Drug Safety 2009; 32 (6): 509-525
0114-5916/09/0006-0509/$49.95/0

An Evaluation of Three Signal-Detection Algorithms Using a Highly Inclusive Reference Event Database

Alan M. Hochberg,[1] *Manfred Hauben*,[2,3,4,5] *Ronald K. Pearson*,[1] *Donald J. O'Hara*,[1] *Stephanie J. Reisinger*,[1] *David I. Goldsmith*,[6] *A. Lawrence Gould*[7] and *David Madigan*[8]

1 ProSanos Corporation, Harrisburg, Pennsylvania, USA
2 Pfizer Inc., New York, New York, USA
3 New York University School of Medicine, New York, New York, USA
4 New York Medical College, Valhalla, New York, USA
5 Brunel University, West London, United Kingdom
6 Goldsmith Pharmacovigilance and Systems, New York, New York, USA
7 Merck Research Laboratories, West Point, Pennsylvania, USA
8 Columbia University, New York, New York, USA

Abstract

Background: Pharmacovigilance data-mining algorithms (DMAs) are known to generate significant numbers of false-positive signals of disproportionate reporting (SDRs), using various standards to define the terms 'true positive' and 'false positive'.

Objective: To construct a highly inclusive reference event database of reported adverse events for a limited set of drugs, and to utilize that database to evaluate three DMAs for their overall yield of scientifically supported adverse drug effects, with an emphasis on ascertaining false-positive rates as defined by matching to the database, and to assess the overlap among SDRs detected by various DMAs.

Methods: A sample of 35 drugs approved by the US FDA between 2000 and 2004 was selected, including three drugs added to cover therapeutic categories not included in the original sample. We compiled a reference event database of adverse event information for these drugs from historical and current US prescribing information, from peer-reviewed literature covering 1999 through March 2006, from regulatory actions announced by the FDA and from adverse event listings in the British National Formulary. Every adverse event mentioned in these sources was entered into the database, even those with minimal evidence for causality. To provide some selectivity regarding causality, each entry was assigned a level of evidence based on the source of the information, using rules developed by the authors. Using the FDA adverse event reporting system data for 2002 through 2005, SDRs were identified for each drug using three DMAs: an urn-model based algorithm, the Gamma

Poisson Shrinker (GPS) and proportional reporting ratio (PRR), using previously published signalling thresholds. The absolute number and fraction of SDRs matching the reference event database at each level of evidence was determined for each report source and the data-mining method. Overlap of the SDR lists among the various methods and report sources was tabulated as well.

Results: The GPS algorithm had the lowest overall yield of SDRs (763), with the highest fraction of events matching the reference event database (89 SDRs, 11.7%), excluding events described in the prescribing information at the time of drug approval. The urn model yielded more SDRs (1562), with a non-significantly lower fraction matching (175 SDRs, 11.2%). PRR detected still more SDRs (3616), but with a lower fraction matching (296 SDRs, 8.2%). In terms of overlap of SDRs among algorithms, PRR uniquely detected the highest number of SDRs (2231, with 144, or 6.5%, matching), followed by the urn model (212, with 26, or 12.3%, matching) and then GPS (0 SDRs uniquely detected).

Conclusions: The three DMAs studied offer significantly different tradeoffs between the number of SDRs detected and the degree to which those SDRs are supported by external evidence. Those differences may reflect choices of detection thresholds as well as features of the algorithms themselves. For all three algorithms, there is a substantial fraction of SDRs for which no external supporting evidence can be found, even when a highly inclusive search for such evidence is conducted.

Background

Since 1968, the US FDA has maintained a computerized database of adverse event reports.[1] The database collects spontaneous reports of adverse events made directly to the FDA or to drug manufacturers without regard to causality assessment. In recent years, computerized data-mining algorithms (DMAs) have been applied to look for patterns in the aggregate data that are not readily apparent from the scrutiny of individual case reports.[2,3] Since spontaneous reporting databases do not contain an explicit estimate of the exposed population, these DMAs cannot calculate incidence or prevalence rates for adverse events, and instead look for a disproportionate number of reports for a given adverse event in association with a particular drug, with respect to all other adverse events reported for that drug. A variety of statistical methods have been used for this, differing in the way that they determine statistical significance, and in the way that they apply corrections, either Bayesian or frequentist, to mitigate false signals when the numbers of reports are small and ratios therefore imprecise. Algorithms vary in the way that the input data are or are not stratified on various demographic variables, and in their use of multidimensional analysis (e.g. looking for drug interactions, or for syndromes of multiple co-occurring adverse events). Nevertheless, all the algorithms commonly in use today are variations on the theme of disproportionality analysis, which accept a contingency table of drug-event combinations as input, and produce a list of signals of disproportionate reporting (SDRs).[4]

Consensus has not yet emerged regarding the optimal role of data mining in pharmacovigilance.[5] One factor in understanding the capabilities and limitations of data mining is the development of standardized lists of adverse events, against which the lists of SDRs for a given algorithm can be compared. Several such standards have been developed, including those based on specific literature sources. A study by Lindquist et al.[6] utilized *Martindale's Extra Pharmacopoeia*,[7] and the *Physician's Desk Reference*,[8] while a study by Hauben and Reich[9] utilized a list of drug label changes. In general, efforts to create the standards have focused on establishing a high degree of certainty that the drug-event relationship is a causal one. This involves a great deal of effort from individuals with a high level of skill in interpreting pharmacovigilance data, and may be impractical in many cases, given the residual uncertainty typical of pharmacovigilance scenarios. Therefore, such databases are generally not very large. They are primarily useful for establishing the sensitivity of data mining, i.e. for determining whether a particular set of 'gold standard' drug-event combinations are detected by a given algorithm.

While sensitivity is an important aspect of signal detection, there is a great deal of interest in determining the specificity of DMAs and the positive predictive value of an SDR. DMAs appear to produce a significant number of false positives, that is, SDRs that do not have counterparts in a gold standard database. Is there external evidence to support these SDRs? To establish whether any external evidence supports these false positives, it is desirable to have a highly inclusive reference event database. This paper describes the construction of such a database, containing reported evidence for drug-event relationships from a variety of sources, including pharmaceutical product labels (prescribing information), formularies, FDA announcements and peer-reviewed publications. Thus, we attempted to broadly capture the published safety information for a set of drugs, rather than to create a more limited reference dataset only from those sources that can meet rigorous criteria for validation.

As a consequence, references in our database vary widely in the extent to which they establish, or attempt to establish, a causal relationship between the drug and the event that they report. Therefore, to be able to say something about sensitivity as well as specificity, we established and utilized rules for assigning each drug-event combination to an approximate 'level of evidence', to describe the strength of evidence for causality contained in the corresponding information source.

Since much of the interest in pharmacovigilance is in the surveillance of drugs during the first few years following their introduction, drugs for this study were selected from a list of new molecular entities (NMEs) initially approved by the FDA during 2000–2004. The intent of the study was to include typical drugs, rather than focusing on drugs that are known *a priori* to have serious safety issues. Thirty-two of the 35 drugs were chosen objectively, and the remaining three were chosen on the basis of their age in the marketplace and their therapeutic category, not on the basis of their safety profiles.

Methods

Data Sources and Selection of Drugs for Study

Input data for the study was taken from the public release known as the Freedom of Information Act version of the FDA adverse event reporting system (AERS) database, covering the period from the first quarter of 2001 through the end of 2005. In the AERS database, adverse events are described at the Medical Dictionary for Regulatory Activities (MedDRA®)[10] preferred term (PT) level, and terms at this level were used throughout the study to describe SDRs. Obsolete MedDRA PTs were updated to Version 10.1.

We obtained a list of 35 drugs in a pre-specified manner, by considering the NMEs approved in even years (2000, 2002 and 2004), and by ordering this list alphabetically and choosing every second drug, starting with the first drug. We then eliminated one insulin formulation, two multi-ingredient drugs and one pure-isomeric form of a previously marketed racemic compound because of the ambiguity in determining the representation of these drugs in the AERS database. In order to include some older drugs and to make up for the lack of certain key therapeutic categories in our drug set, we added on an *ad hoc*

Table I. List of the 35 drugs included in the study[1-3]

Adefovir pivoxil	Nitazoxanide
Amlodipine	Olmesartan medoxomil
Apomorphine HCl	Ovine hyaluronidase
Aripiprazole	Oxcarbazepine
Azacitidine	Pegaptanib sodium
Bivalrudin	Pregabalin
Cevimiline	Rivastigmine
Clofarabine	Rosiglitazone
Darifenacin HBr	Solifenacin
Docosanol	Telithromycin
Eletriptan HBr	Tinzaparin sodium
Erlotinib HCl	Treprostinil sodium
Ezetimibe	Trospium chloride
Gadobenate meglumine	Unoprostone isopropyl
Human secretin[a]	Venlafaxine
Icodextrin	Voriconazole
Lanthanum carbonate	Zonisamide
Mifepristone	

[a] Human secretin is used in diagnostic testing. It was one of the new molecular entities chosen for this study, but no records appeared for it in the adverse event reporting system database in the time period studied.

basis: amlodipine, a cardiovascular drug approved in 1996; rosiglitazone, an antidiabetic agent approved in 2000; and venlafaxine, an antidepressant approved in 1993. Table I shows the complete list of 35 drugs for the study. The following points describe further eliminations for unavoidable technical reasons, which brought the number of drugs actually analyzed to 27. These unavoidable eliminations took place after the data was generated, but before data analysis were conducted. Thus the choice to eliminate these drugs was not influenced by the results of the study.

- The study originally included Kaletra® (lopinavir/ritonavir) and Septocaine® (articaine/epinephrine). These drugs were eliminated after the study began, because validation revealed that the coding of the AERS data does not unambiguously distinguish between combination products and individual drugs used concomitantly. Data reported here do not include those drugs.
- Human secretin is used in diagnostic testing. It was one of the NMEs chosen for this study, but no records appeared for it in the AERS database in the time period studied.
- The Gamma Poisson Shrinker (GPS) algorithm requires a minimum of 100 adverse event reports in order to include a drug in its calculations. Rather than use different sets of drugs for the three algorithms, we eliminated drugs that failed to meet this criterion. The eliminated drugs were apomorphine, azacitidine, cevimeline, hyaluronidase, nitazoxanide, pegaptanib and unoprostone.

Construction of the Reference Event Database

For each of the drugs in the study, we derived reference event database entries from the following information sources: (i) the original package insert at the time of approval and subsequent revised packaged inserts, archived at the FDA web site; (ii) cautions and side effects information

Table II. Contents of the reference event database

Field	Contents
DrugName	Generic name of the drug for this entry
TradeName	Trade name of the drug
ApprovalDate	Date of first US FDA approval for the drug
SourceDate	Publication date of the source of information cited
SourceType	Type of publication (label, formulary, article, etc)
SourceTypeDetail	Type of publication detail (clinical trial report, case report, etc)
SourceReference	Bibliographical reference for the source of information
SourceTitle	Title of the source document
AdverseEventTerm	Verbatim term describing an adverse event
AdverseEventNotes	Verbatim text used to assign level of evidence, e.g. "Rate for drug was 2 × that of placebo"
EvidenceLevel	The assigned level of evidence
FromOrigLabel	(True/false) Is this entry taken from the prescribing information for the drug at the time of initial approval?
Reviewer	Name of individual assigning level of evidence
ReviewDate	Date of assignment of level of evidence

from the British National Formulary;[11] (iii) announcements of label changes and regulatory actions at the FDA website; and (iv) peer-reviewed publications reporting clinical trial reports, case series, pharmacoepidemiological studies and individual case reports, obtained through a MEDLINE search using the drug generic name. If >500 references were obtained by searching a drug name alone, then the search was narrowed by using the keywords ‘safety’, ‘adverse’ and ‘event’. We included only peer-reviewed MEDLINE publications and did not look for adverse event information on the Internet at large (e.g. by Google searches). To avoid ‘circular reasoning’, we eliminated reports if they utilized data mining of the FDA AERS database as the primary evidence of a drug-event association. This restriction affected only six information sources, while 378 information sources were used to create the reference event database. The contents of the reference event database are listed in table II.

We used the reference event database and additional information provided by adjudicators to assign a category to each SDR produced by a DMA. Table III lists the categories. Table IV provides the rules used for assignment of a level of evidence for each entry in the reference event database. Note that, while the terms ‘definite’, ‘probable’ and ‘possible’ were used, these levels of evidence do not correspond to the Naranjo criteria or other criteria that have appeared in the literature for assessing the causality of adverse events, which also use this terminology.[13-15]

Table V shows the types of information sources used to create the reference event database, and table VI shows the level of evidence associated with each of the 6207 reference event database entries that we created.

The blanket rules for assigning level of evidence caused the reference event database to skew heavily toward the ‘possible’ category. This reduced our statistical power to detect differences among the algorithms in how well they detected SDRs at the various levels (e.g. did one algorithm preferentially detect ‘definites’, etc). Therefore, we performed a second level of review to increase

Table III. Descriptions of the various categories to which a signal of disproportionate reporting (SDR) can be assigned. For SDR terms that match more than one entry in the reference event database, any entry in the I, G or D category takes priority, otherwise, the lowest numbered category is assigned

Category	Description
I-Indication	In the opinion of the adjudicator, the SDR is clearly related to the indication for the drug, rather than having a causal relationship to the drug (e.g. fungal infection for an antifungal agent) or to an indication-related confounder (e.g. fungal infection for an antiretroviral drug)
G-Generic	In the opinion of the adjudicator, the term in the SDR is generic or indecipherable in a way that makes it impossible to determine whether or not the term matches an item in the reference event database (e.g. unexpected therapeutic effect, blood test abnormal)
D-Duplicates	The SDR is due to the presence of duplicated reports in the adverse event reporting system database. Duplicate records were detected by automated matching on three of four of the following fields: sex, age, weight and event date
0-Regulatory action	The SDR matches a reference event database entry for a US FDA or manufacturer post-approval regulatory actions reflected at the FDA MedWatch website, including drug recalls, Dear Healthcare Provider letters, black-box warnings and enhancements to the WARNINGS, PRECAUTIONS or similar sections of drug labelling. This does not include simply the listing of an adverse event term in a list or table in the adverse events section of a label
1-Original label	The term for the SDR matches a reference event database entry that corresponds to a precaution, warning or adverse event already listed in the prescribing information for the drug at the time of its initial FDA approval
2-Definite	The term for the SDR matches a reference event database entry that has a level of evidence of ‘definite’, according to the criteria of table IV
3-Probable	The term for the SDR matches a reference event database entry that has a level of evidence of ‘probable’, according to the criteria of table IV
4-Possible	The term for the SDR matches a reference event database entry that has a level of evidence of ‘possible’, according to the criteria of table IV
5-Minimal	The term for the SDR matches a reference event database entry that has a level of evidence of ‘minimal’ according to the criteria of table IV, meaning that no causality-related information is provided
6-Lacking	The term for the SDR does not match any entry in the reference event database

Table IV. Rules for the assignment of level of evidence for adverse event terms included in the reference event database

Criterion	Level of evidence assigned
Added to the US label after initial drug approval	Possible
Was the subject of manufacturer or regulatory action, beyond simple listing in an adverse event table (Dear Doctor letter, black-box warning, withdrawal from market or restrictions on prescription)?	Regulatory
Listed in a non-US National Formulary	Possible
Reported as "Incidence greater than placebo (or comparator drug)" in a trial[a]	Definite
Adverse events whose characteristics are sufficiently objective and specific to serve as proof positive gold-standard reference events on the basis of individual case reports[12]	Definite
Mentioned in the adverse event table of a trial, but not noted as having higher incidence for study drug than for placebo or comparator	Minimal
Stated as 'definitely', 'probably' or 'possibly' drug-related in a clinical trial report or case report that utilized the Naranjo criteria or similar criteria to establish causality	Corresponding designation applied in the reference event database
Stated as 'probably' or 'likely' drug-related in a trial article or case report without specific reference to the Naranjo criteria or similar criteria	Probable
Stated as 'possibly' drug related in a trial article or case report	Possible
Stated by the author as 'treatment emergent' in a case report or an article describing a trial	Possible
Subject of a peer-reviewed case report	Naranjo criteria applied[13]
Subject of a peer-reviewed case series	Naranjo score for the highest-scoring individual case, when the Naranjo criteria are applied to each case separately
Reported as a statistically significant association in a peer-reviewed article describing a pharmacoepidemiological study of spontaneous reports, not relying on the AERS database (AERS-based studies are excluded, since an AERS-derived SDR should not be considered as supporting evidence for the same AERS-derived SDR)	Possible
Is reported as a statistically significant association in a peer-reviewed article describing a pharmacoepidemiological study that used chart review	Possible

[a] We considered a requirement that these comparisons be corrected for multiplicity. However, we found that such corrections are rarely used in safety analysis because their use in that context is not conservative. Only approximately 10% of the reference sources in this category utilized multiplicity correction. We also considered whether these uncorrected findings should be downgraded from 'definite' to 'probable'. We felt that the availability of a clean, controlled experimental comparison (target drug vs control) compared with more typical observational origins of signals justified leaving them in the 'definite' category.

AERS = adverse event reporting system; **SDR** = signal of disproportionate reporting.

the number of 'definite and 'probable' drug-event combinations. Specifically for the 268 case reports and case series contained in the database, the level of evidence for case reports and case series was assigned individually using the Naranjo criteria.[13] The specifics of this assignment were as follows. For the question "Are there alternative causes other than the incriminated drug?", this was scored as 'No' if the article rules out alternative causes based on laboratory results or discussion; 'Yes' if the article specifically discusses alternate causes; and 'Blank' otherwise. This was over-ridden to 'Yes' on rare occasions when the adjudicator identified an obvious alternative cause not discussed in the article, or conversely to 'No' if the drug-event relationship was so well established (based on the remaining criteria) that a discussion of alternative causes would have been superfluous. The question, "Are there previous conclusive reports of this reaction?" was scored based on whether or not the article under consideration mentioned previous reports. The adjudicator did not conduct an independent literature search to proactively identify previous reports for each article scored. For these peer-reviewed articles, nearly all articles

Table V. Number of reference event database entries by level of evidence

Level of evidence/status	Number of reference event database entries
0-Regulatory action	66
1-Original label	3776
2-Definite	79
3-Probable	152
4-Possible	1744
5-Minimal	390
Total	6207

either cited previous reports or stated that a search had been done. For articles that described case series, the Naranjo score was obtained by scoring each case individually and taking the highest one. The adjudicator was not permitted to combine features of several cases into a single score. So if one case described re-challenge and another described a dose-response relationship, these points were not added together to raise the score as if both features were described in a single case.

The reference event database is available through the Pharmaceutical Research and Manufacturers of America by contacting the corresponding author, subject to a simple (one page) data use agreement.

Adjudication of Data-Mining Signals

The process of matching SDRs from data mining against items from the reference event database is not straightforward since the reference event database contains verbatim terminology from various sources, while the SDRs employ coded MedDRA terminology. For example, a reference event database entry might contain the reported verbatim term 'anemia', while a corresponding SDR might involve the MedDRA PT 'haematocrit decreased'. To carry out the matching, we developed clinical criteria for a match, and created adjudication software, which allows a user to match SDRs to reference events in a blinded fashion, so as not to favour one DMA over another. Finally, we performed an inter-rater study to determine the consistency of results among adjudicators.

The adjudication software presents the user with a list of MedDRA PTs representing SDRs for a given drug, alongside a list of reference event database entries for the same drug. The adjudication software is capable of aggregating lists of PTs for a drug from several DMAs, and presenting them to the user in a manner that blinds the user as to which algorithm(s) detected the PT.

The adjudicator in this study was a non-clinician researcher. In the experiment described in Appendix A, this individual's results were compared with those from two drug safety physicians with a minimum of 18 years' experience who participated as adjudicators only in the inter-rater experiment and not in the main body of this study.

The first step of adjudication was to identify those SDRs for which comparison to the reference event database would be meaningless or inappropriate in the context of measuring the performance of various methodologies. These 'flagged' SDRs included those that relate to the indication for the drug rather than to an adverse event (e.g. *fungal infection* for an antifungal agent), indecipherable PTs such as *unexpected therapeutic effect* and those that were detected by an automated algorithm as a result of duplicate reporting in AERS. (De-duplication was carried

Table VI. Contents of the reference event database by information source type

Source type	Count of reference event database entries
Original prescribing information at time of drug approval	3776
British National Formulary	899
Clinical trial report	641
Case report	231
Label changes	198
Review article	68
Case series	37
Observational cohort study	31
Dear Healthcare Provider letter	21
Other (cohort study, meta-analysis, chart review, etc.)	305
Total	6207

out through an algorithm applied to the list of case reports for each SDR individually. This process is both more specific and more computationally efficient than attempting to compare each of the approximately 1 million records of the AERS database with every other record in the database to ascertain duplicates).

In the second step of adjudication, the adjudicator was instructed to match SDRs (MedDRA PTs) to reference event terms when the "nature, severity, specificity, and outcome" of the SDR were, in their judgement, "consistent with the information" embodied in the reference event database term. This excerpt is taken from the FDA publication of the International Conference on Harmonisation E2C guidance, regarding identification of safety signals of interest.[16] In many cases, the matching was trivial. For example, the PT 'thrombocytopenia' trivially matched the reference event database term 'thrombocytopenia'. In other cases, the matching involved simple interpretation, for example, 'thrombocytopenia' versus 'low platelet count'. In other cases, more complex adjudicator judgement was required. For example, an adjudicator might choose to match the PT 'platelet count decreased' to the reference event database term 'thrombocytopenia', on the basis that "the nature, severity, specificity and outcome of '*platelet count decreased*' is consistent with the information embodied in the term '*thrombocytopenia'*." Note that these instructions are not symmetric: an adjudicator might match PT 'platelet count decreased' to reference event 'thrombocytopenia', saying that a decrease in platelet count is consistent with thrombocytopenia, but might not match PT 'thrombocytopenia' to reference event 'decreased platelet count', since thrombocytopenia (i.e. an absolute drop in platelet count below a particular threshold) cannot be inferred from decreased platelet count. This approach was designed to make the matching process as consistent and objective as possible, given the inherent variability in signal evaluation practices. An additional rationale for this approach is that it is liberal in matching SDRs to terms that are already on the label for the product so that they do not count incorrectly as having contributed new safety information. This results in a conservative estimate of the informational contribution of data mining.

Adjudicators were supposed to find all matches for a given SDR term, so if 'thrombocytopenia' was matched to 'thrombocytopenia', they were instructed to continue looking for additional terms, e.g. 'low platelet count'. However, "satisfaction of search" is a well recognized phenomenon,[17] and we did not study, for instance, whether the same matches would have been made if the terms had been presented in a different order than the alphabetical order used by the software. In any case, the blinding ensures that omissions as a result of satisfaction of search would not bias the results toward any particular algorithm.

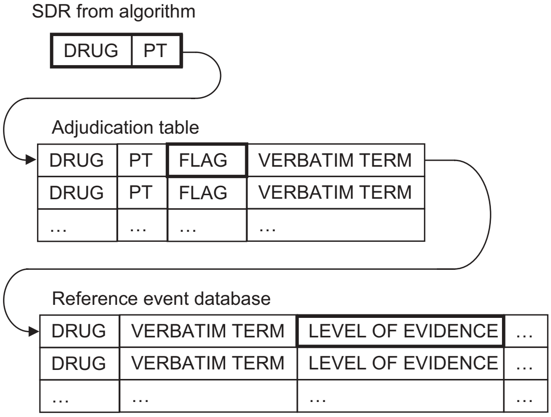
To determine inter-adjudicator reproducibility, three adjudicators evaluated all SDRs for a pilot data-mining exercise employing three methods and five drugs. Appendix A provides a description of the inter-rater experiment.

The database layout and methods for assigning a category to an SDR based on the adjudication results and the reference event database are shown schematically in figure 1.

Data Mining

The proportional reporting ratio (PRR) algorithm as described by Evans et al.[18] and the urn-model algorithm as described by Hochberg et al.[19] were computed in S-Plus (Insightful Corporation, Seattle, WA, USA). The GPS algorithm, as described by DuMouchel,[20] was implemented in SAS/IML (SAS version 9.1; SAS Institute, Cary, NC, USA), using a code provided by Dr Ivan Zorych of Rutgers University. The code was modified to calculate the score designated by DuMouchel as EB05 (the lower 95% confidence interval limit of the Empirical Bayes Geometric Mean [EBGM]), in addition to the EBGM score originally calculated in the Rutgers implementation.

The GPS algorithm requires an input matrix of drug and event counts for all drugs and all event terms that are involved in at least 100 cases ($N_{j.}$ ≥100 or $N_{i.}$ ≥100 in the terminology of DuMouchel[20]) to calculate hyperparameters.



SDR from algorithm

DRUG	PT

Adjudication table

DRUG	PT	FLAG	VERBATIM TERM
DRUG	PT	FLAG	VERBATIM TERM
…	…	…	…

Reference event database

DRUG	VERBATIM TERM	LEVEL OF EVIDENCE	…
DRUG	VERBATIM TERM	LEVEL OF EVIDENCE	…
…	…	…	…

Fig. 1. Illustration of how the category is assigned for a signal of disproportionate reporting (SDR). An SDR may have a flag value of 'indication', 'generic', 'duplicate' or blank. If the flag is not blank, the flag is used for the categorization. If the flag is blank, the level of evidence is used for the categorization. If there is no entry in the adjudication table for the SDR, then the categorization of '6-lacking' is assigned. When there is more than one reference event database entry for a given drug-event combination, then a non-blank flag value takes priority over level of evidence. If all flags are blank, then the lowest-numbered level of evidence is assigned as the category for the SDR. **PT** = preferred term.

It cannot be run using input for only a small set of drugs, such as the 35 drugs in this study. Therefore, we ran the GPS algorithm on this full matrix of drugs and events, as is typically done, and then extracted the information for our 35 drugs from the resulting output dataset. We considered removing the $N \geq 100$ constraints to better match the input to the urn model and PRR algorithms, but found that it is necessary for the convergence of the iterative calculation, so we enforced it, and wish to acknowledge it. In particular, we do not know if unpublished improvements to the algorithm in commercial software have overcome this limitation.

GPS analyses were stratified on age, sex and year of report source as described by DuMouchel.[20] Categories for age were 0–17 years, 18–39 years, 40–64 years, 65–130 years and 'missing'. Categories for sex were 'male', 'female' and 'missing'. About 25% of ages and 7% of sexes are missing. Analyses for the urn model and for PRR were not stratified. In a separate study, we found that the choice to stratify GPS and not PRR or the urn model gave the best sensitivity and specificity for each algorithm (this study is being prepared as a manuscript for publication). While stratification of PRR has been used by some, and is recommended by Woo et al.[21] in a paper that appeared after our study was conducted, the effect of stratification on PRR is modest in terms of the overall number of SDRs produced, and even smaller in terms of sensitivity and specificity.

Scoring thresholds for the reporting of a significant drug-event combination were taken directly from the references that describe each of the DMAs, and were as follows: for GPS, a score of EB05 >2.0, where EB05 is the lower 95% confidence interval limit of the EBGM computed by the GPS algorithm; for the PRR method, a PRR >2.0, an event count >2 and an associated chi-squared value >4.0; for the urn model, Reporting Ratio (RR) >1.0, an event count >2 and a statistical unexpectedness >N/0.05, where N is the total number of MedDRA PTs present in the AERS database for the drug in question. The decision to limit the urn model and PRR to an event count greater than two was made to match the conditions used in previous studies.

We chose thresholds for this study that are in widespread use, and have appeared in previous publications, but these are not 'official' thresholds sanctioned by any expert group, and we are not suggesting that drug safety professionals should necessarily use these same thresholds in practice for all purposes.

Analysis of Data-Mining Results

When an SDR matched a reference event term, it was assigned a 'regulatory action', 'definite', 'probable', 'possible' or 'minimal' level of evidence, according to the information in the reference event database. SDRs that did not match any reference event term for their drug were designated as 'lacking' evidence. For each report source dataset, the number of SDRs in each of these categories was tallied for each data-mining method and for each drug, using software written in S-Plus (S-Plus Version 7.0; Insightful Corp., Seattle, WA, USA).

In addition to examining the numbers of SDRs in various categories from the various algorithms, we wanted to study the degree to which the different algorithms yielded different sets of SDRs. To approach this question, we calculated the overlap among various algorithms and tallied the results.

Scoring was performed by counting matches to individual terms, not to more broadly defined medical concepts, nor to individual reference sources in the reference event database. There is no standard for how finely to draw distinctions among related terms when scoring studies such as this, and the effects of this issue on data-mining evaluations have been discussed in detail by Hauben et al.[22]

As an adjunct to the analyses presented here, SDR counts in the contingency tables were fitted to a generalized linear model (GLM) of the Poisson family (log-linear model), using the 'glm()' and associated functions in S-Plus Version 7.0. This model quantified the influence of DMA and level of evidence categories on the proportion of SDRs detected.

Results

Signal of Disproportionate Reporting Counts and Categories for the Three Algorithms

Table VII shows numbers of SDRs in the various reference level categories for the various report sources and DMAs. In total, the urn model yielded 1562 SDRs, GPS yielded 763 SDRs and PRR yielded 3616 SDRs. Table VIII shows the percentage of SDRs in each category. We have also provided totals for the numbers of 'unlabelled supported' SDRs. These are SDRs that correspond to an entry in the reference event database at the 'minimal' level or above, but did not appear in the original prescribing information for the drug product at the time of its approval. In other words, these are findings from data mining that are 'novel' (with respect to the original label) and are supported by at least minimal external evidence. PRR produced the highest number of unlabelled supported SDRs (296), followed by the urn model (175), with GPS highlighting the fewest unlabelled supported SDRs (89). Considering the fraction of total SDRs that fell in the unlabelled supported category for each algorithm, the highest fraction was obtained for GPS (11.7%) followed by the urn model (11.2%) and then by PRR (8.2%).

Table VII. Classification of signals of disproportionate reporting detected by the three algorithms

Category	Urn model[19]	GPS[20]	PRR[18]
I-Indication	80	52	108
G-Generic	68	36	176
D-Duplicates	3	3	5
0-Regulatory action	40	24	49
1-Original label	486	265	790
2-Definite	4	3	10
3-Probable	19	3	33
4-Possible	97	56	178
5-Minimal	15	3	26
6-Lacking	750	318	2241
Total	1562	763	3616
Unlabelled supported (0+2 through 5)	175	89	296

GPS = Gamma Poisson Shrinker; **PRR** = proportional reporting ratio.

When overlap among the various algorithms is accounted for, a total of 322 unlabelled supported SDRs were detected using all three DMAs combined. Table IX shows the overlap among SDRs detected by various combinations of methods and shows the number needed to detect (NND) an unlabelled supported SDR for each combination of algorithms. The NND is the reciprocal of the fraction of SDRs that are unlabelled and supported.[23] It reflects in some sense the analyst workload required to detect an SDR with at least 'minimal' support in the literature. The NND represents the total number of SDRs that an analyst would need to review per unlabelled supported SDR detected. SDRs that were flagged as 'indication-related', 'generic' or 'duplicate', etc., were not counted in the NND calculation, since these determinations are made by inspection and do not require case review, which accounts for the largest share of the workload in reviewing an SDR.

Generalized Linear Model Results

The data in table VII were fitted to a Poisson-type GLM with two terms, one for the category

Table VIII. Classification of signals of disproportionate reporting (SDRs) detected by the three algorithms, as a percentage of total SDRs detected. For the percentage of unlabelled supported, 95% CIs are shown

Classification	Urn model[19] (%)	GPS[20] (%)	PRR[18] (%)
I-Indication	5.1	6.8	3.0
G-Generic	4.4	4.7	4.9
D-Duplicates	0.2	0.4	0.1
0-Regulatory action	2.6	3.1	1.4
1-Original label	31.1	34.7	21.8
2-Definite	0.3	0.4	0.3
3-Probable	1.2	0.4	0.9
4-Possible	6.2	7.3	4.9
5-Minimal	1.0	0.4	0.7
6-Lacking	48.0	41.7	62.0
Total	100	100	100
Unlabelled supported (classification 0+2 through 5)	11.2 (9.6, 12.8)	11.7 (9.4, 13.9)	8.2 (7.3, 9.1)
Number needed to detect	8.9	8.5	12.2

GPS = Gamma Poisson Shrinker; **PRR** = proportional reporting ratio.

of adverse event ('category'), i.e. the level of evidence, and one for the algorithm employed ('algorithm'). SDRs in the 'indication', 'generic' and 'duplicates' categories were excluded from this analysis. The reason for testing this model was to see if the algorithms differed strongly in their propensity to find higher versus lower levels of evidence. Such a tendency would have resulted in a large residual deviance for this two-term model. We found that the residual deviance was only 1.3% of the total deviance, indicating that the three algorithms produce similar mixtures of levels of evidence.

Inter-Rater Adjudication Study

Results of the inter-rater adjudication study are described in Appendix A.

Discussion

Data-Mining Algorithms

In this study, SDRs detected by data mining were compared with an extensive list of drug-associated events in a reference event database. This list was extremely inclusive of drug safety information available at the time of its completion (March 2006). The mere fact of the occurrence of an adverse event during a trial was enough for inclusion at the 'minimal' level. Nevertheless, a large fraction of the SDRs did not match any reference event database entry. The fractions in the 'lacking' category were 48.0% for the urn model, 41.7% for GPS and 62.0% for PRR. If we remove SDRs that appear on the original label from consideration, these percentages are considerably higher. These non-matching SDRs may include spurious false-positive associations; however, they may also include legitimate drug safety signals for rare events, which have not been recognized by other means. We note that this is a 'young' cohort of drugs, that is, they have not been on the market for a prolonged time. Some of the reference-'lacking' SDRs may be recognized in the literature as time goes on.

Table IX. Overlap of signals of disproportionate reporting (SDRs) detected by the Urn model[19], Gamma Poisson Shrinker (GPS)[20] and proportional reporting ratio (PRR)[18], and number needed to detect an SDR matching the reference event database for each combination

Detected by	Unlabelled supported SDRs	Total SDRs	Percentage unlabelled supported ± 95% CI width	Number needed to detect an unlabelled/supported SDR
Urn model only	26	212	12.3 ± 4.4	8.2
GPS only	0	0		
PRR only	144	2231	6.5 ± 1.0	15.5
Urn model and GPS	0	0		
Urn model and PRR	63	622	10.1 ± 2.4	9.9
GPS and PRR	3	35	8.6 ± 9.3	11.7
All three algorithms	86	728	11.8 ± 2.3	8.5

Still, these numbers provide a useful perspective on the utility of data mining.

These percentages should be placed in the context that data mining is essentially a screening application. This figure means that the NND, the number of SDRs that must be examined to find one SDR that is unlabelled and supported by external evidence, is a relevant figure. For the algorithms studied, this number ranges from 8 to 15.5. Of course, the actual workload would depend on additional factors. Adjudicating SDRs that are described on the original label can be carried out by simply noting the SDR and noting the label terms, and does not involve a detailed review of cases. Assigning an SDR to the category of 'confounding by indication' is often fairly straightforward (e.g. voriconazole and fungal infection), although doing so sometimes presents a difficult and complex decision; for example, in the case of suicidal ideation and psychiatric medication. Also, when it comes to case review, it should be kept in mind that reviewer workload may be more closely related to the number of unique reports that would have to be reviewed, rather than to the number of SDRs considered in the study.[23]

Therefore, our results do not provide rules for determining whether a method that provides more unlabelled supported SDRs but a lower positive predictive value (e.g. the PRR) is universally better or worse than a method that has a higher predictive value but generates fewer unlabelled supported SDRs. Using previously published thresholds, the three algorithms represent three distinct points along a spectrum of sensitivity versus specificity tradeoffs, and it is neither the goal of this study, nor do the authors necessarily believe it is possible, to declare a single 'winner'. The GPS algorithm gives the highest rate of matching against the reference event database for drug-event associations with at least minimal external supporting evidence. However, this algorithm also detects the smallest number of SDRs, and did not uniquely highlight any unlabelled supported SDRs. PRRs identified the highest number of unlabelled supported and unique unlabelled supported SDRs followed by the urn model. The relative premium on specificity over sensitivity of the related multi-item GPS algorithm, compared with PRR, has been previously noted.[24] It should also be noted that the ability of an algorithm to detect unique supported SDRs does not imply that it is uniquely detecting causal drug-event relationships.

It would be interesting to ask, "What fraction of the reference event database was detected by data mining?" We avoided this because the terminology in this database is not normalized in the way that MedDRA PTs are. 'Fever' and 'pyrexia' can both appear, and in fact can appear multiple times for a particular drug, from different sources. Thus, the idea of counting reference event database entries the way we count MedDRA PTs appears not to be meaningful.

Further Limitations of the Study

We acknowledge a number of significant limitations to this study.

Many data limitations and deficiencies bedevil analyses of AERS; the literature has discussed these limitations at length and we will not elaborate here. In our view, DMAs can, at best, suggest drug-event associations that can potentially be confirmed on the basis of controlled studies, medical evidence and fundamental pharmacological considerations.[25]

We studied detection of drug-event combinations and did not consider higher-order associations, e.g. drug interactions.

This study did not consider the time lag from market introduction detection of an SDR for a safety signal that subsequently became well established. This requires different methodology and was the subject of a subsequent experimental study, the results of which are described in a separate publication.[26] We performed all analyses on aggregate data from 2001 to 2005, and did not consider the appearance or disappearance of transient SDRs that might appear above threshold for some period of time and then fall below threshold. We note also that this study was done with peri-approval data, which spanned approximately the first 5 years of a drug's life on the market. The conclusions from this study do not necessarily apply to drugs that have been on the market for many years or decades.

Although many of the reference event database entries were the result of traditional pharmacovigilance investigations, we did not explicitly compare data mining with traditional, qualitative methods of pharmacovigilance, and therefore cannot make direct statements regarding the incremental utility of data mining in general over such methods.

There were additional limitations to the study introduced by the implementation of the reference event database. It is possible that using such a highly inclusive reference set may have biased the results in favour of more sensitive and less specific signal-detection methods, compared with studies where more selective standards were used. Because of the volume of data to be included, practical considerations limited us to blanket rules for assigning levels of evidence, based on broad features of the articles and other information we reviewed. Criteria involving a more complete review of each individual study for factors such as de-challenge/re-challenge information, standards for case definition and other criteria might have resulted in a higher number of reference events classified as ‘definite’ or ‘probable’, and might have increased our power to detect subtle differences among algorithms at those levels. The blanket rules for ‘definite’, ‘probable’ and ‘possible’ provoked considerable debate among the authors during the consensus-building process, and it is clear that further extensive debate among reasonable investigators would be possible. Even though the Naranjo criteria were used in some cases, it should be noted that none of the published causality systems have been thoroughly validated for routine causality assessment of adverse event reports. However, the GLM shows that differences among algorithms in the way they treat the various levels do not affect the overall detection results; 98.7% of the variability in the data is explained without including such terms in the model. Furthermore, our finding of a large proportion of SDRs that lacked a match in the reference event database would not be altered by changing the criteria used to assign drug-event combinations to the ‘definite’, ‘probable’ and ‘possible’ categories.

Subjective judgement was required to identify indication-related SDRs and SDRs due to confounders. Indication-related SDRs and confounders are arguably two distinct phenomena. They were grouped together in this analysis to place them in a category of SDRs that are statistically significant but of no practical utility in identifying candidate drug-event combinations for further safety investigation. We note that the ‘indication’ category of indication-related and confounded SDRs made up only 3–7% of total SDRs; therefore, it is not likely that even a significant misclassification rate with respect to this category would have affected the overall results in a way that favoured one algorithm over another.

The adjudicator who made decisions regarding the flagging of SDRs and the matching of MedDRA terms against the reference event database was a drug-safety researcher with 31 years’ experience in medical informatics, although was not a physician, and would therefore not necessarily exercise judgement in the same way as a clinician. With regard to this concern, we note that (i) the results of the inter-rater experiment showed that the differences among raters, while statistically significant, were small compared with the differences among algorithms or levels of evidence; (ii) the differences between the non-clinician rater 1 and the two physicians, raters 2 and 3, were similar in magnitude to those between raters 2 and 3; and (iii) the blinded aspect of the adjudication make it unlikely that errors in judgement on the part of the adjudicator would serve to favour one algorithm over another in these results.

We relied on detection thresholds that have been used in several published studies, but did not conduct a sensitivity analysis to examine the effect of varying thresholds for any of the methods. The importance of separating intrinsic methodology differences from threshold implementation differences and of testing multiple metrics and thresholds in diverse pharmacovigilance scenarios, was emphasized by Chan and Hauben[27] and Hauben.[28]

We did not consider the degree of overlap or relatedness among various PTs; therefore, we cannot make statements about the numbers of distinct medical concepts detected by the various algorithms under the various conditions studied.

It was pointed out during review of this manuscript that we have no proof that matching of an SDR to a reference event database is in and of itself a good thing, or that fraction matching should necessarily be considered a figure of merit for a DMA. Cases in the literature may have been entered into the AERS database, producing an SDR that will automatically be supported in the reference event database. Stimulated reporting to the AERS database following the publication of a case report could produce a similar phenomenon. A small number of cases with weak evidence for causality could be amplified into an SDR by these mechanisms. Support in the literature does not necessarily imply causality. Perhaps it is the non-matching SDRs that provide causal, novel and important findings.

Finally, it is difficult to make prescriptive statements about the use of data mining and comparative performance of DMAs given the absence of a clear decision theoretic calculus of utilities and consequences of reducing false-positive versus false-negative findings. In general, the greater the imbalance between too much data and constrained resources, the more important is the task of reducing false positives, all other factors being equal. However, an exclusive focus on reducing false positives may prevent useful knowledge discovery, just as an exclusive focus on reducing false negatives may be self-defeating by flooding the user with large amounts of information of limited or no practical use from a public health perspective. The question that remains is how to determine the optimum range of sensitivity and specificity for a given pharmacovigilance organization or scenario.

Conclusions

We were able to obtain a great deal of information about the DMAs we employed. Overall, we found that a sizeable fraction of SDRs do not match information newly added to the original labels or found in the literature, even in a broadly inclusive search of those sources.

We demonstrated differences among the algorithms in the numbers of SDRs that they detect under recommended threshold settings, and we demonstrated differences in the degree to which those SDRs match a reference event database of safety information derived from sources other than AERS data mining. These differences in the extent of matching may be a function of the thresholds used with the various algorithms, rather than the algorithms themselves.

In addition, during the course of this project, we developed methodology and software for the creation of a reference event database and additional software for blinded adjudication of SDRs. We hope that these will evolve into useful resources for future studies of drug safety data-mining methodology.

Acknowledgements

This work was funded by a grant from the Pharmaceutical Research and Manufacturers of America (PhRMA) to ProSanos Corporation. Dr Manfred Hauben, as a representative of the funding committee of PhRMA, participated in the design of the study, the interpretation of data and the editing of the manuscript. Alan Hochberg, Ronald Pearson, Donald O'Hara and Stephanie Reisinger are employees of ProSanos and their work was funded in part by the PhRMA grant. They were responsible for the design and conduct of the study, data collection management and analysis, and interpretation of the data. Dr Manfred Hauben is a full-time employee of Pfizer Inc. and owns stock in this and other pharmaceutical companies that may market/manufacture drugs mentioned in this article or competing drugs. David Goldsmith, Lawrence Gould and David Madigan participated as members of a project steering committee in the design and interpretation of the study and in the editing of the manuscript. They did not receive funding from PhRMA except for nominal reimbursement of incidental expenses related to the project. All authors participated in the preparation, review and approval of the manuscript. The authors thank Dr Lester Reich for his participation in the signal-adjudication process, and Dr Ivan Zorych of Rutgers University for the software implementation of the GPS algorithm. Patents are pending on technology discussed in this paper (rights assigned to ProSanos Corporation). The authors also thank the anonymous reviewers for many helpful comments received during the review of this manuscript.

Independence model with no rater interactions:
[Frequency count] ~ [Rater] + [Reference Level] + [Algorithm] + [Report Source]
+ [Reference Level] : [Algorithm] + [Reference Level] : [Report Source] +
[Algorithm] : [Report Source]

Deviance residuals:

Min	First quartile	Median	Third quartile	Max
–2.986	–0.572	0.000	0.224	2.407

Null deviance: 22294.13 on 485 df
Residual deviance: 403.70 on 402 df

Analysis of deviance table for Poisson model (terms added sequentially):

Term	df	Deviance	Residual df	Residual deviance	p-Value (χ^2)
<null>			485	22 294.13	
Rater	2	129.70	483	22 164.43	<0.0001
Reference level	5	12 741.92	478	9 422.51	<0.0001
Algorithm	2	2 122.64	476	7 299.87	<0.0001
Report source	8	5 120.60	468	2 179.27	<0.0001
[Ref. Lev.] : [Alg.]	10	89.50	458	2 089.77	<0.0001
[Ref. Lev.] : [Source]	40	107.38	418	1 982.39	<0.0001
[Alg.] : [Source]	16	1 578.69	402	403.70	<0.0001

Model with rater interactions:
[Frequency count] ~ [Rater] + [Reference Level] + [Algorithm] + [Report Source]
+ [Reference Level] : [Algorithm] + [Reference Level] : [Report Source] +
[Algorithm] : [Report Source] + [Rater] : [Algorithm] + [Rater] : [Report Source]

Deviance residuals:

Min	First quartile	Median	Third quartile	Max
–2.507	–0.584	0.000	0.145	2.251

Null deviance: 22294.13 on 485 df
Residual deviance: 370.83 on 382 df

Analysis of deviance table for Poisson model (terms added sequentially):

Term	df	Deviance	Residual df	Residual deviance	p-Value (χ^2)
<null>			485	22 294.13	
Rater	2	129.70	483	22 164.43	<0.0001
Reference level	5	12 741.92	478	9 422.51	<0.0001
Algorithm	2	2 122.64	476	7 299.87	<0.0001
Report source	8	5 120.60	468	2 179.27	<0.0001
[Ref. Lev.] : [Alg.]	10	89.50	458	2 089.77	<0.0001
[Ref. Lev.] : [Source]	40	107.38	418	1 982.39	<0.0001
[Alg.] : [Source]	16	1 578.69	402	403.70	<0.0001
[Rater] : [Algorithm]	4	18.24	398	385.46	0.0011
[Rater] : [Source]	16	14.64	382	370.83	0.551 (NS)

χ^2 test for significance of rater interaction terms:

Deviance: 32.88 on 20 df: p-Value (χ^2) <0.0348

Appendix Fig. 1. Poisson models for the inter-rater variability experiment. **df** = degrees of freedom; **Max** = maximum; **Min** = minimum; **ns** = not significant; χ^2 = chi-square.

Appendix A

Results of Inter-Rater Adjudication Study

An experiment was performed to assess the effect of inter-rater variability in the adjudication process on the number of SDRs detected for various algorithms and report sources. Data-mining results for a subset of five drugs from the main study were selected and presented to three individuals for adjudication, as described in the Methods section. All three individuals had ≥2 years' experience in drug safety data mining. Results of the adjudication and scoring process were tabulated, and GLMs of the Poisson family were constructed with 'rater' in addition to 'algorithm', 'evidence level' and 'report source' as stimulus factors, where 'report source' is derived from the 'report source' field in the AERS database. In a baseline model, 'rater' was included as a non-interacting factor, which simply accounted for the overall difference in the number of SDRs available for scoring from the three adjudicators; in other words, a scale factor. In the full model, interactions between rater and other variables ('reference-match', 'category', 'algorithm' and 'report source') were included. The results of adjudication of reference event database entries by the three individuals are shown in Appendix table 1. Note that rater 3 chose not to assign any SDRs to the categories of 'confounding with demographic/clinical factors' or 'confounding with indication'. Results for the Poisson models are shown in Appendix figure 1. The interaction of 'rater' and 'report source' was non-significant. The interaction of 'rater' and 'algorithm' was statistically significant and the origin of this interaction is not known, since the raters were blind to algorithm. While statistically significant, this interaction accounts for a deviance of only 18.24, which is <0.1% of the total model deviance, and thus is of negligible magnitude. The conclusion is that inter-rater variability should have a negligible effect on conclusions regarding various algorithms.

Appendix Table I. Adjudication of reference event database terms by three raters. 'Matching' refers to those terms that were matched to a signal of disproportionate reporting for at least one of the three algorithms (urn model[19], Gamma Poisson Shrinker[20] and proportional reporting ratio[18]) in a pilot study of five drugs

Status	Term category	Rater 1	Rater 2	Rater 3
Matching	Term is on original US label	327	356	286
	Term not on label	112	122	86
Not matching	Term is on original US label	371	382	450
	Term not on label	94	100	136
Flagged	Confounding	95	181	0
(matching not applicable)	Generic/unable to interpret	203	98	20

References

1. Syed RA, Marks NS, Goetsch RA. Spontaneous reporting in the United States. In: Strom BL, Kimmel SE, editors. Textbook of pharmacoepidemiology. West Sussex: John Wiley & Sons, Ltd, 2006: 91-116
2. Gould AL. Practial pharmacovigilance analysis strategies. Pharmacoepidemiol Drug Saf 2003; 12: 559-74
3. Meyboom RHB, Lindquist M, Egberts ACG, et al. Signal selection and follow-up in pharmacovigilance. Drug Saf 2002; 25 (6): 459-65
4. Hauben M, Reich L. Communication of findings in pharmacovigilance: use of the term "signal" and the need for precision in its use. Eur J Clin Pharmacol 2005; 61 (5-6): 479-80
5. Almenoff J, Tonning JM, Gould AL, et al. Perspectives on the use of data mining in pharmaco-vigilance. Drug Saf 2005; 28 (11): 981-1007
6. 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 Dec; 23 (6): 533-42
7. Martindale W, Reynolds JEF, editors. Martindale: the extra pharmacopoeia. 36th ed. London: The Pharmaceutical Press, 2009
8. Physician's desk reference. 54th ed. Montvale (NJ): Medical Economics Company, 1999
9. Hauben M, Reich L. Safety related drug-labelling changes: findings from two data mining algorithms. Drug Saf 2004; 27 (10): 735-44
10. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf 1999 Feb; 20 (2): 109-17
11. Joint Formulary Committee. British national formulary. 52nd ed. London: British Medical Association and Royal Pharmaceutical Society of Great Britain, 2006
12. Hauben M, Aronson JK. Gold standards in pharmacovigilance: the use of definitive anecdotal reports of adverse drug reactions as pure gold and high-grade ore. Drug Saf 2007; 30 (8): 645-55
13. Naranjo CA, Busto U, Sellers EM. A method for estimating the probability of adverse drug reactions. Clin Pharmacol Ther 1981 Aug; 30 (2): 239-45

14. Venulet J, Ciucci A, Berneker GC. Standardized assessment of drug-adverse reaction associations: rationale and experience. Int J Clin Pharmacol Ther Toxicol 1980 Sep; 18 (9): 381-8

15. Karch FE, Lasagna L. Toward the operational identification of adverse drug reactions. Clin Pharmacol Ther 1977 Mar; 21 (3): 247-54

16. US Food and Drug Administration. Guidance for industry. E2C clinical safety data management: periodic safety update reports for marketed drugs [online]. Available from URL: http://www.fda.gov/cder/guidance/1351fnl.pdf [Accessed 2007 Mar 22]

17. Ashman CJ, Yu JS, Wolfman D. Satisfaction of search in osteoradiology. Am J Roentgenology 2000; 175: 541-4

18. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001 Oct-Nov; 10 (6): 483-6

19. Hochberg AM, Reisinger SJ, Pearson RK, et al. Using data mining to predict safety actions from FDA adverse event reporting system data. Drug Inf J 2007; 41 (5): 633-44

20. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system (with discussion). Am Stat 1999; 53 (3): 177-90

21. Woo EJ, Ball R, Burwen DR, et al. Effects of stratification on data mining in the us vaccine adverse event reporting system (VAERS). Drug Saf 2008; 31 (8): 667-74

22. Hauben M, Patadia VK, Goldsmith D. What counts in data mining? Drug Saf 2006; 29: 827-32

23. Hauben M, Vogel U, Maignen F. Number needed to detect: nuances in the use of a simple and intuitive signal detection metric. Pharm Med 2008; 13: 1178-2595

24. Hauben M, Madigan D, Gerrits CM, et al. The role of data mining in pharmacovigilance. Expert Opin Drug Saf 2005; 4 (5): 929-48

25. Aronson JK, Hauben M. Anecdotes as evidence. BMJ 2003; 326: 1346

26. Hochberg AM, Hauben M. Time-to-signal comparison for drug safety data mining algorithms versus traditional signaling criteria. Clin Pharmacol Ther. Epub 2009 Mar 25

27. Chan KA, Hauben M. Signal detection in pharmacovigilance: empirical evaluation of data mining tools. Pharmacoepidemiol Drug Saf 2005 Sep; 14 (9): 597-9

28. Hauben M. Trimethoprim-induced hyperkalaemia: lessons in data mining. Br J Clin Pharmacol 2004 Sep; 58 (3): 338-9

Correspondence: *Mr Alan Hochberg*, ProSanos Corporation, 225 Market St, Suite 502, Harrisburg, PA 17102, USA.
E-mail: alan.hochberg@prosanos.com