Drug Safety 2002; 25 (6): 393-397
0114-5916/02/0006-0393/$25.00/0

A Data Mining Approach for Signal Detection and Analysis

Andrew Bate,[1,2] *Marie Lindquist*,[1] *I. Ralph. Edwards*[1] and *Roland Orre* [3]

1 The Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring, Uppsala, Sweden
2 Division of Clinical Pharmacology, Umea University, Umea, Sweden
3 Department of Mathematical Statistics, Stockholm University, Stockholm, Sweden

Abstract

The WHO database contains over 2.5 million case reports, analysis of this data set is performed with the intention of signal detection. This paper presents an overview of the quantitative method used to highlight dependencies in this data set.

The method Bayesian confidence propagation neural network (BCPNN) is used to highlight dependencies in the data set. The method uses Bayesian statistics implemented in a neural network architecture to analyse all reported drug adverse reaction combinations.

This method is now in routine use for drug adverse reaction signal detection. Also this approach has been extended to highlight drug group effects and look for higher order dependencies in the WHO data.

Quantitatively unexpectedly strong relationships in the data are highlighted relative to general reporting of suspected adverse effects; these associations are then clinically assessed.

According to the WHO, an adverse drug reaction signal is 'Reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously'.[1] A signal is an evaluated association between a drug and an adverse drug reaction, which is considered important to investigate further, and which may refer to new information on an already known association. Usually more than a single report is required to generate a signal, depending upon the seriousness of the event and the quality of the information.

The role of the Uppsala Monitoring Centre[2] is to detect such signals from a database of spontaneously reported case reports of suspected adverse drug reactions (ADRs). The case reports are sent from the 67 member countries of the WHO Programme for International Drug Monitoring.

1. Old System of Signal Detection

Previously, every 3 months, lists of new drug-ADR combinations were generated from all new reports received during that time period. This data was sent to a panel of experts for review, and the summaries of the signals detected by the experts were circulated to National Centres. The large volume of generated data made all the listed information impossible to review effectively, and there was a lack of automation and follow-up within the signalling system.

The large size of the data set (more than 2.5 million cases) makes individual case by case assessment impossible when entering cases into the database, therefore clinical review of case series has to be performed. As the volume of data received on a quarterly basis continued to increase it was decided that a new approach to signal detection was needed.[3] In order to focus the clinical analysis of case series on those drug-ADR combinations most likely to be signals, a quantitative approach could be used to filter out noise, and to highlight potential signals.

2. Objectives

A new method was needed that would provide:

- Objective initial assessment of all drug-ADR combinations
- Transparent selection of drug-ADR combinations for review
- A quantitative aid to signal assessment.

To meet these objectives a Bayesian neural network was developed for quantitative signal detection on the WHO database. The method is called a Bayesian confidence propagation neural network (BCPNN) and implements Bayesian statistics within a neural network architecture.

3. Use of a Neural Network

The term ‘neural network’ is used to describe a wide range of different computational architectures. These architectures are used for a diverse range of tasks, including prediction, classification of data sets, and data mining. Neural networks are made up of many simple processors (‘units’), where each unit has a small amount of local memory. Communication channels (‘connections’), which carry numerical data, link the units to each other, so that each unit only operates on local data and on inputs received via connections from other units.

In this application, a neural network is used to search for dependencies in the data set. The network is routinely used to analyse the strength of connections between drugs and adverse reactions and is a very computationally efficient architecture for considering all combinations. The network is used to count all reports in the database: all occurrences of variable x (for example a drug), all occurrences of variable y (for example an ADR) and all occurrences of x and y together. The method is readily adapted to count occurrences of combinations of any pair of variables.

While the routine use of the method does not strictly require a neural network, the neural network architecture properties are used fully when performing unsupervised pattern recognition to search for previously unknown higher order dependencies in the data set.[4]

4. Choice of Measure of Disproportionality

The strength of dependency between a drug and adverse reaction is defined by a logarithmic measure of disproportionality called the information component (IC). This measure is defined as:

$$\mathrm{IC} = \log_2 \frac{p(x,y)}{p(x)\,p(y)}$$

where:
p (x) = probability of a specific drug ‘x’ being listed on a case report;
p (y) = probability of a specific ADR ‘y’ being listed on a case report;
p (x, y) = probability that a specific drug-ADR combination ‘x’ and ’y’, is listed on a case report.
This equation can also be written as:

$$\mathrm{IC} = \log_2 \frac{p(y|\,x)}{p(y)}$$

where:
p (y|x) = conditional probability of ‘y’ given ‘x’, i.e. the probability of a specific ADR ‘y’ being listed on a case report given the information that a specific drug ‘x’ is listed as suspected on that case report.
Thus the IC value is based on:

- the number of case reports with drug ‘x’; and
- the number of case reports with ADR ‘y’; and

- the number of reports with the specific combination of drug and ADR; and
- the total number of reports.

The IC measure can be seen as the calculation of the logarithm of the ratio of observed rate of reporting of a specific drug-adverse drug reaction combination, to the expected rate, under the null hypothesis of no association between drug and ADR.

Thus when a drug-ADR combination is reported more often than expected relative to general reporting of the drug, and general reporting of the ADR, it results in positive values of the IC.

This IC measure is calculated within a Bayesian framework; rather than the IC being a point estimate, the IC is a distribution that changes on addition of new data. It is assumed that a drug and an ADR are independent in the absence of data of either the drug or ADR. A 'prior distribution' is set up based on this assumption. Therefore in the absence of data, the expectation value of the IC distribution E [IC] equals zero. The distribution shrinks (the variance of the IC diminishes in magnitude), as data is received, and the expectation value of the IC distribution either increases or decreases on addition of this data.

Bayesian statistics is used:

- To motivate the use of the information component as a measure of disproportionality useful in highlighting unexpected dependencies: as the IC is the logarithm of ratio of the posterior and prior probabilities, it represents the change in probability on addition of new data.
- To calculate the information component as a distribution, rather than just a point estimate, based on prior and posterior distributions.
- When the network is used for classification or in pattern recognition, as the outputs in the neural network are posterior probabilities.

5. Use of Bayesian Method

A Bayesian statistical approach has been used for a number of reasons:

- Ready applicability to low (and zero) counter values.
- Results can be calculated despite missing data, by providing increased uncertainty in the results.
- Analysis of other variables and multi-variable analyses can be performed.
- There is an intuitive relationship between the point estimate of the IC and its confidence interval. These are merely two properties of the IC distribution, where the point estimate of the IC is calculated from the expectation value, and the confidence interval is calculated from the variance.

6. Use and Implementation of a Bayesian Confidence Propagation Neural Network

A Bayesian statistical approach is used to highlight unexpected dependencies in the data set;[5] implementation in a neural network architecture allows the routine calculation of the strength (or weakness) of all drug-ADR dependencies in the data set. The calculations performed routinely will ultimately be extended to calculate dependencies between all variables in the data set, although currently such calculations are done on an *ad hoc* basis.

These calculations result in an IC and confidence intervals for each IC (more accurately 'probability intervals'), for each relationship in the network. As the IC is also the weight in the neural network, the computational efficiency of the method is further optimised.

This process is used to highlight drug-ADR combinations most likely to be signals, and to look for more complex patterns in the data. This quantitative signal detection methodology is now in routine use for drug adverse reaction signal detection in the WHO database.

The method has been implemented into a new systematic signalling process.[6] Every 3 months the IC and its confidence interval are calculated for all drug-ADR combinations. Combinations, where the lower 95% confidence limit is newly greater than zero, are highlighted for clinical review. If considered signals after clinical assessment these are then communicated externally. Follow-up processes are also used to provide further information on signals, or allow those potentially clinically interesting combinations to be signalled at a later stage when sufficient data has accumulated.

The BCPNN approach is:

- transparent – easy to see what has been calculated
- robust – valid results can be produced despite missing data
- producing reproducible results – making validation and checking easy
- time efficient – network only needs one pass across data.

7. Evaluation of Method

Initial testing of the approach was done by examining whether the IC was positive for drug-ADR combinations we would hope to signal, such as captopril-coughing, while negative for drug-ADR combinations we would hope not to highlight, such as the combinations digoxin-rash, or digoxin-acne. Captopril-coughing is a signal that would have been highlighted much earlier using the BCPNN method.[7] In contrast, neither digoxin-rash nor digoxin-acne was highlighted with this method.[7]

Tests to determine whether we would have found signals earlier with the new method were made both against general reference sources (Physicians' Desk Reference and Martindale) and extant literature reports in another sensitive international signalling system (database of the publication Reactions Weekly).

In the investigation against general reference sources, a retrospective evaluation of 95 combinations highlighted by BCPNN, and 13 not highlighted by BCPNN from 1993 were made. In this study the positive predictive value reached 44.2% since 42 of the 95 combinations would have been found with the BCPNN. The negative predictive value was 84.6%, since only two of the 13 combinations not highlighted, were found to be signals. For the purposes of the study signals were defined as combinations unknown in 1993, but listed in the literature 7 years later. Both of these studies are described in detail in earlier publications.[6,8]

8. Detection of Group Effects

Beyond detecting specific drug–specific ADR signals, the BCPNN method can also be used to examine the relation between a group of similar drugs and the same ADR.[9] Grouping of drugs can be done for example using the WHO Anatomical Therapeutic Chemical (ATC) classification.[10]

An investigation is performed to look at both:

- IC [drug-ADR]
- IC [(ATC group-drug)-ADR] and the two values compared.

The comparison group must be clinically valid and selected objectively before the analysis is performed. The use of this method relies on the amount of relevant data available and the proportion of cases of multiple suspected drugs on case reports. The types of drugs investigated may also effect the usefulness of the approach; for example drugs with multiple drug indications (at potentially different doses) may be particularly problematic. Also differences in use of ADR terminology, such as inter-country variation, will influence the results.

9. Conclusions

This quantitative method has been developed to improve the usefulness of the WHO database of spontaneously reported suspected drug-adverse reactions. The BCPNN method has been shown to be of use in routine signal detection[8] and has also been used in classification tasks, such as analysing organophosphate poisoning cases, and quality in pulp manufacturing process.

The usefulness of the output is influenced by the quality of the data in the database, therefore this method should be used to detect, rather than evaluate signals. The need for clinical analyses of case series remains crucial.

Since there will always be some drug-ADR combinations that can be considered clinically interesting and important, but that are not being reported unexpectedly frequently, other approaches are also needed. The BCPNN is a tool developed to enhance rather than replace traditional signal detection procedures used on the WHO database by: (i) initial highlighting of associations for clinical review; (ii) providing a tool for further analysis of potential signals; and (iii) seamless extension of the method for detection of complex dependencies in the data set.

References

1. Edwards IR, Biriell C. Harmonisation in pharmacovigilance. Drug Saf 1994; 10 (2): 93-102
2. Olsson S. The role of the WHO programme on international drug monitoring in coordinating worldwide drug safety efforts. Drug Saf 1998; 19 (1): 1-10
3. Bate A, Orre R, Lindquist M, et al. Explanation of data mining methods [online]. Available from URL: http://www.bmj.com/cgi/content/full/322/7296/1207/DC1 [Accessed 2001 Oct 15]
4. Bate A, Orre R, Lindquist M, et al. Pattern recognition using a recurrent neural network and its application to the WHO database [abstract]. Pharmacoepidemiol Drug Saf 2001; 10 (S1): S163
5. Orre R, Lansner A, Bate A, et al. Bayesian neural networks with confidence estimations applied to data mining. Computational Statistics Data Anal 2000; 34 (8): 473-93
6. Lindquist M, Edwards IR, Bate A, et al. From association to alert: a revised approach to international signal analysis. Pharmacoepidemiol Drug Saf 1999; 8: S15-25
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: 315-21
8. 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; 23 (6): 533-42
9. Bate A, Lindquist M, Orre R, et al. Automated classification of signals as group effects or drug specific on the WHO database [abstract]. 8th Annual Meeting European Society of Pharmacovigilance; 2000 Sep; Verona; Elsevier, 2000
10. Guidelines for ATC classification and DDD assignment. 3rd ed. Oslo: WHO Collaborating Centre for Drug Statistics Methodology, 2000

Correspondence and offprints: *Andrew Bate,* The Uppsala Monitoring Centre, Stora Torget, Uppsala, S-75320, Sweden.
E-mail: andrew.bate@who-umc.org