CONFERENCE PAPER

Drug Safety 2007; 30 (7): 623-625
0114-5916/07/0007-0623/$44.95/0

Bayesian Confidence Propagation Neural Network

Andrew Bate

WHO Collaborating Centre for International Drug Monitoring, Uppsala Monitoring Centre (UMC), Uppsala, Sweden

Abstract

A Bayesian confidence propagation neural network (BCPNN)-based technique has been in routine use for data mining the 3 million suspected adverse drug reactions (ADRs) in the WHO database of suspected ADRs of as part of the signal-detection process since 1998. Data mining is used to enhance the early detection of previously unknown possible drug-ADR relationships, by highlighting combinations that stand out quantitatively for clinical review. Now-established signals prospectively detected from routine data mining include topiramate associated glaucoma, and the SSRIs with neonatal withdrawal syndrome. Recent advances in the method and its use will be discussed: (i) the recurrent neural network approach used to analyse cyclo-oxygenase 2 inhibitor data, isolating patterns for both rofecoxib and celecoxib; (ii) the use of data-mining methods to improve data quality, especially the detection of duplicate reports; and (iii) the application of BCPNN to the 2 million patient-record IMS Disease Analyzer.

"*Clearly quantitative methods assist in focussing the attention on those areas of the WHO database likely to contain previously undetected signals.*"

Data mining of the WHO database of suspected adverse drug reactions (ADRs) was originally started at the Uppsala Monitoring Centre (UMC) in 1995. Since 1998, data-mining analyses have been routinely performed as part of the overall signal detection process.[1] Its principal function is the objective initial assessment of all the drug-adverse event combinations, highlighting a subset that is then subjected to clinical review.

The information component (IC) is the measure of disproportionality used for that purpose in a tool called the Bayesian confidence propagation neural network (BCPNN). When the drug-event combination is reported as often as expected, based on the overall reporting of the drug and event in the database, the IC value is close to zero. Confidence intervals are calculated and those combinations for which the lower limit of the 95% confidence interval has become newly positive are then reviewed clinically. A dampening effect of the IC value towards zero, particularly influential with lower numbers of cases and expected accounts, is a desirable property that was realised by a Bayesian implementation of the method. Recent developments on the method are described below.

1. Pattern Recognition

In order to extract additional knowledge from the WHO database, methods have been used to find complex patterns in the data.[2]

In the context of drug safety, complex patterns include syndrome-like ADRs where various interrelated components (e.g. signs or symptoms) constitute a unique clinical entity. A syndrome may be

reported as a group of a few symptoms in one case, and then in another case as a different or partially overlapping group of symptoms. It is unlikely that all components of a syndrome are ever listed in each individual case report.

The method aims at automatic pattern recognition, that is, to find possible new clusters of adverse events that are not known to occur together, and not just to find previously known syndromes in new data. It is also intended to find as many patterns as possible within a particular dataset. It is important for the method to be robust in the presence of noise and missing values and, because of the volume of data; it needs to be computationally efficient.

A study was done to identify patterns of suspected adverse events reports with celecoxib and rofecoxib in the database,[3] and the findings were in line with previous publications on the safety profile of these drugs. Work is in progress at the UMC to develop the technique further: to find unknown clusters of adverse events with other variables, different drugs and using other datasets.

2. Data Quality

Duplicates, different case reports describing the same ADR incident, constitute an enormous problem in the WHO database; as for all large datasets of spontaneous reports. The extent of duplication is not well established; but their presence clearly reduces signal-detection capability.

Anonymised records, as present in the WHO database, increase the difficulty of finding duplicates. Duplicates can occur by several mechanisms, such as: (i) the same incident being reported by health professionals (one or several) and also by a pharmaceutical company (the advent of patient reporting is expected to aggravate this problem); (ii) foreign reports that escape the screening and selective importation of domestic reports from a country to the WHO database; (iii) follow-up reports lacking appropriate linkage to the original case reports.

A duplicate detection method based on a hit-miss model was developed at the UMC to identify duplicated reports.[4] The body of evidence on methods for automated duplicates detection is sparse and the proposed model resembles those referred in record linkage literature used for connecting datasets. The IC plays a critical role in the method.

Matches, or 'hits', receive a positive weight, for example, if the drug is listed on the two reports or if the dates of onset are the same, this is a 'hit'. Mismatches, or 'misses', receive negative weight, e.g. if the same drug is not listed on the two reports or the dates of onset are different, that gets a negative score. When information is missing, the weight is zero.

The weight for each field is computed and the totals are added together to provide an overall pair score. The score accounts for both the level of agreement and the amount of information in the reports.

The following variables were included in the matching process: age, gender, country, date of onset, drug or substance listed, the event terms and the outcome.

The hit-miss model was piloted in a dataset from Norway with approximately 1600 reports from 2004 including 19 known pairs of duplicates. Seventeen pairs were highlighted by the method as possible duplicates and, of those, 12 were among the confirmed duplicates. The other five highlighted pairs were not labelled as known duplicates (assumed false positives), although at least one of these was later confirmed as a duplicate.

The ability to account for near match on age and date helped in the overall performance. Although clusters of reports submitted by the same individual are not necessarily duplicates, they are interesting in themselves and detecting this type of report clustering can be useful.

3. Data Mining of Clinical Records

The IMS UK Disease Analyzer contains monthly, updated, computerised clinical records of patients, as maintained by general practitioners, which have been followed, in theory, over their whole life time. The records include diagnoses, dates, test results, hospital referrals and admissions, surgical procedures, notes, symptoms, signs and administrative

data. This is certainly a rich data source for finding drug safety signals.

The dataset of about 2.4 million patients was analysed, including 113 million prescriptions.[5] By rolling back the database, it was possible to follow the evolution of the IC value for the combination of terbinafine and angioedema (an ADR that is now labelled). The first hint of association occurs in 1992 and rises over time to a significant disproportionality. A labelling change for this combination was approved by the US FDA in January 2004. The pilot investigation suggested that routine analysis of IMS data could have highlighted the combination for clinical review in 1999 or even earlier. This demonstrates that there is potential for the method to find signals early in this type of data.

The benefits of quantitative data mining of patient records rather than performing a formal pharmacoepidemiological study are that: (i) hypothesis generating can be done very fast; (ii) the process is robust and can be applied to large volume of combinations; (iii) there is no pre-defined hypothesis and therefore there is more potential to discover the really unexpected.

Weaknesses of data mining of patient records rather than performing formal studies include: (i) when large numbers of multiple tests are executed; and (ii) the data capture and analysis is not optimised for the potential association highlighted. Critical review is needed for such associations, as well as further evaluation, including formal epidemiological studies.

4. Summary

In conclusion, the BCPNN was developed to enhance rather than replace traditional signal procedures in the WHO database. It has been routinely used as an initial filter to focus the clinical review on those associations that are more likely to represent true signals. In all instances, the method aims to generate hypotheses and not test them. Prospective data mining now routinely leads to the highlighting of important signals, e.g. SSRIs and neonatal convulsions.[6]

Quantitative methods have been shown to improve the data quality of the dataset by identification of duplicate reports.

Extension of the data-mining methods to further analyse potential signals is underway, looking for more complex dependencies in the dataset that may be difficult to detect through individual clinical review and in more complex types of data, such as clinical records.

Acknowledgements

No sources of funding were used to assist in the preparation of this paper. The author has no conflicts of interest that are directly relevant to the content of this paper.

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. Orre R, Bate A, Noren GN, et al. A Bayesian recurrent neural network for unsupervised pattern recognition in large incomplete data sets. Int J Neural Syst 2005; 15 (3): 207-22
3. Bate A, Noren N, Orre R, et al. Pattern detection for celecoxib and rofecoxib in the WHO database. Pharmacoepidemiol Drug Saf 2004; 13 (S1): S323
4. Noren GN, Orre R, Bate A. A hit-miss model for duplicate detection in the WHO drug safety database. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2005: 459-68
5. Bate A, Edwards IR, Edwards J, et al. Knowledge finding in IMS disease analyser Mediplus UK database – effective data mining in longitudinal patient safety data. Drug Saf 2004; 27 (12): 917-8
6. Sanz EJ, De-las-Cuevas C, Kiuru A, et al. Selective serotonin reuptake inhibitors in pregnant women and neonatal withdrawal syndrome: a database analysis. Lancet 2005; 365 (9458): 482-7

Correspondence: Dr *Andrew Bate*, Collaborating Centre for International Drug Monitoring, Uppsala Monitoring Centre (UMC), Stora Torget 3, Uppsala, S-753 20, Sweden.
E-mail: andrew.bate@who-umc.org