ORIGINAL RESEARCH ARTICLE

Drug Safety 2007; 30 (2): 143-155
0114-5916/07/0002-0143/$44.95/0

Potential Use of Data-Mining Algorithms for the Detection of 'Surprise' Adverse Drug Reactions

Manfred Hauben,[1,2,3] *Sebastian Horn*[4] and *Lester Reich*[1]

1 Risk Management Strategy, Pfizer Inc., New York, New York, USA
2 Department of Medicine, New York University School of Medicine, New York, New York, USA
3 Departments of Pharmacology and Community and Preventive Medicine, New York Medical College, Valhalla, NY, USA
4 Corporate Drug Safety, Boehringer Ingelheim GmbH, Ingelheim, Rheinland-Pfalz, Germany

Abstract

Background and objective: Various data mining algorithms (DMAs) that perform disporportionality analysis on spontaneous reporting system (SRS) data are being heavily promoted to improve drug safety surveillance. The incremental value of DMAs is ultimately related to their ability to detect truly unexpected associations that would have escaped traditional surveillance and/or their ability to identify the same associations as traditional methods but with greater scientific efficiency. As to the former potential benefit, in the course of evaluating DMAs, we have observed what we call 'surprise reactions'. These adverse reactions may be discounted in manual review of adverse drug reaction (ADR) lists because they are less clinically dramatic, less characteristic of drug effects in general, less serious than the classical type B hypersensitivity reactions or may have subtle pharmacological explanations. Thus these reactions may only become recognised when *post hoc* explanations are sought based on more refined pharmacological knowledge of the formulation. The objective of this study was to explore notions of 'unexpectedness' as relates to signal detection and data mining by introducing the concept of 'surprise reactions' and to determine if the latter associations, often first reported in the literature, represent a type of ADR amenable to detection with the assistance of adjunctive statistical calculations on SRS data.

Methods: Using commonly cited thresholds, the multi-item gamma Poisson shrinker (MGPS) and proportional reporting ratios (PRRs) were applied to reports in the US FDA Adverse Event Reporting System (AERS) database of well documented 'surprise reactions' compiled by the authors.

Results: There were 34 relevant surprise reactions involving 29 separate drugs in 17 different drug classes. Using PRRs (PRR >2, χ^2 >4, N >2), 12 drug-event combinations were signalled before the first ADR citation appeared in MEDLINE, four occurred concurrently and 11 after. With empirical Bayes geometric mean (EBGM) analysis (EBGM >2, N >0), 12 signals occurred before, three concurrently and 11 after publication of the first literature citation. With EB_{05}

(EB_{05} ≥2, N >0), six occurred before, two concurrently and 14 after MEDLINE citation.

Discussion: Pharmacovigilance is rather unique in terms of the number and variety of events under surveillance. Some events may be more appropriate targets for statistical approaches than others. The experience of many organisations is that most statistical disproportionalities represent known associations but our findings suggest there could be events that may be discounted on manual review of adverse event lists, which may be usefully highlighted via DMAs.

Conclusions: Identification of surprise reactions may serve as an important niche for DMAs.

Background

A principle concern of pharmacovigilance is the detection of adverse reactions that are novel by virtue of their nature, severity and/or frequency as soon as possible after licensing with minimum patient exposure after widespread clinical use. This is an extremely challenging task and can aptly be characterised (in terms similar to those applied to medical decisions making) as both an art and a science of making acceptable decisions in an imperfectly understood problem space using incomplete and often erroneous information.

The safety profile of a medicine unfolds over time. David Finney,[1] who originally delineated numerical approaches to spontaneous reporting system (SRS) data said "the essence is to collect facts that individually tell little, but collectively form a clue to drug dangers". Accordingly, pharmacovigilance employs multiple techniques, tools and datasets to monitor this evolution. Despite well known limitations of each of the datasets and techniques, they hopefully compensate for each other's deficiencies and ultimately help illuminate the 'developmental anatomy' of a medicine's safety profile.

Faced with vast and growing databases of spontaneously reported adverse drug reactions (ADRs), safety reviewers find themselves drowning in data but thirsty for knowledge. This represents a strain on the traditional and largely heuristic approaches to signal detection based on the scientific acumen of 'the prepared mind'.[2] It is therefore understandable that health authorities, drug monitoring centres and pharmaceutical companies are studying computational signal detection algorithms, also know as data mining algorithms (DMAs), as potential adjuncts to standard approaches to signal detection, although the prepared mind remains the cornerstone of signal detection and evaluation.[3]

The most commonly used DMAs project high dimensional SRS data onto 2×2 contingency tables to calculate a pseudo-observed-to-expected ratio that reflects how distinctive the reporting experience is for each reported drug-event combination relative to the background reporting experience across all/most drugs and reactions as reflected in an independence model.[3] However, data mining in pharmacovigilance is a dynamic field and there are additional tools under development or being tested, including algorithms based on correlation analysis, sequential probability ratio testing and multivariate regression.[3]

As the authors have gained more experience in data mining, we have come to realise that DMAs can be used to assist in the initial detection of signals and/or can be used as one of many pieces of information to refine an index of suspicion attached to a signal identified by other means.

It is not uncommon to hear rather 'fuzzy' statements that DMAs are 'useful' but to move the field forward it is important to have clear benchmarks by which to assess the utility of these tools.[3] Briefly, DMAs may help to (i) detect ADRs that would have escaped detection with traditional approaches; (ii) identify the same ADRs that would have been identified with traditional approaches but at earlier time points; (iii) identify the same ADRs at the same time as traditional methods but with greater scientific

efficiency (i.e. decreased person-time per credible signal detected); and/or (iv) as an additional safety net.[3]

The focus on the data mining aspects of this paper is on the first two benchmarks. These two benchmarks are related in another somewhat vague statement that we have heard on multiple occasions, i.e. the greatest value of DMAs is in directing attention to associations that are truly 'unexpected', in the sense that the mere reporting of the association would not trigger a cognitive process leading to an hypothesis. But which associations might be unexpected to the pharmacovigilance expert in this context? We make no pretence to formulate foolproof criteria of what is 'unexpected' because this is obviously contingent on numerous factors but we offer a few relevant considerations. We emphasise that we are exploring 'unexpectedness' in specific scientific context and not in the context of product labelling.

Traditionally, ADRs have been categorised by (i) general pathomechanism (i.e. type A reactions, which can be explained by the pharmacological action of the drug, or type B reactions, which describe immune system or idiosyncratic ADRs); (ii) serious versus non-serious reactions; or (iii) by body system. The latter two categories have been used for regulatory submissions (i.e. Periodic Safety Update Reports, Annual Reports, MedWatch/CIOMS forms). Parenthetically, Aronson and Ferner[4] recently published a cogent critique of the category classification and offer alternatives.

As for unexpectedness, some types of ADRs are generally not considered *a priori* unexpected, at least prior to truly prolonged and extensive experience without relevant reports in real-world clinical settings, given their overall high drug-attributable risk, the numerous and diverse potential pathomechanistic pathways and, consequently, the extremely wide range of implicated drugs. This is true even if the mechanism cannot be linked to the primary pharmacological effect of the drug. A classic example is hepatic injury. 'Idiosyncratic reactions' are also generally not considered unexpected *a priori* by pharmacovigilance experts, given their ill-defined pathogenesis and the diverse range of drugs that may be involved.

Designated medical events are generally not dismissed as possible ADRs *a priori* given their rarity, seriousness, high drug-attributable risk and the differential penalty associated with false-negative versus false-positive findings. On the less serious side of the spectrum, suspected ADRs involving subjective symptoms such as headache, nausea and abdominal pain are not generally considered unexpected *a priori* and may be possible with any drug. On the other hand, some drug reactions may be more likely to be discounted as a possibility. For example, drug reactions that may be considered 'paradoxical' in nature may be dismissed on manual review of adverse event lists because they are attributed to the natural history/complications of the treatment indication (e.g. a reported association of an anti-inflammatory drug with an inflammatory disorder).

In the course of evaluating DMAs, we have been impressed by some unexpected reactions and what we now call, for want of a better phrase, 'surprise reactions'. We began to use the descriptor 'surprise', not only to capture the quality of unexpectedness previously described in this section, but also our surprise at the ability of DMAs to highlight some of these associations on the basis of small numbers of reports well in advance of traditional surveillance data sources (e.g. the published literature) and/or regulatory actions. Recognising that it would represent a sample of convenience, we decided to maintain a 'registry' of such ADRs as we encountered them in the course of our data-mining activities because of their scientific interest and in hopes of usefully adding to the cumulative experience with these tools. These surprise reactions could be easily overlooked or discounted at the time of periodic reviews where line listings without clinical details are evaluated. Understanding this, we would suggest that if the drug safety expert were to incorporate the concept of surprise reactions into his/her pharmacovigilance logic, it might improve pharmacovigilance practice but it is difficult to make generalisations on preferred approaches.

We may characterise surprise reactions as reactions with a low drug-attributable risk that can be infrequently or rarely reported for the drug of interest or for drugs in general, are often of intermediate seriousness, do not have the classical hallmarks of 'idiosyncratic' ADRs and are not obviously explainable based on the primary pharmacological/therapeutic activity of the drug. They may only become recognised when *post hoc* explanations are sought based on more sophisticated pharmacological knowledge of the formulation, and at that point in time in hindsight they could qualify as a surprise reaction. Anecdotally, we have found that they can be associated with early statistical disproportionalities based on small numbers of reports and, although not qualifying as designated medical events (those reactions that are rare, serious and have a high drug-attributable risk), they are objective medical reactions that can have significant medical implications (e.g. cost of medical evaluation and sometimes medically serious also). The motivation for the descriptive typology is not to create a precise and all-inclusive definition that can prospectively identify all surprise reactions (e.g. as a standard medical query) because if there were such a set of clinical characteristics they would be readily identified in advance (e.g. like a designated medical event) and no surprise would be involved. Rather, the motivation is to increase understanding of the 'sample space' of ADRs in hopes of refining overall pharmacovigilance logic for optimum deployment of tools and strategies.

The objective of this paper is to explore the notion of 'unexpectedness' as it relates to signal detection and data mining in pharmacovigilance, by introducing the concept of surprise reactions and examining our anecdotal observations in a more methodical fashion. Formally, the question is whether these surprise reactions represent the type of ADR that could be amenable to detection with the assistance of adjunctive statistical calculations on SRS data. Our hypothesis is that DMAs could be of help in this circumstance.

We report the findings from our registry of well documented examples that we have encountered in the published literature or studied with DMAs to illustrate the potential usefulness of data mining in identifying these phenomena. For each of these examples, the drug-event combination might have escaped detection, either absolutely or relatively in terms of timing, in a real-life pharmacovigilance setting during periodic review of adverse event line listings for the reasons delineated above. Although not a systematic sample, the ADRs represent a diverse set of drugs and reactions and reflects a significant segment of real-world pharmacovigilance. It is important to remember that although these reactions are often quite well documented, causality is not established for most of the associations and that one could very rarely/never conclude that a causal relationship exists based on information contained in or derived from, SRS databases.

Methods

The US FDA Adverse Event Reporting System (AERS) database is a computerised-information database for post-approval safety surveillance. It functions as an early-warning system for ADRs not detected during pre-approval testing. It contains suspected ADR reports with approved drugs and therapeutic biological products submitted in accordance with mandatory reporting obligations by pharmaceutical companies and voluntarily by healthcare professionals and consumers. Suspected ADR reports are reviewed and coded for data entry in accordance with the standardised terminology of the Medical Dictionary for Regulatory Activities (MedDRA). Quarterly extracts are available through the National Technical Information Service. These quarterly updates are subjected to extensive cleaning (i.e. removal of redundant drug nomenclature and duplicate reports) by proprietary software vendors prior to data mining. The data extract used for the current analysis was from 1968 (year the safety database was created) through the second quarter of 2005.

The two DMAs chosen for this analysis were proportional reporting ratios (PRRs) and the empirical Bayesian algorithm multi-item gamma Poisson

shrinker (MGPS [Lincoln Technologies, Waltham, MA, USA]).[5]

The PRR is a simple metric relating the proportional representation of a reaction of interest with a drug of interest compared with the proportional representation of that reaction among all other drugs in the database. For this analysis, a PRR >2 with an associated χ^2 >4 and a case count threshold >2, which has been frequently cited in published studies of data mining, was considered a signal of disproportionate reporting (SDR).

The theoretical basis of MGPS has been described in detail elsewhere[3,6] but briefly is as follows. Expected counts for item sets (i.e. drug-event combinations) are based on the product of the marginal probabilities of each item (drug and reaction) in the database. The observed-to-expected (O/E) ratio is initially calculated as a crude disproportionality metric. Since the same ratio could be obtained from cell counts (frequencies) of markedly different sizes (O/E ratios based on smaller cell counts being considered more variable or imprecise) further modelling using maximum likelihood estimation and Bayesian inference are used to adjust the crude O/E ratios based on the respective drug-event combination counts. Each drug-event combination is modelled as a Poisson random variable. In turn, the Poisson parameter is itself related to a random variable that is distributed *a priori* according to a 2-gamma mixture. A Poisson model with a gamma distributed parameter gives a negative binomial distribution for the marginal prior distribution of O/Es. The gamma parameters of the prior probability distribution are obtained by applying an iterative maximum likelihood algorithm to a negative binomial mixture likelihood. Posterior estimates of the gamma parameters are obtained by updating the prior with the individual cell counts via Bayes theorem.

Using logarithmic transformations and taking the lower 5% cut-off of the posterior distribution of the observed/expected ratios (i.e. fifth percentile of the empirical Bayes gamma mixture) [EB_{05}], an expectation value that adjusts for the variability by down weighting or 'shrinking' the parameters associated with low observed and/or expected cell counts is obtained. These metrics are known as the empirical Bayes geometric mean (EBGM) and the EB_{05}. Therefore, an EB_{05} of eight may be interpreted to mean that reports of the particular drug-event combination occur in the database with a 95% likelihood at least eight times more frequently than would be expected if drug and reaction were independently distributed in the database. The signal metrics used for thresholds in the current analysis were the frequently cited $EB_{05} \geq 2$ as well as EBGM >2. FDA investigators have reported that in an internal validation procedure 'signals' obtained using a threshold of EB_{05} >2, based on FDA data "have high enough specificity to deserve further investigation".[6]

A variety of data-mining options and parameters exist including basic covariate adjustment (stratification by age, sex and year of report) and cumulative sub-setting. Stratification may reduce spurious associations due to confounding and markedly decreases the volume of disproportionalities.[7] However, since the precise prevalence and distribution of measured and unmeasured confounders and biological-effect modifiers within SRSs is not well understood, the overall clinical as opposed to statistical effect of stratification is unclear.

The aforementioned registry of well documented surprise reactions was compiled by the authors based on previous data-mining exercises or as they were encountered in the literature. Data mining was performed on suspect drug-ADR pairs using stratification by age, sex and FDA year of report with cumulative sub-setting by year to determine the first year in which each drug-event combination exceeded the commonly cited statistical thresholds used with the relevant DMAs and the corresponding number of reports.

The specific ADR used for data mining was either the verbatim MedDRA preferred term or the preferred term that was considered clinically most equivalent. Reactions for which no single preferred term could be found or for which there were no reports in the database were excluded from the analysis.

We manually reviewed published citations for these drug event combinations (restricted to

humans, English language, case reports supplemented by letters, 1950 through October 2006) generated through a search of MEDLINE (Dialogue DataStar, Cary, NC, USA) to identify relevant references that we were not already aware of. Articles where full text was available were reviewed in order to identify any earlier citations. A standard drug compendium Facts and Comparisons 4.0[8] was reviewed to determine whether ADRs were listed or unlisted.

Results

The authors identified a total of 40 surprise reactions. In 39 of the reactions, a suitable MedDRA term could be found. For one drug-event combination, clofazimine-enteropathy, no single preferred term could be found that would be suitable for data mining. For one drug-event combination, chlormethiazole-parotitis there were no reports in AERS. For four drug-event combinations, phenytoin-hypothermia; cotrimoxazole (trimethoprim/sulfamethoxazole)-aseptic meningitis; heparin-osteoporosis; prednisone-pancreatitis, the first literature report identified was published before 1968 and there were no cases in the start-up year of the FDA database (1968). These six reactions were excluded from further analysis. The 34 reactions involved 29 drugs in 17 different drug classes: anti-infectives (6), anticonvulsants (3), immunomodulators (3), corticosteroids (2), bisphosphonates (2), diuretics (2), anticoagulants (1), vasopressor (1), antidiabetic (1), antihypertensive (1), antipsychotic (1), analgesic (1), antiacne (1), anaesthetic (1), broncholytic (1), proton pump inhibitor (1) and contrast medium (1). The set of drugs were also temporally diverse, by which we mean their year of approval in the US and/or initial reporting activity generally occurred in all decades from the 1960s.

Published citations were found for the 34 reactions and the number of citations ranged from 1 to 66. The earliest publication for hydrochlorothiazide-pulmonary oedema was published in 1968 and the latest first publication was found in 2006 for propofol-priapism. Twenty-three of the reactions were listed in a standard drug compendium, Facts and Comparisons 4.0.[8] One drug, sodium iopodate, was not listed in the compendium. A summary of the findings is provided in table I.

There was no pattern in terms of nature of reactions, body systems in which they were contained, pharmacological/therapeutic classes of drugs or seriousness of reactions. Some reactions were inherently serious and the remainder could be considered serious under certain circumstances. Most drug-event combinations were listed in Facts and Comparisons 4.0.[8] Most SDRs appeared with two or all three metrics. For the seven drug-event combinations with reports but no SDRs with any of the metrics, the total number of reports as of the data-lock point ranged from 1 to 120 reports. The number of cases required to signal with PRR (N >2) and EBGM were in many cases the same or similar.

For those drug-event combinations associated with SDRs, performance similarities and differences were observed in the time to appearance of the SDR between the three metrics and their relationship to the publication date of the first MEDLINE citation. But overall, for all three metrics there was either no SDR or an SDR occurred concurrently or after the first literature report for a majority of the 34 drug-event combinations. A summary of these findings is provided in table II.

As displayed in table II, all three metrics provided an SDR for a particular drug-event combination in a majority of instances (≥65%). With the selected PRR and EBGM thresholds, the initial SDR for a particular drug-event combination usually preceded or coincided with the first literature citation, whereas with the selected EB_{05} threshold, the initial SDR usually occurred afterwards. Both PRRs and EBGM provided the same number of SDRs in advance of the first MEDLINE citation. From the perspective of sensitivity, both PRRs and EBGM seemed to outperform the EB_{05} metric in terms of the number of surprise reactions highlighted as well as the relative timing for those surprise reactions highlighted by more than one metric. This is not surprising since the EB_{05} metric includes both Bayesian shrinkage and an additional element of non-Bayesian shrinkage from the use of the lower 5% threshold.

Table I. 'Surprise reactions': first identified literature citation/drug compendium review/statistical disproportionalities (SDRs)

Drug-event combination		Year first literature	No. of citations[a]	Listed in Facts and	PRR >2, χ^2 >4, N >2		EBGM >2, N >0		EB_{05} ≥2, N >0	
drug	event (preferred term)	report identified		Comparisons 4.0[8]	year	no. of reports in year of first SDR	year	no. of reports in year of first SDR	year	no. of reports in year of first SDR
Acarbose	Lymphocytic colitis (colitis)	2000	1	No	1997	3	1998	5	2000	8
Bumetanide	Severe musculoskeletal signs and symptoms (myalgia)	1987	2	No	1987	6	1987	6	1988	12
Captopril	Cough (cough)	1985	26	Yes	1982	13	1982	13	1982	13
Carbamazepine	Hypertension (hypertension)	2002	1	Yes		NA		NA		NA
	Immunodeficiency (immunodeficiency common variable)	1983	5	No	1983	3	1981	2	1990	7
Ceftriaxone	Biliary pseudolithiasis (cholelithiasis)	1986	23	Yes	1987	4	1987	4	1988	7
Ciprofloxacin	Tendinopathy (tendon disorder)	1988	15	Yes	1993	3	1995	7	1996	31
Clozapine	Hypercholesterolemia (hypercholesterolaemia)	2005	1	Yes		NA		NA		NA
Cotrimoxazole (trimethoprim/ sulfamethoxazole)	Hyperkalemia (hyperkalaemia)	1993	17	Yes	1979	3	1979	3	1994	56
	Tremor (tremor)	1988	6	No		NA		NA		NA
Dexamethasone	Hiccups (hiccups)	1982	7	Yes	1980	3	1980	3	1980	3
Dobutamine	Eosinophilia (eosinophilia)	1995	4	Yes	1985	3	1985	3	1991	9
Dobutamine	Fever (pyrexia)	1992	2	Yes	1984	4	1982	2	1993	31
Erythromycin	Pyloric stenosis (pyloric stenosis)	1976	6	No	2004	3	2000	2		NA

Continued next page

Table I. Contd

Drug-event combination		Year first literature report identified	No. of citations[a]	Listed in Facts and Comparisons 4.0[8]	PRR >2, χ^2 >4, N >2		EBGM >2, N >0		EB_{05} ≥2, N >0	
drug	event (preferred term)				year	no. of reports in year of first SDR	year	no. of reports in year of first SDR	year	no. of reports in year of first SDR
Hydrochlorothiazide	Noncardiogenic pulmonary oedema (pulmonary oedema)	1968	26	Yes	1976	3	1974	2	1979	8
Ibuprofen	Aseptic meningitis (meningitis)	1978	30	Yes	1981	5	1981	5	1983	11
Interferon-α	Akathisia (akathisia)	1999	2	No		NA		NA		NA
Interferon-α	Hypertriglyceridemia (hypertriglyceridaemia)	1992	6	Yes	1998	11	1998	11	1998	11
Interferon α-2b	Parkinsonism (Parkinson's disease)	2002	1	No	2001	7	2001	7		NA
Iopodate sodium	Sialoadenitis (sialoadenitis)	1969	1	NA	1986	3	1971	1		NA
Isotretinoin	Hoarseness (dysphonia)	2005	1	Yes		NA		NA		NA
Itraconazole	Heart failure (cardiac failure)	2001	1	Yes	1995	9	2002	51		NA
Mefloquine	Pneumonitis (pneumonitis)	1998	2	No	1998	3		NA		NA
Methylprednisolone	Anaphylaxis (anaphylactoid reaction)	1974	22	Yes	1973	10	1969	2	1978	30
Montelukast	Churg Strauss syndrome (allergic granulomatous angiitis)	1999	21	Yes	1998	15	1998	15	1998	15
Omeprazole	Dry mouth (dry mouth)	1995	1	Yes	1990	10	1990	10	1991	18
Pamidronic acid	Ocular inflammation (uveitis)	1994	5	Yes	1995	3	1994	2	2000	5
Pegylated interferon-α	Dystonia (dystonia)	2004	1	No		NA		NA		NA
Propofol	Priapism (priapism)	2006	1	No		NA		NA		NA

Continued next page

Table I. Contd

Drug-event combination		Year first literature report identified	No. of citations[a]	Listed in Facts and Comparisons 4.0[8]	PRR >2, χ^2 >4, N >2		EBGM >2, N >0		EB_{05} ≥2, N >0	
drug	event (preferred term)				year	no. of reports in year of first SDR	year	no. of reports in year of first SDR	year	no. of reports in year of first SDR
Topiramate	Myopia (myopia)	1998	12	Yes	2001	17	1999	2	2001	17
	Oligohidrosis (anhidrosis)	2001	4	Yes	1997	4	1997	4	1997	4
Warfarin	Skin necrosis (skin necrosis)	1969	66	Yes	1973	3	1970	1	1976	7
Zoledronic acid	Osteonecrosis (osteonecrosis)	2003	27	Yes	2003	19	2003	19	2003	25
Zonisamide	Oligohidrosis (anhidrosis)	1997	4	Yes	2000	3	2000	3	2001	8

a 1968 through October 2006.

EB_{05}: 5th percentile of the posterior empirical Bayes gamma mixture; **EBGM** = empirical Bayes geometric mean; **NA** = not applicable; **PRR** = proportional reporting ratio.

An interesting observation was that the drug-event combinations that were highlighted earlier by EBGM relative to PRR were concentrated in old drugs reflected in data prior to 1985. This may reflect less intense reporting (or equivalently more under-reporting) prior to the change in requirements for reporting that were effected in the rewrite of section 21CFR 314.80 of the US Code of Federal Regulations in 1985. The additional case-count threshold requirement frequently imposed on the PRRs could exert a higher toll in periods of less frequent reporting, though we have observed a lack of consistency in the use/non-use of case-count thresholds with MGPS in various venues. The differential between PRRs and EBGM largely disappeared for drugs first marketed in the 1980s onward. However, given the limited number of observations and the sampling mechanism, our results should not be construed as a rigorous or complete comparison of the algorithms and associated metrics.

Discussion

Although this was not a systematic study and more along the lines of a concept paper, it appears that identification of surprise reactions may serve as an important niche for DMAs.

Our findings are broadly consistent with some more structured evaluations that we,[5] as well as other researchers,[9] have performed, in which conventional pharmacovigilance procedures identified a majority of suspected ADRs of interest prior to/concurrently with what would have been the first appearance of an SDR if one had indeed been identified (see table II). A question that may arise in the minds of some readers is whether these findings somehow refute the potential value of DMAs in pharmacovigilance since, in most instances, there was no SDR or an SDR was preceded by or occurred concurrently with a published literature report(s). However, we feel that the number of SDRs that preceded initial literature citation, although not a majority of the total, was not insignificant either, and suggests that DMAs may usefully supplement conventional procedures, such as literature surveillance. Furthermore, there are two potential uses of

DMAs in pharmacovigilance corresponding to the fact that identification of a credible signal is often the end result of a continuous process over a time interval, rather than a discrete event. Specifically, DMAs may not only help in the initial identification of a signal, but they can help refine an index of suspicion associated with signals initially detected by other means. Therefore, it would be a mistake to conclude that an SDR that appeared after initial publication of an association in the literature (or in SRS data) was useless. In real life, the safety reviewer might continuously weave together both clinical- and quantitative-data streams regardless of the order in which they appeared until a critical evidentiary mass is achieved that prompts definitive action. Nevertheless, some SDRs appeared quite early. In one example, an SDR for topiramate-oligohidrosis was identified with all three metrics/thresholds based on four cases, 4 years before first literature report and 6 years before dissemination of a safety alert in the US.[10]

Despite the fact that some of these surprise reactions may at first sight not seem to be of much regulatory or clinical importance, because the official criteria for a 'serious' reaction largely exclude reactions that are subjective or functional disorders and insufficiently take into account the pain experienced by or inconvenience to the patients (e.g. cough), many of these reactions are frequent causes of non-compliance or treatment withdrawal because they are important from the perspective of the patient's perception of effective treatment. Many of the adverse effects we studied are in fact inherently serious (e.g. noncardiogenic pulmonary oedema, pancreatitis, aseptic meningitis) and also might involve costly workups to determine the aetiology of these reactions. These ADRs are more likely to be overlooked by manual review of ADR lists because of the characteristics delineated previously (see background section).

An additional characteristic of surprise reactions can be the situation that there is more than one factor (than only the suspect drug) needed for their development, such as in ibuprofen-associated meningitis (which mainly occurs in patients with systemic lupus erythematosus), sialoadenitis with sodium iopodate in patients with renal failure, a drug interaction or a drug-reaction constituting a complex syndrome. In these circumstances, DMAs may identify what the human mind may be more likely to overlook. In recent publications, we retrospectively examined the potential utility of DMAs for early detection of potentially fatal/disabling ADRs.[5,11] The previous and the present publication when taken together may give a picture of the sensitivity of these algorithms across a range of serious and less serious ADRs.

Based on this non-systematic analysis, we would suggest that from the perspective of initial signal detection, DMAs may have value by focusing the attention of safety reviewers on reports of surprise reactions that might otherwise be overlooked or discounted based on manual review of ADR lists because of the characteristics delineated previously. They may also help refine an index of suspicion based on potential signals identified through traditional pharamcovigilance practices.

Our findings emphasise the importance of a holistic approach to signal detection using a comprehensive suite of signal detection strategies, including the importance of focused reviews of the published literature as a source of early 'signals' since DMAs did not always highlight these associations or

Table II. Timing of signals of disproportionate reporting (SDRs) in relation to first citation in MEDLINE for each metric studied

Relative timing of SDR vs first citation in MEDLINE (n = 34)	SDRs detected with PRR [n (%)]	SDRs detected with EBGM [n (%)]	SDRs detected with EB_{05} [n (%)]
SDR before	12 (35)	12 (35)	6 (18)
SDR concurrently	4 (12)	3 (9)	2 (6)
SDR after	11 (32)	11 (32)	14 (41)
Total	27 (79)	26 (76)	22 (65)

EB_{05}: 5th percentile of the posterior empirical Bayes gamma mixture; **EBGM** = empirical Bayes geometric mean; **PRR** = proportional reporting ratio.

highlighted them after initial publication. As we reviewed the literature reports, we found that convincing *post hoc* explanations could be found for some of these associations. For example, for cases of dobutamine-induced eosinophilia, the responsible agent might have been a preservative in the dobutamine solution, sodium bisulfite. Hypothermia with phenytoin overdose might be mediated by its effect on the hypothalamus, a well known effect of most of drugs with anaesthetic or sedative properties. Hypohidrosis with topiramate and zonisamide could be due to a carbonic anhydrase block at the level of sweat gland, thereby influencing pH dynamics by altering the hydrogen ion concentration. Sialoadenitis with sodium iopodate could be due a direct concentration-related toxic effect of iodine. For some drug-event combinations, no convincing pharmacological or other explanation could be identified either by authors in those references that we had access to through MEDLINE or on review.

Hand et al.[12] has stated that initially unsuspected patterns that can be credibly explained are more likely to be real. Therefore, we hope that future scientific research will not focus excessively on statistical algorithms, but will include attempts to harness technology that allows the safety reviewer to organise, access and integrate extra-statistical scientific knowledge relevant to the aetiology of ADRs with the statistical calculations, i.e. a marriage of data mining and knowledge mining. It is also important not to neglect optimising data quality and the application of clinical cognition to signal detection and evaluation.

Comparison of Data Mining Methods

Although we stress that our findings should not be construed as a systematic comparison between the DMAs/metrics studied and that our main message is the potential added value of all contemporary DMAs, we would be remiss if we totally avoided discussing the observed performance gradients.

First, we note that in both regulatory and industry settings some users like frequentist approaches (e.g. PRRs) because the method and output is clear, easy to understand and they feel more confident that they are not missing credible associations because of their higher 'sensitivity' when common implementations are used. Others in the same settings prefer the Bayesian approaches because they present the user with fewer spurious associations, which may be a critical factor in organisations with marked imbalance between data and resources. We feel both viewpoints merit respectful consideration and no criticism or endorsement of any one method or metric is intended or implied. In this regard, our general experience to date is consistent with one group of researchers from the FDA that concluded that frequentist and empirical Bayesian approaches perform competitively with each other with judicious threshold election, a given method may be superior in some situations and inferior in others, each method may have strengths and limitations and that statistical properties may not ensure superior performance.[13]

Of the two most commonly used implementations, we found the frequentist PRRs highlighted more surprise reactions than the most commonly used and recommended empirical Bayesian metric ($EB_{05} \geq 2$, $N > 0$). This is consistent with previous findings from regulatory, academic and industrial settings.[14] There may be a significant price to pay for this increased sensitivity in the form of an overabundance of highlighted associations that require additional triage criteria for practical implementation.[3,15-17] This price or opportunity cost will be dependent on multiple factors, including the number of reports involved and the ease with which spurious associations are identified as such (e.g. false alarms should be truly alarming) and other local organisational factors. The other empirical Bayesian metric (EGBM >2, N >0) is less commonly but still frequently used and in a sense more comparable to the PRR metric we studied (i.e. expected value versus expected value rather than a lower 5% cut-off versus an expected value).

The curiosity about the relative performance of common implementations of various DMAs and the research this curiosity inspired is both understandable and scientifically valuable; however, comparing DMAs using a common threshold for algorithms

that are designed to produce a different range of 'scores' requires judicious interpretation from a purely statistical perspective.[18] An alternative option is to compare the yield of credible associations of those metrics/thresholds that highlight the same overall number of associations. Under these circumstances, one may get a different perspective of which metric/threshold combination provides the 'biggest bang for the buck'. Ultimately, the significance of the comparative findings presented in this paper would depend critically on the opportunity costs associated with false-positive associations highlighted by each data-mining implementation. Although some authors have endorsed specific thresholds (e.g. $EB_{O5} \geq 2$),[5] systematic comparison of DMAs requires testing numerous thresholds with each DMA in diverse pharmacovigilance scenarios.[15-17]

In a recent state-of-the-art review of data mining using spontaneous reports, Bate and Edwards[19] note researchers persist in holding viewpoints that range from extremes of "unbridled optimism" to "considerable scepticism". Our cumulative experience to date, including our experience with surprise reactions, leads us to a more moderate view than those observed by Bate and Edwards[19] and acknowledge both the strengths and limitations of these tools. There are gold nuggets that can be effectively mined with emerging technologies but effective pharmacovigilance will be impaired by either overestimating or underestimating the value of any single approach.

Limitations

Our study had some limitations. First, because we know little about the properties of the publication process, relative to other processes, as a signal detection method, we recognise that it might not be a reliable reference standard. Pharmacovigilance organisations regularly survey the published literature for signals and, like every other pharmacovigilance tool, it remains unvalidated with specific performance characteristics undefined. It is known that the peer-review process does not provide immunity against spurious associations and indeed may be linked to SDRs with spurious associations.[20] Not withstanding the lack of validation because literature surveillance is routinely done in real-life pharmacovigilance settings, it is relevant if DMAs can usefully supplement, but not substitute for, this and other existing pharmacovigilance practices.

In addition, our non-systematic analysis does not fully replicate the manner in which DMAs would be deployed in naturalistic pharmacovigilance settings. For example, we did not exhaustively review all ADR terms associated with an SDR for clinical relevance with each drug, but merely chose a single preferred term that was deemed a direct representation of the medical concept. This is one particularly important limitation given the hypergranular nature of the MedDRA dictionary used to encode reactions in the AERS database because a given medical concept may be represented by numerous clinically overlapping preferred terms.[21] Also, results might have been different if we had used other restrictions for our MEDLINE search.

Conclusions

For many but not all organisations, the potential of DMAs to help detect truly novel events or detect events well in advance of conventional surveillance methods has so far not been fulfilled, with most statistical associations affirming what is already known, under review or judged to be spurious upon further scrutiny. This is not a negative reflection of the methods, but may reflect the fact that we already have very credible hypothesis-forming procedures in place. The ability of DMAs to detect the same events at the same time as conventional procedures but with enhanced resource efficiency may therefore receive greater emphasis. However, our anecdotal experience suggests that the 'sample space' of adverse events could contain associations that may be discounted and missed, absolutely or relatively in terms of timing, during conventional screening procedures yet are highlighted in a timely manner via data mining. The incremental utility in this regard may be modest, but still important if an appropriate balance is to be achieved between the benefits of additional true alarms and opportunity costs of addi-

tional false alarms generated as a result of data mining. Not withstanding the potential utility of DMAs in detecting surprise reactions, to us the most instructive lesson that we have learned from our compilation of these surprise reactions is that we must not be too quick to dismiss potential drug-event associations as being due to non-drug factors; in other words, SDRs may cause the reviewer to pause and think for a while and even possibly review and contemplate the pharmacology of the drug. Statistically, disproportionality should not be assessed in a biological vacuum but should trigger a thought process that incorporates scientific information about the drug, event and relevant patient population.

Acknowledgements

This manuscript was produced during work time and by using the software infrastructure of Pfizer Inc., New York, USA and of Boehringer Ingelheim GmbH, Ingelheim, Germany. Other than this, no funding, sponsorship and/or conflicts of interest exist. The authors also wish to thank Dr Muhammad Younis at the Department of Epidemiology, Michigan State University, East Lansing, Michigan, USA, for his help in the preparation of this manuscript.

References

1. Finney DJ. The detection of adverse reactions to therapeutic drugs. Stat Med 1982; 1 (2): 153-61
2. Trontell A. Expecting the unexpected: drug safety, pharmacovigilance and the prepared mind. N Engl J Med 2004; 351 (14): 1385-7
3. Hauben M, Madigan D, Gerrits CM, et al. The role of data mining in pharmacovigilance. Expert Opin Drug Saf 2005; 4 (5): 929-48
4. Aronson JK, Ferner RE. Clarification of terminology in drug safety. Drug Saf 2005; 28 (10): 851-70
5. Hauben M, Reich L. Potential utility of data-mining algorithms for early detection of potentially fatal / disabling adverse drug reactions: a retrospective evaluation. J Clin Pharmacol 2005; 45: 378-84
6. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and reactions in the US FDA's spontaneous reports database. Drug Saf 2002; 25 (6): 381-92
7. Hauben M, Reich L. Reports of acute angle closure glaucoma-related adverse events with SSRIs: results of a disproportionality analysis. CNS Drugs 2006; 20 (4): 327-9
8. Facts and comparisons 4.0 [online]. Available from URL: http://online.factsandcomparisons.com/ [Accessed 2006 Sep 25]
9. Lehman H, Chen J, Gould LA. An evaluation of computer-aided disproportionality analysis for post-marketing signal detection. Pharmacoepidemiol Drug Saf 2006 Aug; 15 Suppl. 1: S106-7
10. U.S. Food and Drug Administration. 2003 Safety Alert: topamax (topiramate) tablets/sprinkle capsules [online]. Available from URL: http://www.fda.gov/medwatch/SAFETY/2003/topamax.htm [Accessed 2006 Sep 25]
11. Hauben M, Reich L, Chung S. Postmarketing surveillance of potentially fatal reactions to oncology drugs: potential utility of two signal-detection algorithms. Eur J Clin Pharmacol 2004; 60 (10): 747-50
12. Hand DJ, Blunt G, Kelly MG, et al. Data mining for fun and profit. Stat Sci 2000; 15 (2): 111-31
13. Banks D, Woo EJ, Burwen DR, et al. Comparing data mining methods on the VAERS database. Pharmacoepidemiol Drug Saf 2005; 14 (9): 601-9
14. Hauben M, Reich L. Response to letter by Levine, et al. Br J Clin Pharmacol 2006; 61 (1): 115-7
15. Hauben M. Trimethoprim-induced hyperkalaemia: lessons in data mining. Br J Clin Pharmacol 2004; 58 (3): 338-9
16. Hauben M, Reich L. A case report of rhabdomyolysis with pentamidine that prompted a retrospective evaluation of a pharmacovigilance tool under investigation. Br J Clin Pharmacol 2004; 58 (6): 675-6
17. Hauben M, Reich L. Drug-induced pancreatitis: lessons in data mining. Br J Clin Pharmacol 2004; 58 (5): 560-2
18. Fawcett T. An introduction to ROC analysis. Pattern Recognition Lett 2006; 27 (8): 861-74
19. Bate A, Edwards IR. Data mining in spontaneous reports. Basic Clin Pharmacol Toxicol 2006; 98 (3): 324-30
20. Hauben M, Reich L, van Puijenbroek EP, et al. Data mining in pharmacovigilance: lessons from phantom ships. Eur J Clin Pharmacol 2006; 62 (11): 967-70
21. Hauben M, Patadia VK, Goldsmith D. What counts in data mining. Drug Saf 2006; 29 (10): 827-32

Correspondence and offprints: Dr *Sebastian Horn*, Boehringer Ingelheim GmbH, Corporate Drug Safety, Binger Straβe 173, 55216 Ingelheim, Germany.
E-mail: HORNS@ing.boehringer-ingelheim.com