COMMENTARY

Drug Saf 2012; 35 (1): 79-84
0114-5916/12/0001-0079/$49.95/0

© 2012 Adis Data Information BV. All rights reserved.

Terminological Challenges in Safety Surveillance

Andrew Bate,[1,2,3] *Elliot G. Brown*,[4,5] *Stephen A. Goldman*[6] and *Manfred Hauben*[1,2,3,7]

1 Pfizer, Inc., New York, NY, USA
2 New York University School of Medicine, New York, NY, USA
3 Brunel University, West London, UK
4 Elliot Brown (Consulting) Ltd, Leeds, UK
5 PrimeVigilance Ltd, Guildford, UK
6 Stephen A. Goldman Consulting Services, L.L.C., Morris Plains, NJ, USA
7 New York Medical College, Valhalla, NY, USA

1. Introduction

In this issue of *Drug Safety*, Crooks et al.[1] present a novel approach to quantitative signal detection in spontaneous reports of suspected adverse reactions following immunization. The authors depict several novel elements in their research paper, and in particular describe the first application of a Bayesian borrowing algorithm to a spontaneous report dataset. This is noteworthy as it has been argued that the process of classification of disease and health outcomes using hierarchical terminologies (such as Medical Dictionary for Regulatory Activities [MedDRA®]) to populate databases is a major potential limitation to performing quantitative signal detection using algorithms in which terms in an adverse event (AE) dictionary are considered individually.

Most healthcare datasets have structured vocabularies to classify disease and diagnoses, such as Read codes in UK Electronic Medical Record (EMR) datasets, *International Classification of Diseases, 9th edition* (ICD-9) in some US transactional claims databases, and MedDRA® in spontaneous report datasets. While such hierarchical terminologies have clear value in structuring medical information for analysis, there are general limitations with such an approach, and use of the specific terminologies such as MedDRA® can be challenging.[2,3] Surveillance across multiple data streams is complicated by the use of different terminologies in diverse datasets with variable conceptual content and linkages between terms. In such circumstances, efforts to map or link medical terminologies such as the Unified Medical Language System[4] and, more specific to safety surveillance, the Observational Medical Outcomes Partnership,[5] are helpful as the mappings are made available for use by the wider research community.

Quantitative methods are now routinely used in many organizations as part of their signal detection process in order to highlight signals of disproportionate reporting (SDRs) for possible clinical review.[6]

The most commonly used methods for quantitative signal detection in drug safety have focused on individual datasets with a specific hierarchical AE terminology.[7-9] In each approach, an overall appropriate level of precision in medical terminology is pre-specified, with analysis of each AE performed at that level – in the case of spontaneous reports, this is often at a Preferred Term (PT) level. At least in theory, such an approach limits quantitative signal detection capability as the presence of a rare disease that has two or more literally distinct but conceptually similar representations in the database could require increased power from the quantitative approach to detect an underlying

signal. Increased power would also be required if rare diseases are erroneously coded under inappropriate representations in the database. Similar methods have predominated in the analysis of randomized controlled trials (RCTs) or observational data screening[10,11] using different medical terminology (such as ICD-9 codes).

The concept of evaluating different levels of specificity of medical terms in analyses has been examined in the data mining literature. Bate et al.[12] investigated selected case studies and proposed that the level of specificity of terms in the WHO Adverse Reaction Terminology dictionary could be best defined by case-specific clinical judgement, rather than by rigid use of one level of terminology. This approach is useful when further investigating a particular drug-AE combination, but is impractical for large-scale screening of all possible drug-AE combinations. Pearson et al.[13] systematically compared quantitative signal detection performance using different levels of the hierarchy for MedDRA®, and showed that performance metrics of several algorithms improved consistently from PT level to High-Level Term (HLT) level to Standardized MedDRA® Query (SMQ). However, the authors commented that broader-level terms might have significant limitations – some SDRs might be generated as a result of the grouping of disparate PTs, with not all of them being clinically relevant to the generated signal. This would necessitate an additional step to examine the PTs located under the group terms to ensure that they contribute appropriately to the SDRs. This is because the groupings in MedDRA® (such as HLTs) were designed for the purpose of data retrieval (for finding similar terms within a database) rather than for representing medically related concepts for data mining. The PTs within SMQs are likely to be more pertinent in this regard as the SMQs were designed specifically to bring together PTs across System Organ Classes (SOCs) in order to identify cases relevant to pre-specified medical conditions that occur as adverse reactions. However, broad (as distinct from narrow) SMQs may include terms that are very general in nature and not specific to a given medical condition. This is particularly important with syndromes with signs and symptoms that overlap across a clinical spectrum, such as neurotoxicity associated with use of psychiatric medications.[14]

The medical informatics literature is another source of approaches that have not fully diffused into the drug safety literature, and are important to consider. These other approaches include the use of compositional terminologies and semantic reasoning approaches,[15,16] which as targeted surveillance in spontaneous reports is an alternative to that performed through the use of SMQs. Compositional terminologies include a relatively small number of fixed primitive concepts (at least compared with a hierarchical terminology such as MedDRA®), which are defined with the intention that they will be combined by the terminology user as needed to classify a medical condition appropriately. The very large numbers of permutations that are possible in such terminologies offer potential greater expressivity in coding, but can also create greater challenges in analysis. Grouping semantically similar terms by means of expert opinion to construct a case definition is important in second-pass analysis when conducting case review, often the initial formal step in evaluation of a signal, but is impractical in first-pass screening. This is due to the sheer number of potentially important adverse reaction concepts precluding individual consideration prior to initiating data mining runs.[17] Any approach that could, in a semi-automated manner, appropriately consider separate recorded terms as semantically associated could therefore have the potential to improve signal detection and represent the next generation in the use of quantitative signal detection algorithms.

In fact, some of the aforementioned medical informatics work report exactly that. A hierarchy of concepts and relations is constructed to achieve more clinically meaningful linkages and groupings of AE codes than is possible when based solely on placement within the dictionary hierarchy. In this way, data sparsity can be mitigated, as the average case count in the database is increased, ideally in a clinically meaningful way. Bousquet et al.[15] reported such an increase in a subset of the French pharmacovigilance database with use of this tool, along with an attendant enhancement in performance of data mining

© 2012 Adis Data Information BV. All rights reserved.

techniques. Illustrative examples might include considering both myocarditis and cardiomyopathy PTs when screening for heart muscle disorder,[12] or using MedDRA® primary and secondary SOC locations for data mining within a single body system. In the latter example, cases with cardiac arrhythmias (with terms primarily located in the Cardiac disorders SOC) would then be automatically associated with terms for sudden death (with a primary location in the General disorders SOC) and counted together.

2. Bayesian Borrowing

Bayesian approaches have been applied to quantitative signal detection to reduce the number of false positive SDRs for rarely reported drug-event combinations when reporting of both the drug and AE itself are rare. This is accomplished by combining the observed data with prior assumptions within a so-called Bayesian framework. Similar to the null hypothesis in an observational study, the prior assumption is that there is no relationship between drug and AE in the absence of any data to the contrary. A prior assumption can be just a subjective assessment of independence of drug and AE (supported by empirical evaluations) or may be generated by using the global patterns of reporting of drugs and AEs, where naturally overall there is no relationship between most drugs and AEs. The combination of a naïve or raw estimate with other information, with the intention of improving the statistic, is referred to as 'shrinkage' in the statistical literature. Either way, such a Bayesian implementation reduces extreme variability in the statistical algorithm scores when a relative lack of data can result in spurious signals being highlighted due to chance alone based on very few cases.

However, in both methods, all AEs are considered equally. In the elegant work presented in the current issue,[1] the authors suggest that instead of developing a data-driven prior that considers all AEs to be semantically equidistant from one another, i.e. all AE terms are considered equally similar or different from one another, why not create separate priors for each group of similar AEs? These can be defined by SOC, which leads to variable levels of shrinkage depending on the scores of semantically similar AEs. For example, a cardiomyopathy SDR for one drug would make the same drug with myocarditis have less strong shrinkage, and thus require less data to be an SDR in itself. The clinical, knowledge-based reasoning used by Crooks et al.[1] is similar to the aforementioned semantic terminological reasoning, albeit to a different end. Crooks et al.[1] ultimately use semantic reasoning to fine tune prior probabilities of ratios of observed to expected reporting frequencies, while the cited medical informatics studies use the reasoning to increase case counts to mitigate sparsity. This borrowing approach has been used with RCT data,[18] but Crooks and colleagues[1] present an adaptation of the method and have applied it to spontaneous report data.

3. Strengths and Weaknesses of Approach and Paper

As the authors themselves admit, there is much further research needed to see if the theoretical promise of their work (i) translates to practical benefit, and (ii) generalizes to drugs (as distinct from vaccines). It will also be interesting and valuable to determine if such an approach can be effectively applied to surveillance of other forms of real-world data, such as EMR databases. Further study would need to demonstrate that the performance characteristics of the new approach are better than the simpler, routine quantitative signal detection algorithms in use currently. Specifically, it would need to confirm that the new approach returns unique and credible associations that would not have otherwise been highlighted, or produces earlier, consistent highlighting of SDRs. At the same time, the false positive rate needs to be minimized, along with clear demonstration that implementation of this new approach can significantly and positively impact effectiveness of an overall signal management programme that utilizes data mining.

The paper by Crooks et al.,[1] as with much of the data mining literature, does not address in detail the difference between statistical and clinical significance in pharmacovigilance. This is

© 2012 Adis Data Information BV. All rights reserved.

magnified if one does not consider signal detection findings in the context of the overall real-world pharmacovigilance process, which is a continuum from signal detection through signal evaluation and then usually risk assessment, followed by risk minimization actions as deemed appropriate. As previously discussed, the granular nature of the MedDRA® dictionary is an important aspect in this regard[19] because, typically in pharmacovigilance, when an SDR is identified and survives an initial first-pass triage, a case definition based on MedDRA® terminology is constructed to extract all relevant cases for detailed clinical review, not just those events that are statistically highlighted.

The results of Crooks et al.[1] include differential findings between a frequentist, i.e. classical non-Bayesian statistical approach, and Bayesian calculation that well illustrate this point. The Bayesian analysis uniquely identified 'signals' of α-fetoprotein increased (hepatitis vaccine), hepatocellular damage (hepatitis vaccine), liver fatty (hepatitis vaccine) and aphthous stomatitis. To decide whether this represents a real, practical improvement over the frequentist calculation, one would need to know that the uniquely highlighted terms represent significantly new information or result in earlier detection of credible signals necessitating clinical evaluation. The hepatic AEs highlighted by both Bayesian and frequentist calculations would probably prompt the same or similar MedDRA®-based clinical case definitions for detailed clinical review, leaving the incremental utility of these additional SDRs unresolved and probably situation-dependent. However, if further experience demonstrates that such techniques would lead to more focused and precise case definitions based solely on statistically highlighted terms, or to a priority 'short list' case definition that could provide the same answer from a reduced number of cases reviewed, this could increase pharmacovigilance efficiency.

Similarly, did the Bayesian analysis provide a real performance boost by removing the following 'signals' or, expressed a little differently, are the following associations that were selectively highlighted by the frequentist calculations clearly spurious: gingivitis (tetanus toxoid); tooth discolouration (measles vaccine); and increased stool frequency (polio vaccine)? We should avoid being seduced by algorithms with an elegant or extensive mathematical framework, into assuming practical value. The removal of increased stool frequency is perhaps the least likely to be obviously spurious given that oral polio vaccine uses a live, attenuated enteric virus that may establish infection in the gastrointestinal tract (i.e. biological plausibility) and has been reported as associated with diarrhoea in a published observational pharmacovigilance analysis.[20] Tooth discolouration with measles vaccine might appear, at first glance, to challenge credulity, but it is worth noting that the measles vaccine contains a live attenuated virus, and systemic postnatal infections such as measles infection can cause enamel hypoplasia and tooth discolouration.[21] Of course one could use some of these facts to argue that reporting of these associations may reflect a pro-reporting bias. It is rare to read or hear that Bayesian shrinkage does anything but improve performance by removing noise, but there is no currently available method that perfectly removes noise without also removing signal.

Given the paper's focus on vaccines, consideration of the confirmatory use of other databases and further investigations (e.g. observational studies) of signals detected from large spontaneous reporting databases has particular relevance. Such databases can provide information as to a particular vaccine brand, but it is possible that different reporting rates for AEs of concern from one brand versus another may not constitute an actual signal. This was demonstrated by the use of a computerized record linkage system in a retrospective cohort study to investigate US FDA Vaccine Adverse Event Reporting System (VAERS) data that suggested an increased number of serious and death reports in children who received one recombinant hepatitis B vaccine brand.[22] Concluding that actual difference between serious event rates in temporal association with the vaccines was unlikely, the investigators observed that VAERS analysis results, being subject to a passive surveillance system's intrinsic limitations, emphasize the importance of employing other methods to evaluate preliminary findings from any such AE reporting system.

© 2012 Adis Data Information BV. All rights reserved.

While it is beyond the scope of this commentary to fully describe a state-of-the-art signal detection and evaluation strategy for use by regulators and industry, it is always worth emphasizing that data mining methodologies, even the most advanced and sophisticated, may serve as significant parts of such programmes, but do not constitute the programme itself. The limitations of large databases, coding inconsistencies, case report quality variation and other factors addressed earlier emphasize that electronic methods and algorithms are not determinants of the clinical relevance of a serious adverse reaction. Their value lies in bringing a possible signal of concern to the attention of trained clinicians whose medical expertise, experience and judgement are utilized in assessing the reports themselves in consultation with colleagues in other disciplines (e.g. pharmacology, epidemiology), and considering further investigations and possible actions taken in consideration of public health.

We cannot help but contemplate the possible synergies between the work of Crooks et al.[1] and the medical informatics research discussed above. The real nugget in the work by Crook et al.[1] is the elegant Bayesian borrowing, with the terminological reasoning a stepping stone to achieving implementation. In fact, the need for an expert gastroenterologist to construct a hierarchical classification may be an impediment to its use in real-world, routine pharmacovigilance. Semi-automated terminological reasoning, referred to in the literature as 'ontological', and its integration with data mining methodologies was the central focus of the medical informatics literature discussed. Therefore, it seems natural that there might be synergies between these two research domains yielding interesting and useful research findings by allowing semi-automated implementation of Bayesian borrowing and more extensive and systematic head-to-head comparisons of data mining alone versus data mining + terminological reasoning versus data mining + terminological reasoning + Bayesian borrowing.

Finally, we should be dispassionate about the strengths and weaknesses of drug safety surveillance, without exaggeration or minimization. What are appropriate resources for developing and testing new methodologies? In a sense, any unexpected serious adverse reaction is one too many and an opportunity for thought and discussion about process improvement, but issues such as data quality and benefit-risk assessment continue to be major factors in the public health impact of pharmaceutical safety surveillance. The fundamental questions of what is a realistic goal for pharmacovigilance, and how to define its success, remain essential for industry, regulators, healthcare professionals and patients worldwide.

Acknowledgements

No sources of funding were used to prepare this commentary. Andrew Bate and Manfred Hauben are full-time employees of Pfizer, Inc., and, as part of a compensation package, receive stock options from Pfizer, Inc. Manfred Hauben owns stock in other pharmaceutical companies. Elliot Brown and Stephen Goldman have no conflicts of interest that are directly relevant to the content of this commentary.

References

1. Crooks C, Prieto-Marino D, Evans SJW. Identifying adverse events of vaccines using a Bayesian method of medically guided information sharing. Drug Saf 2012; 35 (1): 61-78
2. 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
3. Goldman SA. Adverse event reporting and standardized medical terminologies: strengths and limitations. Drug Inf J 2002; 36 (2): 439-44
4. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med 1993; 32 (4): 281-91
5. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. Ann Intern Med 2010; 153 (9): 600-6
6. Hauben M, Bate A. Decision support methods for the detection of adverse events in post-marketing data. Drug Discov Today 2009; 14 (7-8): 343-57
7. 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
8. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat 1999; 53 (3): 177-90
9. Evans S, Waller P, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001; 10 (6): 483-6
10. Norén GN, Hopstadius J, Bate A, et al. Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Discov 2010; 20 (3): 361-87
11. Norén GN, Hopstadius J, Bate A, et al. Safety surveillance of longitudinal databases: methodological considerations. Pharmacoepidemiol Drug Saf 2011; 20 (7): 714-7

© 2012 Adis Data Information BV. All rights reserved.

12. Bate A, Lindquist M, Orre R, et al. Data-mining analyses of pharmacovigilance signals in relation to relevant comparison drugs. Eur J Clin Pharmacol 2002; 58 (7): 483-90
13. Pearson RK, Hauben M, Goldsmith DI, et al. Influence of the MedDRA hierarchy on pharmacovigilance data mining results. Int J Med Inform 2009; 78 (12): e97-103
14. Goldman SA. Lithium and neuroleptics in combination: the spectrum of neurotoxicity. Psychopharmacol Bull 1996; 32: 299-309
15. Bousquet C, Henegar C, Louët AL, et al. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. Int J Med Inform 2005; 74 (7-8): 563-71
16. Avillach P, Mougin F, Joubert M, et al. A semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European eu-ADR project. Stud Health Technol Inform 2009; 150: 190-4
17. Bate A, Evans S. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf 2009; 18 (6): 427-36
18. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. Biometrics 2004; 60 (2): 418-26
19. Hauben M, Patadia VK, Goldsmith D. What counts in data mining? Drug Saf 2006; 29 (10): 827-32
20. Sugawara T, Ohkusa Y, Taya K, et al. Diarrhea as a minor adverse effect due to oral polio vaccine. Jpn J Infect Dis 2009; 62 (1): 51-3
21. Saraf S. Textbook of Oral Pathology. New Delhi: Jaypee Brothers Medical Publishers (P) Ltd, 2006: 36
22. Niu MT, Rhodes P, Salive M, et al. Comparative safety of two recombinant hepatitis B vaccines in children: data from the Vaccine Adverse Event Reporting System (VAERS) and Vaccine Safety Datalink (VSD). J Clin Epidemiol 1998; 51 (6): 503-10

Correspondence: Dr *Andrew Bate*, Pfizer, Inc., 235 East 42nd Street, New York, NY 10017, USA.
E-mail: andrew.bate@pfizer.com

© 2012 Adis Data Information BV. All rights reserved.