EDITORIAL

Drug Saf 2010; 33 (7): 527-534
0114-5916/10/0007-0527/$49.95/0

A Decade of Data Mining and Still Counting

Manfred Hauben[1,2,3,4] and *G. Niklas Norén*[5,6]

1 Pfizer Inc., New York, New York, USA
2 New York University School of Medicine, New York, New York, USA
3 New York Medical College, Valhalla, New York, USA
4 Brunel University, West London, UK
5 Uppsala Monitoring Centre, Uppsala, Sweden
6 Stockholm University, Stockholm, Sweden

The introduction of database-wide disproportionality screening for signal detection in spontaneous reporting systems (SRS)[1] sparked a renaissance in pharmacovigilance research notable for numerous peer reviewed research articles, three expert working groups/white papers,[2-4] countless meetings, symposia, workshops, graduate school theses and aggressive promotion of proprietary software. In addition to expanding the pharmacovigilance toolkit, this research has yielded ancillary benefits beyond patient safety, including an increased awareness of data quality issues such as case report duplication,[5,6] the importance of adverse event coding terminology,[7,8] the proper definition of *signal* in drug safety,[9] the logic of signal detection[10] and an admonition that conflicts of interest, both intellectual and financial, may not only involve the 'usual suspects' such as software vendors, but also other stakeholders that may not normally come to mind, such as regulatory authorities.[11]

The absence of standard test beds for systematic evaluation of signal detection methodologies is an impediment to progress. A fundamental question is what steps of a comprehensive screening strategy to test and what to compare performance against.[12] Herein, we discuss the value of data mining evaluation studies that continue to appear, and related issues, in the hope of facilitating pharmacovigilance system benchmarking and optimization. As a basis for discussion we focus on three recent performance evaluations published in *Drug Safety*[13-15] (see table I).

1. The Signal Detection Process

As a reference for broader discussion of performance evaluation, figure 1 provides a rudimentary schematic of the signal detection process, which approximates a variety of contexts but is especially apt for signal detection in which first-pass screening is based on quantitative methods. Although *traditional* is often used as a point of comparison for more contemporary computerized approaches, traditional methods have historically included quantitative heuristics, and the concept of disproportionality analysis is not new. The breakthrough of the late 1990s was the advent of computational methods that could support routine disproportionality screening involving hundreds of thousands of drug-adverse drug reaction (ADR) pairs simultaneously.[1]

First-pass screening using traditional quantitative filters or disproportionality analysis yields a list of associations (step I). Because of the numerous potential reporting distortions and the low predictive value of quantitative associations in their own right, these are viewed with a skeptical eye and initially subjected to triages that are

Table I. Characteristics of three recent performance evaluations published in *Drug Safety*

Study characteristic	Hochberg et al.[13]	Bailey et al.[14]	Alvarez et al.[15]
Design	Retrospective	Prospective	Retrospective
No. of drugs	35	6	267
Age of drugs	Relatively new	Mature	Mixture
No. of drug-event pairs	3616 (PRR), 1562 (Urn model), 763 (GPS)	861	6888 (532 + 6356)
Database	AERS (2001–5)	AERS (1968–2006)	EudraVigilance
Methods tested	PRR vs Urn model vs GPS	(M)GPS vs traditional	PRR vs traditional
Concomitant medicines included	Yes	No	No
Stratification	Yes [for (M)GPS only]	Yes	No
Steps tested[a]	I + II	I + II	I (for true positives) I + II (for true negatives)
Gold standard	Multiple levels of evidence (retrospective)	Prospective labelling changes	Retrospective labelling changes
Triages	Yes	Yes	Yes
NND	8.9–12.2[b]	32.1[c]	7.2[d]
Inter-assessor variability studied	Yes	No	Yes
Time-to-signal studied	No	No	Yes
Manual resources measured	No	Yes	No

a Please refer to figure 1 for further details of steps I, II and III of the signal detection process.

b Note the minimal level of evidence required.

c Computed by the authors of this editorial as 225/7 based on information available in the original publication.

d Most likely biased downward because of differential treatment of true and false positives, as further discussed in this editorial.

AERS = Adverse Event Report System; **(M)GPS** = (Multi-Item) Gamma Poisson Shrinker; **NND** = number needed to detect; **PRR** = proportional reporting ratio.

typically clinically oriented (step II).[16,17] This is because cogent clinical information generally trumps quantitative aspects in SRS data, which is never cogent when viewed in a biological vacuum. No computerized method available today mitigates all possible SRS distortions. Triages can include down-prioritizing labelled events and events compatible with the natural history/complications of the disease being treated. Associations passing steps I and II warrant at least a minimal formal investigation in step III – the ubiquitous spontaneous reports case series investigation (case definition, case ascertainment and in-depth clinical analysis). For some this would be a signal of suspected causality.[9] Some will result in regulatory action such as warnings or changes to the core data sheet. Passing step III may also require consideration of complementary sources of information, including existing epidemiological studies, clinical trials and

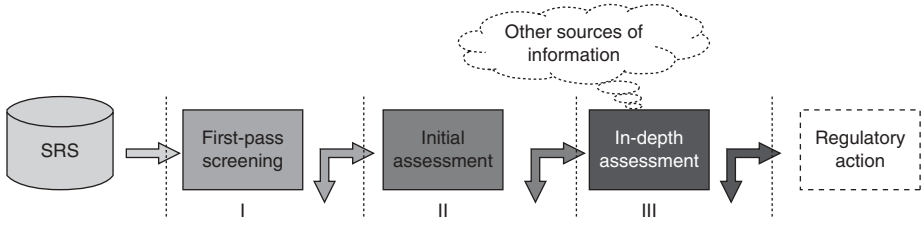

Fig. 1. Schematic overview of the signal detection process. **SRS** = spontaneous reporting system.

toxicological assessments. In the absence of such information, step III might include the generation of such data, i.e. performance of additional studies.

Thus, spontaneous reports are used both for the detection and initial assessment of signals. This seems to violate a paradigm of using independent information sources for hypothesis generation and strengthening. However, to the extent that different aspects of the data (so-called orthogonal information) are considered in each step, it is not so problematic.[18]

2. Replicability

Performance evaluation is typically restricted to a very limited part of the signal detection process. It rarely reflects its inherently exploratory nature, the large space of 'point and click' choices in contemporary data mining software, the rapid query/response cycles between regulators and pharmaceutical companies, and the quantitative and qualitative limitations in the data. Even trying to control for many of the above, there is residual performance variability.[19]

The many design choices affecting data mining performance[20] challenges replicability and detailed convergence of findings. This is not unique to data mining – witness the frequent inability to replicate findings from epidemiological studies.[21] It also makes difficult the identification of key components of a successful but comprehensive signal detection process. At the same time, the application to different datasets is crucial to ensure generalizability.

For example, Hochberg et al.[13] studied 35 relatively new drugs, Bailey et al.[14] studied six mature drugs with established safety profiles, while Alvarez et al.[15] studied 267 drugs in various life-cycle stages. An exclusive focus on a specific life-cycle phase, as in the study of Bailey et al.,[14] may extrapolate poorly to the broad range of scenarios in pharmacovigilance. Hochberg et al.[13] and Bailey et al.[14] used the US FDA Adverse Event Reporting System (AERS) database, containing non-serious as well as serious reports, whereas Alvarez et al.[15] used EudraVigilance, which contains primarily serious reports. Arguably, the absence of reports with only non-serious reactions can affect the results for some serious events as well as non-serious events that are reported and recorded along with serious events.

Nonetheless, while precise replication is unlikely in the absence of canonical datasets and methods, approximate convergent findings from multiple study settings provide insights. For example, studies generally converge in demonstrating that data mining methods tend to identify some credible associations that would otherwise have been missed, and at the same time have limited predictive value in their own right, and are resource intensive in terms of the manual effort required to manage their output. The rational application of triages to downgrade most quantitative associations without further investigation is necessary to make data mining practical.

3. Overall Performance Objectives

Improved signal detection performance is achieved when we identify credible associations that would otherwise escape detection altogether or alternatively detect credible associations at earlier points in time and/or with decreased resources. Systems optimization entails all three aspects. None of the three studies achieves this but each provides pieces of the puzzle that the others do not. For example, Alvarez et al.[15] report a comparison of timeliness for those associations highlighted by both traditional and data mining methods, while Bailey et al.[14] limit their analysis to associations highlighted to date only by data mining and the associated resource costs. Bailey et al.[14] report an investigation in which they complete the entire signal detection process for a selection of quantitatively highlighted associations, and eventually identify seven safety signals that prompted labelling changes. This is a powerful design for optimizing both computerized screening and clinical assessment, in which true positives consist of real safety issues prospectively highlighted in the dataset of interest. Unfortunately, it does not allow for an

evaluation of the number of true safety signals missed, and the set of true positives in their study is very small. Its strong link to the specific data source renders the reference set relevant primarily for the data from which it was generated. The retrospective studies by Hochberg et al.[13] and Alvarez et al.[15] produce reference sets that may prove broadly useful for other validation studies. On the other hand, their points of reference are partially detached from the dataset at hand. A very simple illustration would be an ADR confirmed in a large-scale clinical trial that is not detected by a data-mining algorithm in a specific SRS. Clearly, if there are no spontaneous reports on the drug-ADR pair, the method should not be penalized.

Related to this, the study by Alvarez et al.[15] does not effectively isolate the incremental value of disproportionality analysis over traditional methods, but actually contrasts pairs of methods and data. EudraVigilance was back-populated with reports so that more data from any given time period is available in their retrospective evaluation than was the case at the time in question. This is particularly important to bear in mind when the aim is not to demonstrate the value of an entire signal detection system (including data collection), but to optimize specific screening algorithms within that system.

4. Performance Evaluation within the Signal Detection Process

Is the objective of data mining in pharmacovigilance the identification of associations deemed worthy of investigation, in the sense of not being immediately discountable based on rational scientific grounds, or the identification of associations that are ultimately confirmed (to the extent that we can in pharmacovigilance) on investigation? Expressed a little differently, should the performance of step I be measured relative to best available step II performance, the minimal step III evaluation (in-depth case series review) or, when applicable, a full step III evaluation utilizing all internal and external sources of information? This may be especially pertinent because some organizations have a remit is to identify signals early but not fully evaluate them. On the other hand, other organizations are more directly impacted by full-scale investigation of signals ultimately judged to be false positives, and may tend to judge performance more conservatively. Comparison of step I to step III is perhaps ideal, but requires considerable investigator effort. Optimal performance in this sense would be for first-pass screening to highlight all reporting patterns that survive in-depth assessment (and only those). Under the prevailing notion that computational methods supplement but do not replace clinical review and are implemented in series, an alternative approach is to optimize the algorithms for first-pass screening in step I against the outcome of the initial clinical assessment in step II. The main advantage of this would be lower resource implications. It does require reliable initial assessment, and it accentuates the perspective that computerized screening based on low-dimensional contingency tables is unlikely to outperform clinical review. At the same time, initial assessment can be more or less well supported by systems to highlight relevant reporting patterns, and any improvement of the initial assessment in the future may render an existing reference set outdated.

5. External versus Internal Gold Standards for Performance Evaluation

Constructing a reference set of real associations and non-associations (true positives and negatives) is contentious. One extreme view requires true positives to have a guarantee of causality. However, despite the limitations in SRS data precluding its use for establishing causality except in unusual circumstances,[22,23] the reality is that day-to-day labelling decisions are often made with SRS data. Not every labelled adverse event has been subjected to formal epidemiological assessment (the question of to what extent observational studies truly establish or refute causality notwithstanding[21]). Therefore, a more realistic formulation might include *true probables* and *true possibles* in which the evidence is sufficiently persuasive to merit informing prescribers.[24]

In retrospective studies, true positives may be extracted from various sources, including product labels, standardized drug compendia, published findings from clinical trials and epidemiological studies. Alvarez et al.[15] used product labelling to define their reference set, whereas Hochberg et al.[13] considered a broader range of information sources. Bailey et al.[14] focused on labelling changes originating in excessive reporting rates highlighted prospectively in their own study. Product labels reflect an amalgam of information sources that is rarely accessible in retrospective studies and may reflect truth only for those associations that were detected by legacy methods. Adverse events may be labelled based on any of the above sources of information. Using product labels has the drawbacks of reflecting local signal evaluation practices and leaves the reader in the dark about the strength of evidence available at step I. This is even true in the prospective study by Bailey et al.[14] in which the methods of signal evaluation applied were described only in very general terms. The high rate of inter-assessor variability in the application of simple triages (step II) in the Hochberg et al.[13] and Alvarez et al.[15] studies suggests even greater variability in subsequent signal evaluation (step III). In the Hochberg et al.[13] study, one of three assessors did not attribute any associations to confounding. In the study by Alvarez et al.,[15] one assessor attributed more than 4-fold the number of associations to 'demographic features of the target population' as did the other. Presumably this assessment was done without actually looking at the observed reporting patterns in the data, and higher concordance might have been attained if the assessors had been given information on the empirical reporting patterns across different subsets of the database. Interestingly, disagreement is also seen on less ambiguous questions such as whether an event is on the product label or not.

Hochberg et al.[13] constructed a level of evidence hierarchy based on 378 sources of information in addition to the label. They reported performance of each algorithm at each level of evidence. Notably, for their overall results they used an extremely inclusive reference set (so-called *unlabelled, supported*) in which the true positives included ADRs only supported by reported statistical associations in another SRS, single published case reports and even adverse events reported in clinical trials without a higher incidence for the study drug. Precise documentation of the level of evidence supporting labelledness or the construction of a level of evidence hierarchy in methods testing, as done by Hochberg et al.,[13] is very informative and should be done whenever possible. Detailed information on the origin of specific labelling changes also allows for a more refined analysis of false negatives of both contemporary and traditional methods. For example, Alvarez et al.[15] were able to very precisely determine the source of information that had lead to labelling changes not picked up by disproportionality analysis.

Some favour use of simulated data for performance assessment. An unambiguous reference is naturally an advantage. However, while we may know the truths that we simulate, no study to date has attempted to simulate *all* the truths that we know to exist in SRS data. Most simulations focus on random variation at the expense of the non-random reporting artifacts that plague SRS data, making claims of near 0% false positive rates based on such simulations[25] maybe misleading. Simulation studies reported so far have a narrow scope of applicability, and we will not cover them further in this editorial.

6. A Stricter Structure to Clinical Assessment?

With rare exceptions,[26] studies comparing traditional and contemporary approaches define the information and organizational source of traditional methods, but not what those methods are. This lack of a systematic inventory and assessment of traditional signal detection methods limits comparisons. Our devotion to studying more fashionable computational approaches means less attention spent on understanding and refining the clinical logic and semi-quantitative heuristics that continue to be the core of pharmacovigilance. Indeed, an optimization of

the principles for initial clinical assessment (step II) against the outcome of in-depth clinical review (minimum step III) might well be the most efficient way to improve the overall performance of a comprehensive signal detection process.

7. Assessing Work Impact

An original raison d'être of data mining was reduced labour intensity of effective signal detection. All three studies document the large number of SDRs generated, which entails a significant report review burden in terms of the *number needed to detect*, which quantifies the number of associations requiring in-depth review per credible signal identified. Do note that the definition of what constitutes a credible signal varies substantially between the Hochberg et al.[13] study and the other two studies. In addition, the study by Alvarez et al.[15] subjected only the false positives but not the true positives to initial assessment, introducing a downward bias in the estimated number needed to detect. Initial assessment eliminated over 75% of the false positives in their study and lowered the reported resource cost to the same extent. It would have been interesting to know how many true positives highlighted by the proportional reporting ratio (PRR) would have been filtered out by this assessment, for example as confounded by the underlying disease.

Bailey et al.[14] actually estimate workflow impact in person hours. A one-time data-mining exercise using the most specific of the commonly used metric/threshold combinations involving six products required 184 hours per product assessment/analysis. Before we score one against data mining it must be noted that the latter study did not document the person-time requirement of traditional methods. There is currently no evidence that an association highlighted by data mining is any more or less difficult to evaluate than one highlighted by traditional methods. Therefore, these results provide information on the resources required for signal evaluation, but do not necessarily say more about data mining performance than a false positive rate.

8. Conclusions

Fundamental questions remain. One relates to the advisability of performing intake medical assessment of adverse event reports. Some organizations perform intake medical assessments on all reports, some on serious reports only. Some regard intake medical assessment as wasteful of pharmacovigilance resources. Others think that the benefits of earlier detection of important events still justify its use. An interesting topic for further research is the potential use of computational methods to automatically highlight those cases that have a high probability of informative intake case assessment and reserve clinical review for only those.

The three recent performance evaluation studies[13-15] provide further evidence of how far we have progressed from the extreme views of ''unbridled optimism'' and ''considerable pessimism'' noted by Bate and Edwards.[27] The initial advocacy of quantitative methods in SRS data, which in some instances had an almost evangelical tone, has been displaced by broader experience and the viewpoint that honours these methods as credible additions to the pharmacovigilance toolkit while with significant limitations, just like all available methods and data.

Research initiatives involving signal detection in electronic patient records (e.g. the Sentinel Initiative, Observational Medical Outcomes Partnership [OMOP], Exploring and Understanding Adverse Drug Reactions [EU-ADR] and Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium [PROTECT]) may result in more reliable triages for signals from SRS data using an independent data source. This may allow more reliable differentiation between indication-compatible events from biological effect modification if the intuitive notion that confounding by indication is less prevalent and can be more reliably highlighted in electronic medical records. The EU-ADR programme's incorporation of biological mechanisms may be advantageous in this sense by providing a further filter based on 'orthogonal' information. There are so many nuances to disease etiology and drug action that

technology to display biological pathways would be a great asset.

More than 10 years have passed since database-wide disproportionality analysis first became a realistic option. While our measures of disproportionality have remained largely the same, substantial progress has been made in terms of computerized medical triages and methods to highlight suspected reporting artifacts, even though there is much further to go. More sophisticated screening methods have been proposed, including hierarchical models and shrinkage regression, but have had limited impact on day-to-day pharmacovigilance thus far. This is in part due to the lack of empirical data to properly assess their real-world value. To reliably make progress we must be able to measure it! We hope to see prospective studies of performance against real-world assessment of emerging safety issues, for each step in a comprehensive signal detection strategy. This might drive processes for clinical assessment in the direction of better structure and transparency. We also hope to see more and broader standard reference sets for evaluation of signal detection strategies based on relevant internal and external safety information. Such reference sets will provide standard testbeds for benchmarking and comparison studies that are not restricted to a specific organization, dataset or point in time.

Acknowledgements

No sources of funding were used in the preparation of this editorial. Manfred Hauben is a full-time employee of Pfizer Inc., and owns stock/stock options in Pfizer Inc. and other pharmaceutical companies. Niklas Norén has no conflicts of interest to declare.

References

1. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998; 54 (4): 315-21
2. Almenoff J, Tonning JM, Gould AL, et al. Perspectives on the use of data mining in pharmacovigilance. Drug Saf 2005; 28 (11): 981-1007
3. European Medicines Agency, EudraVigilance Expert Working Group. Guideline on the use of statistical signal detection methods in the EudraVigilance data analysis system [online]. Available from URL: http://www.emea.europa.eu/pdfs/human/phvwp/10646406en.pdf [Accessed 2010 May 5]
4. CIOMS Working Group VIII. Report on practical aspects of signal detection in pharmacovigilance. Geneva: CIOMS. In press
5. Norén GN, Orre R, Bate A, et al. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Discov 2007; 14: 305-28
6. Hauben M, Reich L, DeMicco J, et al. Extreme duplication in the US FDA Adverse Events Reporting System database. Drug Saf 2007; 30 (6): 551-4
7. Henegar C, Bousquet C, Lillo-Le Louët A, et al. Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. Comput Biol Med 2006; 36 (7-8): 748-67
8. Brown EG. Effects of coding dictionary on signal generation: a consideration of use of MedDRA compared with WHO-ART. Drug Saf 2002; 25 (6): 445-52
9. Hauben M, Aronson JK. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. Drug Saf 2009; 32 (2): 99-110
10. Meyboom RH, Lindquist M, Egberts AC, et al. Signal selection and follow-up in pharmacovigilance. Drug Saf 2002; 25 (6): 459-65
11. Erratum. Br J Clin Pharmacol 2007; 64 (1): 118
12. Lindquist M, Ståhl M, Bate A, et al. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. Drug Saf 2000; 23 (6): 533-42
13. Hochberg AM, Hauben M, Pearson RK, et al. An evaluation of three signal-detection algorithms using a highly inclusive reference event database. Drug Saf 2009; 32 (6): 509-25
14. Bailey S, Singh A, Azadian R, et al. Prospective data mining of six products in the US FDA Adverse Event Reporting System: disposition of events identified and impact on product safety profiles. Drug Saf 2010; 33 (2): 139-46
15. Alvarez Y, Hidalgo A, Maignen F, et al. Validation of statistical signal detection procedures in EudraVigilance post-authorisation data: a retrospective evaluation of the potential for earlier signalling. Drug Saf 2010; 33 (6): 475-87
16. Ståhl M, Lindquist M, Edwards IR, et al. Introducing triage logic as a new strategy for the detection of signals in the WHO Drug Monitoring Database. Pharmacoepidemiol Drug Saf 2004; 13 (6): 355-63
17. Levitan B, Yee CL, Russo L, et al. A model for decision support in signal triage. Drug Saf 2008; 31 (9): 727-35
18. Walker AM. Orthogonal predictions: follow-up questions for suggestive data. Pharmacoepidemiol Drug Saf 2010; 19 (5): 529-32
19. Hauben M, Reich L, Gerrits CM, et al. Illusions of objectivity and a recommendation for reporting data mining results. Eur J Clin Pharmacol 2007; 63 (5): 517-21
20. Hauben M, Bate A. Data mining in drug safety: side effects of drugs essay. In: Aronson JK, editor. Side effects of drugs annual. Vol. 29. Amsterdam: Elsevier, 2007: xxxiii-xlvi

21. Taubes G. Epidemiology faces its limits. Science 1995; 269 (5221): 164-9
22. Meyboom RH, Hekster YA, Egberts AC, et al. Causal or casual? The role of causality assessment in pharmacovigilance. Drug Saf 1997; 17 (6): 374-89
23. Aronson JK, Hauben M. Anecdotes that provide definitive evidence. BMJ 2006; 333 (7581): 1267-9
24. Hauben M, Reich L. Response to letter by Levine et al. Br J Clin Pharmacol 2006; 61 (1): 115-7
25. Almenoff JS, LaCroix KK, Yuen NA, et al. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. Drug Saf 2006; 29 (10): 875-87
26. Hochberg AM, Hauben M. Time-to-signal comparison for drug safety data-mining algorithms vs. traditional signaling criteria. Clin Pharmacol Ther 2009; 85 (6): 600-6
27. Bate A, Edwards IR. Data mining in spontaneous reports. Basic Clin Pharmacol Toxicol 2006; 98 (3): 324-30

Correspondence: Dr *Manfred Hauben*, Pfizer Inc., 235 East 42nd Street, New York, NY 10017, USA.
E-mail: manfred.hauben@Pfizer.com