Drug Safety 2006; 29 (10): 827-832
0114-5916/06/0010-0827/$39.95/0

© 2006 Adis Data Information BV. All rights reserved.

What Counts in Data Mining?

Manfred Hauben,[1,2,3,4,5] *Vaishali K. Patadia*[6] and *David Goldsmith*[7]

1 Risk Management Strategy, Pfizer Inc., New York, New York, USA
2 Department of Medicine, New York University School of Medicine, New York, New York, USA
3 Department of Community and Preventive Medicine, New York Medical College, Valhalla, New York, USA
4 Department of Pharmacology, New York Medical College, Valhalla, New York, USA
5 School of Information Systems, Computing, and Mathematics, Brunel University, London, UK
6 Pharmacoepidemiology, Global Safety, Amylin Pharmaceuticals, San Diego, California, USA
7 Goldsmith Pharmacovigilance and Systems, New York, New York, USA

Statistical data mining algorithms (DMAs) are being heavily promoted as valuable adjuncts to conventional pharmacovigilance methods. The majority of contemporary DMAs are variants of disproportionality analysis based upon 2 × 2 contingency tables, the underlying principles, potential benefits and limitations of which have been previously discussed in detail.[1,2]

Disproportionality analysis has been in use for decades in drug safety.[3] However, increased attention has been stimulated by: (i) the commercialisation of data mining software; (ii) the improving technological capacity for calculating measures of association for the millions of 2 × 2 tables represented in large spontaneous reporting system databases; and (iii) the use of Bayesian methodology to dampen the calculated disproportionality metrics generated by small numbers of reports. The latter is a common scenario in drug safety that may be particularly pertinent to large, sparse data sets. The corresponding reduction in the number of drug adverse event pairs presented to the user is cited as an advantage of the Bayesian approaches.[4] Therefore, there is a great interest in fully understanding the comparative performance of various DMAs in naturalistic pharmacovigilance scenarios. This holds true especially for comparisons of frequentist (e.g. proportional reporting ratios [PRRs]/reporting odds ratios [RORs]) versus Bayesian approaches, where the differences may be most apparent, but the comparative performance of Bayesian (e.g. Bayesian Confidence Propagating Neural Network [BCPNN]) and empirical Bayesian algorithms (e.g. multi-item gamma Poisson shrinker [MGPS]) also garners attention. In this context, the drug-preferred term (D-PT) pair has assumed a prominent role as a unit of measurement with some variation of the number/percentage of highlighted D-PT pairs being a standard performance metric.[5-7] This may be presented as summary statistics and scatter-plots without detailed clinical context. The prominent use of D-PT in data mining may reflect the historical prominence of the preferred term (PT) in adverse event coding plus the understandable interest in *a priori* database screening.

Interpreting these research findings is complicated by numerous factors, including the complex, multi-disciplinary nature of the subject and the large space of choices available to the data miner when configuring a data mining analysis.[2] Within the large space of available choices are the selection of the terminology for coding the adverse event and the level of the terms within the dictionary hierarchy. Brown[8,9] and Purcell[10] have explained how the anatomy of the adverse event dictionary can impact the ability to identify disproportionately reported adverse events and Bousquet has eloquently addressed fundamental issues in the construction of concept-oriented vocabularies in pharmacovigilance.[11-13] Related effects can occur from the introduction of new PTs and programmatic mapping of legacy data during dictionary conversions.[14]

Increasingly, researchers from various academic, regulatory and pharmaceutical settings are studying whether the grouping of adverse event terms (e.g. standardised medical queries) or drugs within a pharmacological class may enhance the ability of DMAs to highlight credible associations. There has been less attention to the impact of over-coding on data mining, although Walsh has suggested recording suspect versus concomitant adverse events, akin to suspect versus concomitant drugs.[15]

A related aspect, that to the best of our knowledge has not been discussed, is the impact of the same dictionary architecture and coding practices on comparative assessments of DMAs using real pharmacovigilance data. We have been studying how dictionary and coding-related issues may impact performance comparisons between DMAs and assessing the potential impact of DMA selection on pharmacovigilance workflow. Our preliminary experience suggests a trend towards over-coding and that local organisational factors can affect the reported performance of a given DMA, and that these factors also significantly complicate the interpretation of head-to-head comparisons of DMAs. Our main point, which we help illustrate with a combined tabular-graphic display containing data mining output, is that using the D-PT pair as the unit of measurement/observation may not be the optimum metric for comparing the performance and efficiency of different DMAs especially with very fine-grained adverse event dictionaries. This is not only related to dictionary anatomy and over-coding, but also to the fact that such analysis may not fully account for the process that is actually set into motion within effective pharmacovigilance organisations when an association highlighted by any means is scrutinised with careful clinical analysis. Reported performance differentials between DMAs and work-impact assessments may therefore be partly illusory.

Figure 1 displays data from a single year (2000) for the immunosuppressant drug sirolimus. Two DMAs/metrics were applied to data derived from the publicly available subset of the US FDA safety database after pre-processing to screen for duplicate reports and to adjust for redundant drug nomenclature (WebVDME version 5.0). One algorithm calculated a non-Bayesian metric: the lower 5% cut-off of the ROR (ROR05), and the other calculated an empirical Bayesian metric: the lower 5% cut-off of the posterior interval of the empirical Bayes geometric mean – commonly referenced by its acronym the EB05. A threshold of 2 was used for each and an additional case-count threshold (N >2) was applied with the ROR05 threshold. The adverse effects are listed in alphabetical order.

Inserted into the figure is a dense network of interconnecting arrows linking distinct but medically equivalent or related PTs, which should be examined together when evaluating a signal. For example, three events: interstitial lung disease; PO_2 decreased; and PCO_2 increased appeared uniquely with ROR05. However, from a clinical perspective, these cannot be uncoupled *a priori* from other pulmonary pathology terms highlighted by both metrics. The linkages were constructed based upon the knowledge that in conscientious pharmacovigilance organisations, meticulous investigation of any highlighted associations often implies extension of the search beyond a given preferred term to include PTs that are medically equivalent, overlapping or sufficiently related to warrant further study. Expressed a little differently, when evaluating a potential signal, a relatively wide net is typically cast so that the ‘unit of observation’ in signal detection and evaluation is typically a medical concept(s) and not necessarily a given PT. This is exemplified by the fact that careful evaluation of a potential signal for staphylococcal infection should typically prompt review of all cases coded as bacterial infection as well as abscess formation and osteomyelitis. Indeed, it would be rare and probably inappropriate to restrict the analysis to a specific PT unless the potential signal of disproportionate reporting was so specific and strong that

© 2006 Adis Data Information BV. All rights reserved.

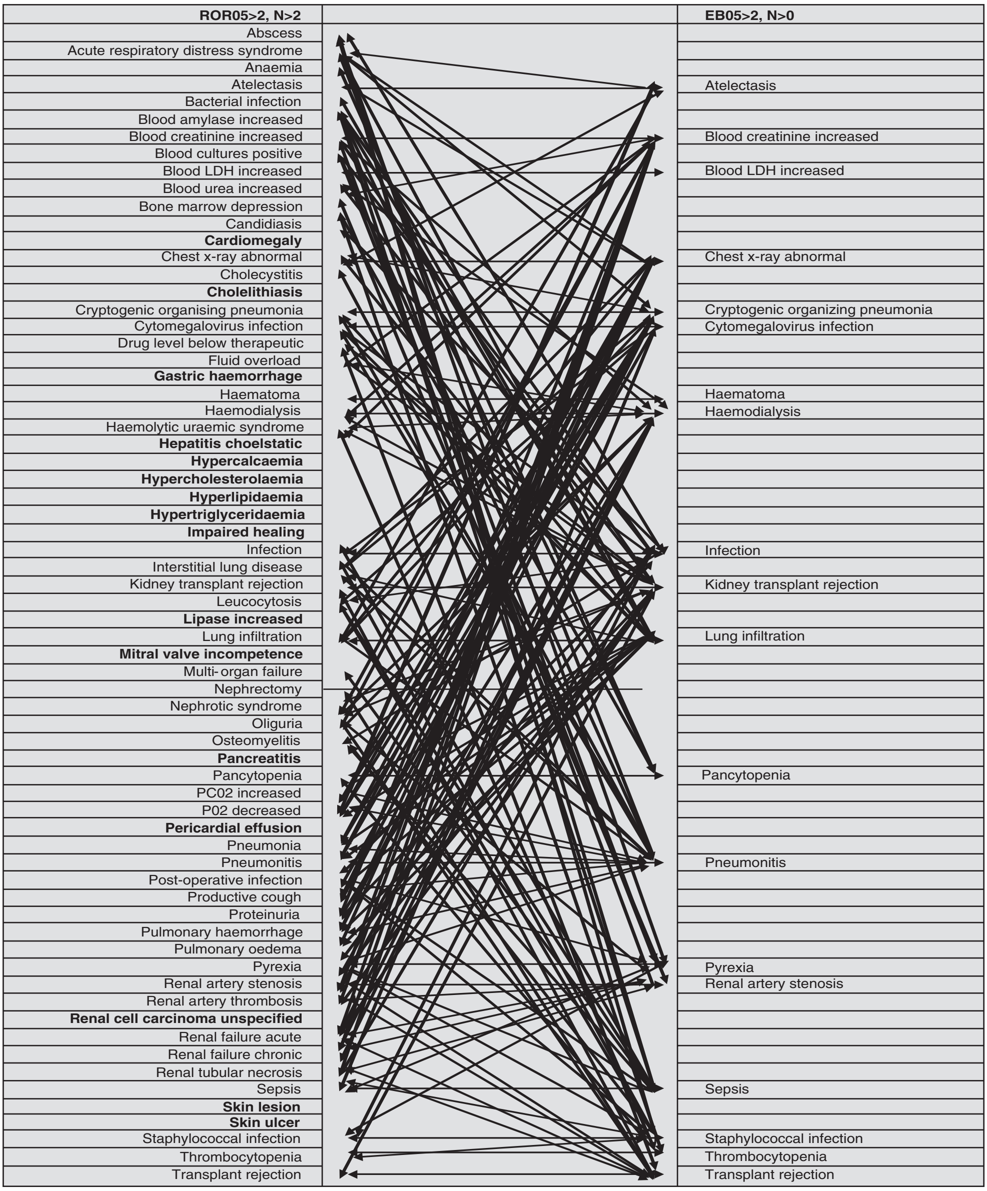

Fig. 1. Data from a single year (2000) for the immunosuppressant drug sirolimus. Drug-preferred term pairs exceeding each statistical threshold are shown and listed alphabetically. **EB05** = empirical Bayesian metric; **LDH** = lactate dehydrogenase; **ROR05** = the lower 5% cut-off of the reporting odds ratio.

further analysis of additional PTs would be unnecessary.

Figure 1 dramatically illustrates the interwoven complexity of preferred terms highlighted by one or both metrics and underscores the caution required in drawing inferences about comparative performance, workflow impact and the great care required in making reference class assignments. The latter may require detailed case-level assessment for complete resolution.

© 2006 Adis Data Information BV. All rights reserved.

It is obvious from the figure that the ROR05 highlights many more D-PTs than the empirical Bayesian metric (EB05) [67 vs 19]. Although the impact of these additional D-PTs would crucially depend on the corresponding number of reports, the time interval for review and how many are already known, the 'penalty' of 49 additional D-PTs may seem impressive. However, when one considers events that are clinically linked, a different view emerges with only 16 additional truly distinct PTs highlighted uniquely by ROR05. Interestingly, in addition to more obvious medical linkages such as hyperlipidaemia-related and pancreatitis-related events, a little thought and prior knowledge reveals linkages between some of the remaining ROR05-specific PTs. For example, refractory bleeding, gastro-duodenal ulcers and pericardial effusions may have a clinically meaningful mechanistic link with impaired wound healing.[16,17] Although some authors report using DMAs to "generate signals without external exposure data, adverse event background information or medical information on adverse drug reactions",[5] our linkages demonstrate the crucial importance of medical information and judgment through all phases of the signal detection and evaluation process.[2] Although the varying perspectives on the existence versus nonexistence of gold standards for causality in pharmacovigilance are beyond the scope of this exposition, we note that there is credible support of several of the associations uniquely highlighted by the ROR05 and they cannot be dismissed out of hand.[16-21] Comparing algorithmic metrics would require a full accounting of the utility of any additional credible associations discovered/discovered earlier against the resources diverted investigating totally spurious associations. Although we have focused on specificity, these factors could result in overestimates of either 'sensitivity' and/or 'specificity', depending on the circumstances. For example, the association of nephrotic syndrome and tubular proteinuria was the subject of considerable attention because of the interest in using sirolimus to mitigate calcineurin-inhibitor nephrotoxicity.[22] Although the PTs nephrotic syndrome and proteinuria were highlighted only by the ROR05, someone using the EB05 might have been directed to the same events when evaluating the signal of disproportionate reporting for other renal events such as blood creatinine increase, although this latter finding is not part of the classic constellation of signs composing nephrotic syndrome. The current exercise involves a comparison of a frequentist metric (ROR05) versus an EB05 but the exact same considerations would apply to a comparison of a Bayesian metric (e.g. the BCPNN's ICLL05) and an EB05. In summary, if meticulous evaluation of a potential signal involves reviewing reports of all medically related events, then the true-performance gradients may be narrower than that reflected by D-PT statistics, particularly when very fine-grained adverse event dictionaries are used.

The extent to which the above delineated effects impact the final analysis may not be fully resolved without case-level clinical review and analysis of the 'developmental anatomy' of the drug's safety profile over time. Additional influential factors include, but are not limited to, the maturity of the drug, the mass of accumulated data, how many events are already known/labelled, the accuracy of the coding process and the distribution of synonymous, equivalent and overlapping PTs across reports. The complexity of the drugs safety profile and treatment populations and prior knowledge of the pharmacovigilance expert will also impact the workload imposed by any signalling algorithm and results for one drug may not be exactly generalisable to all pharmacovigilance scenarios. The influence of dictionary and over-coding may be most significant when trying to design or interpret data mining exercises that take a snapshot of a large, massive and mature database, or on the drug-level, with mature drugs that have substantial mass of accumulated reports. Nevertheless, given the peculiarities of contemporary adverse event dictionaries used in pharmacovigilance, we think it would be beneficial to include a careful accounting of these factors in the methodology section and discussion of comparative assessments of DMAs. This is consistent with one group of researchers from the US FDA who note that, even though specific algorithms are "designed

© 2006 Adis Data Information BV. All rights reserved.

with statistical principles in mind….these properties may not ensure superior performance”.[23]

Just as the drug-PT pair may not be the optimal unit of observation/measurement for purposes of signal detection because it does not reflect the interwoven complexity of contemporary adverse event terminology nor precisely reflect the full spectrum of real-world signal detection and evaluation processes, it may also not be the optimal unit for comparison because its use could possibly lead to overestimates of some performance gradients and inaccurate work-impact assessments. The number of terms and/or cases needing evaluation cannot be estimated from an analysis of only the number of D-PT pairs generated by DMAs. In addition, the analysis of only D-PT pairs by data mining cannot provide a full picture of medically relevant potential signals needed for further analysis. Data mining cannot replace traditional signalling methods but can only be used to supplement standard pharmacovigilance methods. Real-world performance gradients may not be as clear-cut as that suggested by isolated data mining exercises. *A priori* global database screening may still require data mining at the PT level and the number of PTs highlighted by different metrics/thresholds may be interesting and quite dramatic. However, intelligently choosing between DMAs requires a calculus of costs and utilities that may be more tightly correlated with medical concepts/related events than preferred terms and that might make most sense within the framework of real-world pharmacovigilance processes. We recommend that researchers designing and/or reporting systematic comparisons of DMAs avoid vague or imprecise descriptions and provide a detailed and precise accounting of how dictionary architecture, coding practices, medical information and local organisational procedures related to signal detection and evaluation were factored into the analysis and/or its conclusions. Failing to do so, they should at least describe in detail how this omission might limit the interpretation of the findings.

Acknowledgements

No sources of funding were used to assist in the preparation of this commentary. The authors have no conflicts of interest that are directly relevant to the content of this commentary.

References

1. Hauben M, Madigan D, Gerrits C, et al. The role of data mining in pharmacovigilance. Expert Opin Drug Saf 2005; 4 (5): 929-48
2. Hauben M. Data mining in pharmacovigilane the need for a balanced perspective. Drug Safety 2005; 28 (10): 835-42
3. Moore M, Thiessard F, Begaud B. The history of disproportionality analysis measures (reporting odds ratios, proportional reporting rates) in spontaneoue reporting of adverse reactions. Pharmacoepidemiol Drug Saf 2005; 14 (4): 285-6
4. Bate A, Edwards IR. Data mining in spontaneous reports. Basic Clin Pharmacol Toxicol 2006; 98 (3): 324-30
5. Szarfman A, Machado SG, O’Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA database. Drug Saf 2002; 25: 381-92
6. Stahl M, Lindquist M, Edwards IR, et al. Introducing triage logic as a new strategy for the detection of signals in the WHO drug monitoring database. Pharmacoepidemiol Drug Saf 2004; 13: 355-36
7. Kubota K, Koide D, Toshiki H. Comparison of data mining methodologies using Japanese spontaneous reports. Pharmacoepidemiol Drug Saf 2004; 13: 387-94
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. Brown EG. Using MedDRA implications for risk management. Drug Saf 2004; 27 (8): 591-602
10. Purcell P. Data mining in pharmacovigilance. Int J Pharm Med 2003; 17 (2): 63-4
11. Bousquet C, Henegar C, Louet ALL, et al. Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. Comput Biol Med 2006; 36 (7-8): 748-67
12. Bousquet C, Louet ALL, Le Beller C, et al. Appraisal of the MedDRA conceptual structure for describing and grouping adverse drug reactions. Drug Saf 2005; 28 (1): 19-34
13. Bousquet C, Henegar C, Louet ALL, et al. Implementation of automated signal generation using a knowledge-based approach. Int J Med Inform 2005; 74 (7-8): 563-71
14. Hauben M, Reich L. Valproate-induced Parkinsonism: use of newer pharmacovigilance tools to investigate reporting of an unanticipated adverse event with an old drug. Mov Disord 2005; 20 (3): 387
15. Walsh L. Incidental events in spontaneous reports: a proposal for filtering “noise”. Br J Clin Pharmacol 2006; 61 (1): 118-9
16. Smith AD, Bai D, Marroquin CE, et al. Gastrointestinal hemorrhage due to complicated gastroduodenal ulcer disease in liver transplant patients taking sirolimus. Clin Transplant 2005; 19: 250-4
17. Kuppahally S, Al-Khaldi A, Weisshaar D, et al. Wound healing complications with de novo sirolimus versus mycophenolate mofetil-based regimen in cardiac transplant recipients. Am J Transplant 2006; 6 (5 Pt 1): 986-92

© 2006 Adis Data Information BV. All rights reserved.

18. Neff GW, Ruiz P, Madariaga JR, et al. Sirolimus-associated hepatotoxicity in liver transplantation. Ann Pharmacother 2004; 38: 1593-6
19. Fridell JA, Kumar A, Jain B, et al. Phenytoin decreases the blood concentration of sirolimus in a liver transplant recipient. Ther Drug Monit 2003; 25: 117-9
20. Sartelet H, Toupance O, Lorenzato M, et al. Sirolimus-induced thrombotic microangiopathy is associated with decreased expression of vascular endothelial growth factor. Am J Transplant 2005; 5: 2441-7
21. Mahe E, Morelon E, Lechaton S, et al. Cutaneous adverse events in patients receiving sirolimus-based therapy. Transplantation 2005; 79: 476-82
22. Straathof-Galema L, Wetzels JFM, Dijkman HBPM, et al. Sirolimus-associated heavy proteinuria in a renal transplant recipient: evidence for a tubular mechanism. Am J Transplant 2006; 6: 429-33
23. Banks D, Woo EJ, Bowen DR, et al. A comparison of data mining methods in VAERS. Pharmacoepimdeiol Drug Saf 2005; 14 (9): 601-4

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

© 2006 Adis Data Information BV. All rights reserved.