antihistamine drugs and cardiac arrhythmias: biased risk estimates from spontaneous reporting systems? Br J Clin Pharmacol 2002; 53: 370-4

57. Tisonova J, Szalayova A, Kriska M. Factors influencing the spontaneous reporting of adverse drug reactions: the experience of the Slovak Republic. Pharmacoepidemiol Drug Saf 2003; 13: 333-7

58. Coster TS, Szarfman A, Tonning J. The application of data mining to analyze pre-publicity psychiatric signals with the use of mefloquine [abstract]. ASCPT Annual Meeting; 2004 Mar 4; Miami (FL)

59. Varricchio F, Iskander J, Destefano F, et al. Understanding vaccine safety information from the vaccine adverse event reporting system (VAERS). Pediatr Infect Dis J 2004; 23: 287-94

60. Institute of Medicine. Immunization safety review [online]. Available from URL: http://www.iom.edu/project.asp?.id=4-705 [Accessed 2005 Sep 26]

61. Yee CL, Klincewicz SL, Knight JF, et al. Practical considerations in developing an automated signaling program within a pharmacovigilance department. Drug Inf J 2004; 38: 293-300

62. US FDA. Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment. US Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, March 2005 [online]. Available from URL: http://www.fda.gov/cder/guidance/6359OCC.htm#_Toc48124197 [Accessed 2005 Sep 27]

63. Niu MT, Erwin DE, Braun MM. Data mining in the US vaccine adverse event reporting system: early detection of intussusception and other events after rota virus vaccine. Vaccine 2001; 19: 4627-34

64. Banks D, Woo EJ, Burwen D, et al. Comparison of 4 data mining methods in the US Vaccine Adverse Event Reporting System (VAERS) [abstract]. Pharmacoepidemiol Drug Saf 2003; 12 Suppl. 1: S138

65. Begier EM, Burwen D, Haber P, et al. Post-marketing safety surveillance for typhoid fever vaccines from the Vaccine Adverse Event Reporting System, July 1990-June 2002. Clin Infect Dis 2004; 38: 771-9

66. McMahon AW, Bryant-Genevier MC, Woo EJ, et al. Photophobia following smallpox vaccination [letter]. Vaccine 2005; 23: 1097-8

67. Zhou W, Pool V, DeStefano F, et al. A potential signal of Bell's palsy after parenteral inactivated influenza vaccines: reports to the vaccine adverse event reporting system (VAERS): United States, 1991-2001. Pharmacoepidemiol Drug Saf 2004; 13: 505-10

68. Shapiro S. Clinical judgment, common sense and adverse reaction reporting. Pharmacoepidemiol Drug Saf 2004; 13: 511-3

69. Zhou W, Pool V, DeStefano F, et al. Reply to the editorial. Pharmacoepidemiol Drug Saf 2004; 13: 515-7

70. van Puijenbroek EP, Egberts ACG, Heerdink ER, et al. 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 2000; 56: 733-8

71. Almenoff JS, DuMouchel W, Kindman A, et al. 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 (6): 517-21

72. Szarfman A. Syndromic surveillance and risk management using multi-item gamma Poisson shrinker. Journal of Urban Health: bulletin of the New York Academy of Medicine 2003; 80 (2 Suppl. 1): i133 [online]. Available from URL: http://www.syndromic.org/syndromicconference/2002/Supplement-pdf/Abstracts_SectionIV.pdf [Accessed 2005 Sep 15]

73. Yuen NA, Almenoff JS, DuMouchel W, et al. Disproportionality analysis to explore patient and treatment related factors associated with adverse events [abstract]. Pharmacoepidemiol Drug Saf 2004; 13: S259

74. Szarfman A. Gender-related 'higher-than-expected' drug-event combinations in spontaneous adverse drug event reports [abstract no. D05]. 2000 FDA Science Forum – FDA and the science of safety: new perspectives; 2000 Feb 14-15; Washington, DC [online]. Available from URL: http://vm.cfsan.fda.gov/~frf/forum00/d05.htm [Accessed 2004 Oct 12]

75. Gogolak VV. The effect of backgrounds in safety analysis: the impact of comparison cases on what you see. Pharmacoepidemiol Drug Saf 2003; 12: 249-52

Correspondence and offprints: Dr *June Almenoff*, Global Clinical Safety and Pharmacovigilance, GlaxoSmithKline, 5 Moore Drive, Mail Stop 5.4214.4C, PO Box 13398, Research Triangle Park, NC 27709-3398, USA.
E-mail: june.s.almenoff@gsk.com