

Quantitative Methods in Pharmacovigilance

Focus on Signal Detection

Manfred Hauben^{1,2,3} and Xiaofeng Zhou⁴

- 1 Safety Evaluation and Epidemiology, 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 Clinical Safety and Risk Management, Worldwide Regulatory Affairs, Pfizer Inc., Ann Arbor, Michigan, USA

Contents

Abstract	159
1. Overview of Signal Detection Methods	160
2. Denominator-Based Methods	162
2.1 Cumulative Sum	162
2.2 Time Scan	162
2.3 Poisson Method	162
3. Observational Epidemiological Methods	164
4. Numerator-Based Methods	165
4.1 Short Memory Schemes	165
5. Proportional Reporting Ratios (PRRs)	165
6. Bayesian Data Mining	167
6.1 Bayesian Confidence Propagation Neural Network (BCPNN)	169
6.2 Empirical Bayes Screening (EBS)	174
7. Discussion of PRRs, BCPNN and EBS	180
8. Summary and Conclusions	183

Abstract

Pharmacovigilance serves to detect previously unrecognised adverse events associated with the use of medicines. The simplest method for detecting signals of such events is crude inspection of lists of spontaneously reported drug-event combinations. Quantitative and automated numerator-based methods such as Bayesian data mining can supplement or supplant these methods. The theoretical basis and limitations of these methods should be understood by drug safety professionals, and automated methods should not be automatically accepted. Published evaluations of these techniques are mainly limited to large regulatory databases, and performance characteristics may differ in smaller safety databases of drug developers. Head-to-head comparisons of the major techniques have not been published. Regardless of previous statistical training, pharmacovigilance practitioners should understand how these methods work. The mathematical basis of these techniques should not obscure the numerous confounders and biases inherent in the data. This article seeks to make automated signal detection meth-

ods transparent to drug safety professionals of various backgrounds. This is accomplished by first providing a brief overview of the evolution of signal detection followed by a series of sections devoted to the methods with the greatest utilisation and evidentiary support: proportional reporting ratios, the Bayesian Confidence Propagation Neural Network and empirical Bayes screening. Sophisticated yet intuitive explanations are provided for each method, supported by figures in which the underlying statistical concepts are explored. Finally the strengths, limitations, pitfalls and outstanding unresolved issues are discussed. Pharmacovigilance specialists should not be intimidated by the mathematics. Understanding the theoretical basis of these methods should enhance the effective assessment and possible implementation of these techniques by drug safety professionals.

The principal objective of pharmacovigilance is the detection of adverse events related to the use of medicines that are unknown or novel in terms of their clinical nature, severity or frequency. This entails the search for preliminary *signals* of such events. In the context of spontaneous reports the WHO defines a signal as 'reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously. Usually more than a single report is required to detect a signal, depending on the seriousness of the event and the quality of the information'.^[1] A single, well documented report with a positive rechallenge could represent a signal, although replication of findings in a series of reports is more often required. A signal does not establish that drug and event are causally related but suggests that further investigation may be warranted to clarify the observed association. Since other sources of data can supply signals, a more general definition is 'an alert from any available source that a drug may be associated with a previously unrecognised hazard or that a known hazard may be quantitatively (more frequent) or qualitatively (e.g. more serious) different from existing expectations'.^[2] The objective of this article is to provide a sophisticated yet intuitive comparison of commonly used signalling methods for postmarketing drug safety surveillance.

1. Overview of Signal Detection Methods

Different signal detection methods have different data requirements. Spontaneous reporting systems constitute the major source of data for this activity but drug utilisation/sales estimates and/or observational epidemiological data may be required according to the method used.^[1,3-6] Pre- and postapproval safety databases have complementary strengths and limitations for signal detection and evaluation. Postmarketing safety databases derived from spontaneous reports contain substantially larger volumes of data than premarketing databases but are uncontrolled, lack exposure data and are vulnerable to numerous systematic biases that complicate signal detection and analysis. Application of quantitative techniques to such uncontrolled and biased data is problematic. Premarketing safety databases include data from well controlled and defined populations and are amenable to quantitative statistical analysis but the size of clinical trial databases precludes reliable detection of rare adverse events, adverse events with prolonged latencies and risk factors. Using quantitative techniques with larger postmarketing safety databases could enhance the detection of previously unrecognised safety issues by objectively utilising the huge amount of data they typically contain.

The simplest methods of signal detection involve periodic review of crude frequency data by

expert human analysts.^[1-3] Within-drug and between-drug crude numerical comparisons of reporting frequency or time trends are made. An example is the former two threshold signalling method used by the WHO Programme for International Drug Monitoring. Suspect drug-event combinations reported at least twice (level 2) or five times (level 5) were sent to expert reviewers for selection of possible signals.^[1,3,5,6] Drawbacks include the usual limitations and biases of voluntarily reporting systems such as underreporting, lack of exposure measurements, incomplete data, the considerable human prejudgement involved and underutilisation of data, since signals for a given drug are sought using only data for that drug. Fully exploiting the information in postmarketing safety databases with quantitative and automated techniques has received increasing attention from drug developers and regulatory agencies.^[2,3,6-14]

Several statistical methods have previously been suggested for surveillance activities including postmarketing safety surveillance.^[15-21] They can be grouped into two categories: denominator-dependent methods and numerator-based methods. The former use some form of exposure estimates and are mostly designed to detect temporal changes in reporting rates or frequencies by constructing a probability model and corresponding test statistic to assess the probability that observed temporal changes reflect random sampling variability. Most have extremely limited or no support in the form of actual use or testing with spontaneous reports. Examples are cumulative sum (cusum) techniques, time scans and Poisson methods. For these techniques tables of critical values are calculated or the respective test statistic is calculated for specific situations.^[16-21] Observational epidemiological databases have also been used for signal detection. Section 3 briefly discusses these methods.

Numerator-based methods, which are self contained in that they do not require access to external data sets for exposure estimates, include short memory schemes, proportional reporting ratios (PRRs)^[1,5,7,9,10] and Bayesian data mining,^[3,6,11-15]

among which, PRRs and Bayesian data mining are currently the focus of attention from regulators and drug developers. Although the underlying models vary considerably, a common feature of these methods is an assessment of how much the observed reporting frequency of a given drug-event combination deviates from that expected, given statistical independence between drug and event. They have the greatest evidentiary support from postmarketing data but still raise many questions about underlying assumptions, probability models, validation and generalisability of findings across databases. Numerous commercial vendors offer 'off-the-shelf' software to implement automated signal detection in postmarketing databases. The method of PRRs has already been incorporated into the routine surveillance activities of the UK Medicines Control Agency (MCA).^[8] The Uppsala Monitoring Center (UMC) which has technical operational responsibility for the WHO database incorporates one of the Bayesian-based signal detection strategies known as the Bayesian Confidence Propagation Neural Network (BCPNN) methodology in routine pharmacovigilance activities.^[6] Other national spontaneous reporting centres and drug safety research units routinely use various numerator-based methods such as empirical Bayes screening (EBS), reporting odds ratios (RORs) and incidence rate ratios (IRRs).^[12,22,23] Despite the increased use and availability of these methods, most pharmacovigilance professionals may not be familiar with many of the underlying theoretical concepts originating in information theory and mathematical statistics. Readers without prior statistical training should be able to understand how these methods work and be able to assess them critically. After a brief review of denominator-based and older numerator-based methods, this article focuses in detail on the three methods of automated signal detection with the most widespread application, namely PRRs and two Bayesian methods (BCPNN and EBS). The objective is to provide a technical yet intuitive evaluation of these methods.

2. Denominator-Based Methods

2.1 Cumulative Sum

Cusum techniques exploit the fact that positive and negative deviations around a mean tend to cancel (which is why variance involves a squared deviation term).^[16,17] Accumulated deviations (cumulative sums) from the mean or control value in sequential observations would be unlikely to exceed 'control limits' unless a nonrandom process truly increases the underlying population mean. If the cusum exceeds a threshold value an alert is detected. The threshold is determined by the average run length (ARL₀) under background incidence K_0 , the standard deviation k_0 , and the difference $K_1 - K_r$, where K_1 is an increased incidence level and $K_1 > K_0$ is considered to be the alert or rejection level, and K_r is the reference level and is usually taken as the mean of K_0 and K_1 , that is, $K_0 < K_r \leq K_1$. The ARL is the average time until the threshold is exceeded and an alert is detected. Cusum techniques have been applied to surveillance of adverse reactions to vaccines, birth defects and other rare diseases but have limited evidentiary support in pharmacovigilance. A closely related sequential sampling technique known as Wald's sequential probability ratio test has demonstrated potential value in the early detection of signals during post-marketing surveillance.^[24]

2.2 Time Scan

Time scan methods compare sales-adjusted adverse event rates between current and historical comparison reporting periods.^[16,17] If the rate in the current reporting interval exceeds that from the historical period an assessment is made of how likely this is due to binomial sampling variation (approximate Poisson process with high sales volume). The three parameters are sales volume, the population probability of a given adverse event per unit sales and the number of reported adverse events. Given the current reporting period ($i = C$), the historical comparator period ($i = H$), the number of reports of adverse events in a given time

period x_i , the sales volume per reporting period s_i , and the population probability per unit sales p_i , X_i follows a Poisson distribution (figure 1) with parameter $s_i p_i$. With large sales this approximates a binomial distribution with (equation 1):

$$P(X_c = x | t) = [t! / ((t-x)! x!)] R^x (1-R)^{t-x}$$

$$x = 0, 1, \dots, t$$

with (equation 2):

$$R = s_C / (s_C + s_H)$$

as the sales ratio, and (equation 3):

$$t = x_H + x$$

This method has been used by regulatory authorities, such as the US FDA, to look for signals of increased frequency of serious labelled adverse events. The FDA amended its regulations on expedited reporting to revoke the requirement for increased frequency reports based on its determination that it did not result in the timely identification of safety problems requiring regulatory action.^[25] Multiple same-region report bias may contribute to the high false-positive rate with standard serial increased frequency analysis.^[26]

2.3 Poisson Method

The Poisson method is a straightforward application of statistical theory to postmarketing safety data with somewhat more support than the above referenced methods.^[20,21,27] While cited as a statistical approach to signal detection, it is more accurately considered a method for evaluating an existing signal.^[20] It requires spontaneous adverse event reports as well as 'external' data in the form of estimated background incidence of the adverse event and level of drug utilisation.

With estimated background incidence of an adverse event and the number of patients treated, rare coincidences of drug and event are modelled by a Poisson distribution with the probability of at

The familiar binomial distribution is given by the following expression:

$$P(x) = [n! / (x!(n-x)!)] p^x (1-p)^{n-x}$$

Where $P(x)$ is probability of x success in n trials, n is number of trials, and p is probability of success on a single trial.

There is an important connection between the binomial distribution and the Poisson distribution. Consider a process where the probability of the success in any one trial is extremely small, the average probability of success in n trials given by $\mu = np$, and with an extremely large number of trials. Substituting $p = \mu/n$ in the binomial formula gives

$$P(x) = [(n)_x / n^x] [\mu^x / x!] [(1 - (\mu/n))^{n-x}]$$

As $n \rightarrow \infty$ for fixed x and μ , the terms in n in the first bracket approach 1 and the terms in the last square bracket $\rightarrow e^{-\mu}$ since $(1 - \mu/n)^n \rightarrow e^{-\mu}$ as $n \rightarrow \infty$. Therefore:

$$\lim_{n \rightarrow \infty} P(x) = e^{-\mu} \mu^x / x!$$

This equation is used in application to approximate the binomial distribution by the Poisson distribution when the success probability p is small and n is large. The important features of the Poisson process are that it is a discrete, one parameter (μ) distribution with mean and variance equal to μ .

Another important property of the Poisson distribution is that it forms a conjugate pair with the gamma distribution. For now this means that the Poisson distribution can be used with another distribution known as the gamma distribution to model uncertainty in the parameter (μ) of a Poisson process and the new distribution will also have the parametric form of a gamma distribution.

We will see in section 6.2 how to model additional uncertainty in a Poisson mean by using related distributions such as the negative binomial distribution. This added uncertainty could be detected if the variance of the data significantly exceeds the mean.

Fig. 1. Poisson distribution.

least x coincidences of statistically independent drug and events per time period given by (equation 4):

$$\text{Prob}(X \geq x) = 1 - \sum_{e}^{-\mu} \mu^x / x!$$

where μ is the expected or average number of cases given the background incidence (b) of the event and the estimated level of drug exposure (n), that is, $\mu = nb$. The critical number of cases for rejecting the null hypothesis of statistical independence between drug and event is $\text{Prob}(X \geq X_{cr} \text{ when } H_0) \leq \alpha$. If the number of adverse event reports is greater than X_{cr} , the hypothesis of independence between drug and event is rejected.^[20,27]

An illustration of the Poisson technique is the association of spinal and epidural haematoma (ex-

tremely rare adverse events after spinal and epidural anaesthesia) after neuroaxial blockade and preoperative thromboprophylaxis with a low-molecular-weight heparin.^[21] The table of critical values shows the need for the number of spontaneous reports, background incidence of the adverse event and the estimated exposure level (table I).

For a given critical alpha level the critical number of adverse event reports can be read off from the table. If 100 000 patients receiving low molecular weight heparin underwent spinal or epidural anaesthesia the probability of at least four reports would be 0.005. The 22 spontaneous reports of this event actually submitted, for example, were therefore much higher than expected for statistical independence. This result is unlikely to be a manifestation of poor baseline risk or drug exposure

Table I. Example of the Poisson method applied to spontaneous reports of spinal haematoma with a low molecular weight heparin (probability of $\geq x$ adverse drug reaction reports of spinal haematoma assuming a background incidence of 1 : 150 000)^[21]

No. of reports	No. of patients receiving low molecular weight heparin undergoing neuroaxial block (in thousands)								
	10	25	50	100	150	200	250	300	500
1	0.064	0.154	0.283	0.487	0.632	0.736	0.811	0.865	0.964
2	0.002	0.012	0.045	0.144	0.264	0.385	0.496	0.594	0.845
3	<0.001	<0.001	0.005	0.030	0.080	0.151	0.234	0.323	0.647
4			<0.001	0.005	0.019	0.046	0.088	0.143	0.427
5				<0.001	0.004	0.012	0.028	0.053	0.243
6					<0.001	0.003	0.007	0.017	0.121
7						<0.001	0.002	0.005	0.053
8							<0.001	0.001	0.021
9								<0.001	0.007
10									0.002
11									<0.001

estimates. Even with a much higher baseline risk of 1/35 000 and with 350 000 patients exposed it would be unusual to receive more than 20 reports. The probabilities of the number of reports observed are actually much lower given underreporting in voluntary surveillance systems.

Since many events detected after approval have low baseline rates, some claim that no more than one to three reports should be coincidental under a Poisson distribution and that for diseases with extremely low baseline incidence such as aplastic anaemia, more than three reports is a strong signal.^[26] The usual limitations of spontaneously reported data, including substantial underreporting, and other extraneous factors apply. Underreporting decreases the statistical power of the method. Accepting higher significance levels can partially compensate for this but given the relatively small variance of the Poisson distribution this would require substantial increases in the significance level.^[20] This method may be most helpful in vaccine safety, where large numbers of healthy individuals receive treatment.

3. Observational Epidemiological Methods

Observational epidemiological databases are used for signal detection.^[4] Although case-control methodology is usually applied to the analysis of

signals already detected by other means, multipurpose case control databases are being used to screen for and clarify signals of drug-event associations. The Slone Epidemiology Unit (SEU) has used this method to detect signals of adult illnesses and fetal malformations using separate databases.^[4] Such data sets may be most useful for drugs involving moderate risk, since relatively common adverse events are often detected in clinical trials and rare events usually are detected with large databases composed of spontaneous reports.

Databases are screened annually by estimating odds ratios for all drugs and drug classes for each adverse event, compared with all other events. Crude and Mantel-Haenszel adjusted odds ratios are calculated. In the adult case-control surveillance system of the SEU this is done for current drug exposure (use within previous 6 months) versus past, and again according to lifetime exposure stratified by treatment duration. Odds ratios are displayed as a printout according to drug and event. To control signal volume additional output criteria are utilised, e.g. at least 50 drug exposures. Associations considered significant are subjected to detailed case-control analysis.^[4]

Two refinements increase the utility of these databases for signal detection. Selective enrolment of cases involving an array of diseases of interest and selective enrolment of controls hospitalised for

diseases that are not usually drug-induced (e.g. appendicitis) can increase efficiency of signal detection for the conditions of interest.

4. Numerator-Based Methods

4.1 Short Memory Schemes

Short memory schemes use conditional binomial tests performed sequentially at fixed time intervals to detect an increase in the mean of a Poisson process.^[17] The number of adverse event reports in the current time period (i.e. $s + 1$ weeks) is compared with the total number of cases during a prior reference time interval (i.e. s weeks where s represents ‘the memory of the scheme’). Since the most recent data are used for memory, it is known as a short memory scheme.^[8] With ‘non-epidemic conditions’ in both memory and current reporting period, the total number of cases should be distributed uniformly within limits of sampling variation. The expected proportion of cases is $1/(s + 1)$ in the current reporting period and $s/(s+1)$ in the memory period. The number of reports in the current reporting period is binomially distributed with parameter $1/(s + 1)$ under the null hypothesis of no difference and $\gamma/(s + \gamma)$ under the alternative hypothesis that there is an epidemic. The null hypothesis of no epidemic would be rejected if (equation 5):

$$\Sigma_y = Y_{,N} [(N!)/((N - y)!y!)]$$
$$[1/(s+1)]^y [s/(s+1)]^{n-y} \leq \alpha$$

where Y = the number of cases/events in the present week and N = sum of cases/events in the preceding s week memory period plus cases in the present week. This technique can be used when the baseline rates are unknown. Application of this technique to drug safety surveillance is limited.

Three numerator-based methods are currently the focus of attention of regulatory agencies and drug developers. These are PRRs, the BCPNN and EBS. The latter two methods are based on the construction of probability models and the application

Table II. 2 × 2 table for proportional reporting ratio calculation

	Specific drug	All other drugs
Specific adverse event	A	B
All other adverse events	C	D

of Bayesian inference. The remainder of this article is devoted to a detailed examination of these three methods.

5. Proportional Reporting Ratios (PRRs)

An early attempt at quantitative analysis of spontaneous reports involved PRRs.^[1,5,7-10] PRRs are analogous to proportional mortality ratios in epidemiology and based on the observation that the proportional frequency of individual adverse reactions reported to the UK Yellow Card system (the spontaneous reporting system in the UK) is relatively constant over time despite the significant increase in total reports. PRRs can be understood by the following 2 × 2 contingency table (table II).

As seen in table II (equation 6):

$$PRR = [a / (a + c)] / [b / (b + d)]$$

Although statistical significance testing in the context of such anecdotal and uncontrolled data is circumspect, it can be used as an automated signal detection criterion. For instance, drug-event combinations with at least three reports, a PRR >3 and a chi-square >5 would represent a signal. An example of PRRs in signal detection is shown using the example of rifabutin-induced uveitis (table III).^[5]

Table III. Proportional reporting ratio (PRR) of uveitis with rifabutin

	Rifabutin	All other drugs
Uveitis	41	754
All other adverse events	14	591 958
Proportion of rifabutin adverse events as uveitis = $0.75 = [41/(41 + 14)]$		
Proportion for all other drugs = $0.0013 = [754/(591, 958 + 754)]$		
PRR = $586 = (0.75/0.0013)$; $\chi^2 = 22\ 740$.		

The performance characteristics of PRRs have been evaluated in both regulatory databases and smaller postmarketing surveillance databases of drug developers.^[7-10] Using a cut-off of $PRR > 2$ and chi-square > 4 for all adverse events occurring with a frequency of > 2 , in the UK Yellow card/ADROIT (Adverse Drug Reaction Online Information Tracking) database about 60% of signals were of known adverse reactions. Approximately 15% were assessed as false-positive signals representing confounding by indication. About 25% of the signals identified by this PRR method were novel and underwent detailed evaluation. A more extensive analysis of the performance of characteristics of PRRs in the MCA database was recently published.^[8] The PRR method was applied to the 15 newly marketed drugs with the highest overall adverse event reporting frequency during 1996–1998 to determine whether it would identify known hazards and possible signals of previously unknown adverse events. In this study signal selection criteria consisted of a $PRR \geq 2$, chi-square ≥ 4 and a crude reporting frequency of ≥ 3 . There were 487 drug-event combinations meeting the specified signal criteria. These potential signals involved 10% of the 'preferred terms'. Six were considered unevaluable. On average, the method detected five unrecognised signals per drug. Of the 481 evaluable potential signals, 339 (70%) were recognised adverse events, 62 (13%) were attributed to underlying comorbidities, and 80 (17%) were judged to warrant further evaluation as potential signals of previously unrecognised adverse events. The details of the individual case causality assessments supporting the latter judgements were not provided. Twenty-two of the 80 newly identified potential signals warranted on-going monitoring, and three singles generated a request to the pharmaceutical manufacturer to amend product information. So, based on these data, the positive predictive value of *any* signal for the 15 selected drugs detected by PRRs was $3/481 = 0.006$, i.e. 0.6% of signals detected triggered regulatory intervention to change product information. The same

three reports constitute 2% [$3/(80 + 62)$] of signals involving *unlabelled events*. The authors did not specify whether these findings were obtained from serial database scans or a single scan covering the entire 3-year time period. Performance characteristics could vary depending on whether serial scan versus a single scan was performed, although the short time interval of 2 years mitigates against a major effect. It should be emphasised that these findings apply to only 15 selected drugs in a database consisting of over 400 000 reports and 600 000 reactions.^[8]

PRRs can be used with other covariates. For a given adverse event-drug combination, the PRR of concomitant medications has been used to detect signals of drug interactions considered well established on the basis of citation in the FDA's 'Dear Healthcare Professional' letters or in the published literature.^[28] This method detected signals for approximately two-thirds of the well established drug interactions, some of which may have been detected prior to the first published literature references. However, no data were provided to document that individual case-causality assessment would have verified the signal and resulted in labelling changes.

PRRs are relatively easy to understand and calculate and are now part of routine surveillance activities at the MCA, so this method has increasing evidentiary support. Computational ease of use is an important advantage considering the dynamic nature of the data and associated sequential scans of increasingly large data sets. Its greatest utility may be in highlighting drug-event combinations with intermediate PRRs, since those with very large scores were noted to involve recognised adverse events (e.g. rifabutin and uveitis), while pairs with PRRs near 1 may be triaged as likely background noise. Care must be taken when strong signals are detected for a given drug, since this will reduce the PRR for other adverse events with that drug. This could be addressed by excluding events with very strong signals.

Two different approaches to statistical inference take fundamentally different viewpoints towards statistical decision-making: Bayesian and frequentist.

'Classical' frequentist seeks the value of a population parameter by assuming it is unknown yet fixed and can be estimated using sample data obtained randomly from the population of interest in a repeatable experiment. It seeks objectivity by restricting the information used to that obtained from a current set of clearly relevant data.

Bayesian inference starts with a preexisting personal assessment of the probability distribution of the parameter in question (the 'prior distribution'). Note that this is a probability distribution of possible probability models. This is why the parameters of the prior distribution are often referred to as hyperparameters (versus parameters of a simple probability distribution). It allows us to weigh various underlying models at the same time. This can be based on previous experience or expert knowledge or may make minimal assumptions (an 'uninformative prior'). This is intuitively plausible since we use previous knowledge and experience to make approximate probability estimates to support our decisions every day. This approximation is made prior to the experiment and mathematically summarises our initial assessment of how likely various values of the unknown parameter are. We then use the data from our experiment to refine this initial ('prior') estimate using Bayes theorem. To understand Bayes theorem it is best to initially look at discrete data that is exchangeable (i.e. the probability of two events E and D occurring together is not dependent on the order in which they occur).

Intuitively the joint probability of two events occurring $P(E,D)$ is related to the probability of E conditional on D occurring multiplied by the probability of D:

$$P(E,D) = P(E|D) P(D) \text{ or } P(D,E) = P(D|E)P(E)$$

Rearranging we have

$$P(E|D) = P(E,D)/P(D)$$

Multiplying both sides by $1/P(E)$, we can thus get the formula in the form of joint probability as follows:

$$P(E|D)/P(E) = P(E, D)/P(E)P(D)$$

From the perspective of signal detection, $P(E|D)$ would be posterior probability of observing a specific adverse event E given that a specific drug D is listed as the suspect drug, $P(D)$ is prior probability that a specific drug D is observed in the entire database, $P(E)$ is prior probability that a specific adverse event E is observed in the entire database, $P(E,D)$ is joint probability that both a specific drug D and a specific adverse event E were observed in the same database coincidentally. The identity of the drug D can be said to provide support for the adverse event E. If the left hand ratio is greater than one this means that the probability of a report listing event E is increased if we know that the identity of the drug is D.

The analogous relationship with continuous probability density functions would be

$$g(E|D) = f(D|E)g(E)/\int_{0,\infty} f(D|E)g(E)dE$$

The updated distribution is the 'posterior distribution'. The posterior probability is in effect telling us how likely a given value of the parameter is given the observed data ($g(E|D)$), based on an initial hypothesis about the parameter of interest ($g(E)$) and the conditional probability distribution of the data. This is the inverse of the usual statistical analysis in which we determine how likely a given value of the data is given the value of the parameter.

Fig. 2. Bayesian inference.

6. Bayesian Data Mining

A major initiative in signal detection is Bayesian data mining.^[3,6,11-15] Data mining integrates several technologies and knowledge domains, including data management, probability and statistics, machine learning, and data visualisation, to

detect novel and potentially useful patterns in large databases in the absence of *a priori* hypothesis.

'Bayesian' refers to the use of Bayesian inference (figure 2) Each Bayesian method seeks to take advantage of the vast amount of information on drugs and adverse events contained in the entire database. Two different Bayesian mining tech-

How much information does the flip of a coin convey? If the coin is rigged so that both faces are the same the flip of the coin does not provide any new information because it does not reduce the level of uncertainty. If the coin is fair the result of a coin flip does convey information because it has reduced the level of uncertainty.

Information theory emerged in the 1940s to describe and quantify information transfer in communication systems. Claude Shannon demonstrated the essential unity of all information media, which could be encoded in a universal language of binary digits or bits. The information component (IC) used in the Bayesian Confidence Propagation Neural Network (BCPNN) can be understood with basic information theory.

Given the connection between information and probability that an event occurs (p), we want our measure of information, $I(p)$, to have several properties: (i) it should be a non-negative number ($I(p) \geq 0$); (ii) if the probability of an event is 1 (the event will definitely occur) the outcome does not convey any information ($I(1) = 0$); (iii) if we observe two independent events, the information communicated is the sum of the individual information quanta ($I(p_1 p_2) = I(p_1) + I(p_2)$) and; (iv) it should be continuous and monotonic, which means that slight changes in probability correspond to slight changes in information.

Logarithmic functions behave this way. From these requirements we justify the following important measure:

$$I(p) = -\log_b(p) = \log_b(1/p)$$

The negative signs ensure that all probabilities have some associated quantity of information. Note that $p = 1/2$ maximises the information content since the absence of any prior data favouring one outcome or another makes the outcome most informative. If we have prior information about the coin (e.g. that it is weighted so that it comes out heads 60% of the time) the outcome is less informative.

That this definition is intuitively appealing can be seen by considering the most elementary unit of information, the binary digit or bit. Let's say we have I bits of information that are used to specify n messages, states, outcomes etc. The number of messages, states, or outcomes is $n = 2^I$. Taking logs we get $I = \log_2 n$. Since the probability of any given outcome or state is inversely proportional to the total number of states we get $I = \log_2 (1/p)$ as above. This also demonstrates that the base 2 logarithm is used since the bit is the most elementary unit of information.

Shannon derived a formula for a quantity known as the 'mutual information' that measures the average reduction in uncertainty or the increase in information about a given variable (e.g. the probability of an adverse event report occurring in the database, $P(X)$) when you learn of the value of another variable (e.g. the suspect drug, Y):

$$\text{Mutual information} = \sum_{ij} P(X_i Y_j) \log_2 [(P(X_i Y_j) / (P(X_i) P(Y_j)))]$$

Notice that given the probability-based definition of information discussed above and the properties of logarithms we could rewrite this expression as the difference in information associated with knowing the joint probability of drug and event appearing on the same report and knowing the product of the individual probabilities (reflecting statistical independence). So the information that X tells us about Y is the reduction in uncertainty about Y due to knowledge about X . If X and Y are statistically independent then $P(X_i Y_j) = P(X_i) P(Y_j)$. Then the mutual information becomes $\log_2 [P(X) P(Y) / (P(X) P(Y))] = \log_2 1 = 0$. This makes sense since if the events are independent the knowledge of one of the variables contributes no new information about the other variable and does not reduce the uncertainty about it.

Fig. 3. Introduction to information theory.

niques have recently been applied to two different postmarketing safety databases.^[3,6,11-15] Both can be conceptualised as consisting of three components: an adverse event data set, a derived data set of signal scores for adverse event-drug associations and visualisation tools designed to graphically highlight the derived signal scores.

One method uses ideas derived from information theory to calculate a quantity known as the information component (IC) for each drug-event

combination in the database (figure 3).^[3,6] The value, precision and time trend of the information content is the basis for signal detection. The other method, which has been applied to the FDA spontaneous reporting system, involves a data-mining technique known as EBS. It ranks drug-event combinations according to how 'interestingly large' the number of reports of that drug-event combination is compared with what would be expected if the drug and event were statistically independent.^[11-15]

A key difference between the two methods is the fact that one provides a stand-alone measure (IC) for each drug-event combination, while the other technique emphasises an overall ranking of drug-event combinations.

These methods visualise the postmarketing safety database as a matrix with rows and columns usually consisting of adverse events and drugs (although in principle other database variables could be used). Cell counts are the number of adverse events reported for each drug-event combination. A model is first chosen to calculate expected cell frequencies (EBS) or probabilities (BCPNN) assuming statistical independence between drug and event. Next, a function is chosen to describe the size of the observed cell counts or probabilities compared with the previously calculated expected cell counts or prior probabilities. The variation in the function is fitted to one of several probability distributions, allowing the analyst to assess whether each value of the function is 'interestingly large' in relation to the expected value of its vari-

ance. The larger the value of this function the greater dependency between drug and adverse event than expected for statistical independence.

6.1 Bayesian Confidence Propagation Neural Network (BCPNN)

Since 1998 the UMC has used Bayesian inference (figure 2) to detect signals in their database of almost 2 million AE reports with 49 data fields per report. The BCPNN operated by the UMC uses neural network architecture (figures 4 and 5) to detect signals.^[3,6]

The BCPNN is imbedded in a neural network because neural networks are self-organising, suited to parallel computation, computationally efficient and provide a simple probabilistic interpretation of network weights.^[3] Computational efficiency may be particularly advantageous with this programme because the BCPNN starts by calculating cell counts for all potential drug-adverse event combinations in the database, not just those that appear together in at least one report. This is ac-

A neural network is a data processing model inspired by the structure and pattern of neurons in living organisms. The fundamental unit of a neural network is an artificial neuron that is an information processing device with multiple inputs and one output. This is analogous to a living neuron with its multiple dendrites and single axon. Neurons are connected by synapses. In biological nervous systems each neuron synapses with multiple inhibitory and excitatory neurons. Whether a neuron fires is determined by the relative weights of inhibitory and excitatory inputs. An example of neural network model of cancer diagnosis is shown in figure 5.

In neural networks large numbers of data processing elements (nodes) are arranged in parallel, highly interconnected layers. Using training datasets this highly interconnected architecture learns to analyse similar but imprecise datasets to solve problems such as pattern analysis, trend recognition or function approximation.

Neural network architecture can be classified by the number of layers of data processing elements and the directionality of signal transmission. The commonest architecture consists of three layers of units: an input layer (where the raw data is fed into network) connected to a middle layer of 'hidden' units that is connected to a layer of output units. Signals travel only from input to output in feed-forward networks. Signals can travel in both directions in a feedback network via feedback loops.

Neural networks learn by example unlike conventional computers that follow a set of unambiguous instructions (algorithm or computer program) to solve problems. This restricts the problem solving capability of conventional computers since the solution to the problem must already be understood. The conventional computer then provides the computational muscle to effect the solution.

When the network is being trained individual neurons are trained to fire or not fire depending on the pattern of inputs. When the neural network is actually being used and a taught input is detected, its associated output becomes the current output. If the input pattern is not exactly one of the teaching patterns, a firing rule that matches the input pattern to the closest training pattern is used to determine if the neuron will fire. Firing rules are important in neural networks and explain their flexibility.

Fig. 4. Introduction to neural networks.

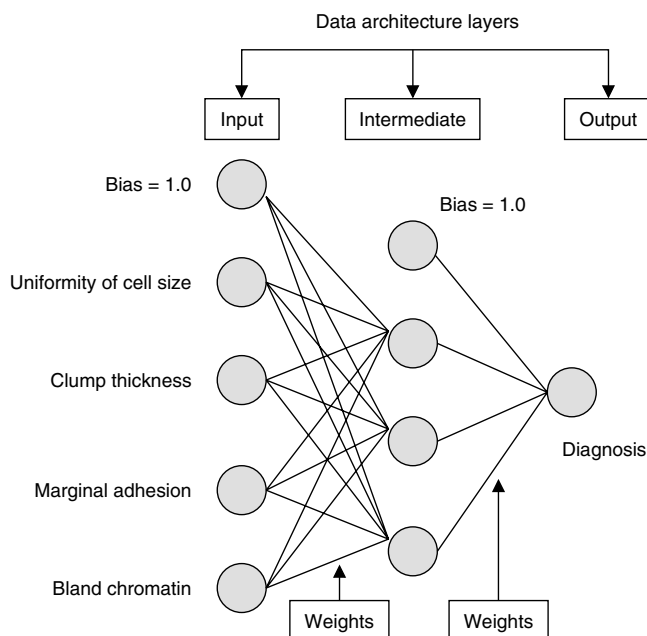


Fig. 5. Neural network model of cancer diagnosis. Potentially predictive clinical variables are input into the first layer of data processing elements. These multiple inputs with their respective weights and a bias value to a specific neuron (a processed clinical variable) result in a single output, which then acts as one of the inputs for the neurons at the next layer until the final diagnosis is achieved.

complicated with two fully interconnected layers, one for all drugs and one for all adverse events.^[3,6]

The IC (figure 3) is calculated for each drug-adverse event combination in the database and monitored with sequential time scans. Denote the overall probability of finding a given event in the database as $P(x)$, the probability for a given drug as $P(y)$ and the probability of finding the drug-event in the same report is $P(x,y)$. The BCPNN calculates the IC for drug-event combinations defined as (equation 7):

$$IC = \log_2 [P(x,y)/(P(x)P(y))]$$

Each of the original probabilities (priors) of drug, event and drug-event combinations are modelled as a beta distribution (figure 6). The joint probability of the drug-event combinations is actu-

ally modelled with a generalisation of the beta distribution known as the Dirichlet distribution. The expected value of each beta distribution is based on the weights of the individual probability models modelled by the beta density. The variance of the beta distributions and the IC (which is a function of the beta distributions) allow the uncertainty of these estimates to be calculated and expressed as 95% confidence intervals (CI).

Using Bayes formula, observed cell counts are used to revise the prior beta distributions. The revised beta distribution is the posterior distribution. The posterior distribution for a given quarterly time scan of the database is then used as the prior distribution for the subsequent quarterly database scan. As more reports are entered into the database over time further revisions are made and the variance of the beta distribution decreases.

If the probability of co-occurrence of drug and event is the same as the product of the individual probability of drug and event (independence between drug and event), the Bayesian likelihood estimator $P(x,y)/P(x)P(y)$ will be equal to 1 (equal prior and posterior probabilities) and since the logarithm of 1 is 0, $IC = 0$. Since $P(x,y)/P(x)P(y)$ is equal to $P(x|y)/P(x)$ (figure 2), as the posterior probability $P(x|y)$ [e.g. probability observing a specific adverse event (x) given a specific suspect drug (y)] exceeds the prior probability $P(x)$ [e.g., probability that a specific adverse event x is observed in the entire database], the IC becomes more positive and, depending on how large it is, may detect a signal. Corresponding 95% CIs are calculated for each IC. In order to detect a signal based on the IC of drug-event associations, sequential time scans of the database are performed given the dynamic nature of the data set. An IC with a lower 95% CI > 0 that increases with se-

quential time scans (e.g. a stabilised positive signal score) is a criterion for signal detection.

Figure 7 shows the IC for captopril and cough over quarterly time scans of the WHO database. Over time the number of adverse event reports of all kinds increase. This increase in the quantity of data increases the statistical precision of the estimated IC as reflected in a narrowing CI (increasing number of adverse event reports). Notice that the lower 95% CI goes from negative to positive and thus detects a signal. Note that the IC became positive in the second time scan in 1983. At that time there were only three reports of captopril and cough in the database.

Similarly, figure 8 shows quarterly time scans for digoxin-acne. Note that the estimated IC becomes more precise and becomes negative with serial time scans.

Figure 9 shows quarterly time scans for the association digoxin and rash. Note that the IC be-

Let us try to understand the nature of the beta distribution and why it is useful in modelling prior probabilities. The beta family of probability distributions are described by two parameters (α, β) and characterised by the following probability density function:

$$f(x) = K, x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1, \alpha > 0, \beta > 0$$

Note that for a beta distribution with parameter values >1, $F(x) = 0$ when $x = 0$ and $x = 1$. This is described as being bounded between $x = 0$ and $x = 1$. Since probabilities (and proportions and percentages) are bounded between 0 and 1(100) this characteristic of the beta distribution is one of the most obvious reasons for using them to model prior probabilities. Beta distributions are also flexible. Altering the parameters provides a rich variety of beta curves with differing peaks, locations, dispersion and skewness. This ability to 'fine tune' the shape of the beta curve is advantageous when modeling samples of probabilities. With multiple outcome variables a distribution known as the Dirichlet is used. This is just a higher dimensional generalisation of the beta distribution applicable when you have more than two joint probabilities to model (e.g. drug+event+, drug-event-, drug+event-, drug-event+).

For a given value of α larger values of β generate curves that are more heavily weighted to low probabilities e.g. the curve peaks at lower probability models and is skewed right. Larger values of the sum $\alpha + \beta$ also make the curve less 'responsive' to observed data. In other words the posterior probability that we obtain by applying Bayes theorem is not as moved away from the prior probability as curves with lower values for this sum. This allows us to choose priors for cells with low cell counts that suggest statistical independence (e.g. information component [IC] approaches 0) and which require an adequate amount of data (e.g. sufficiently large cell counts) to move the posterior away from statistical independence. Thus the probability estimates for cells with low counts is 'shrunk' toward the expected value of the prior distributions by suitable selection of prior hyperparameters contained in IC. This probability estimation is therefore known as shrinkage estimation.

Beta distributions have the added advantage that they form a natural conjugate pair with binomial sampling distributions. So we can make an educated guess of these prior probabilities, update them with the observed binomial sample data, and the updated posterior distribution will have the same parametric form as the prior, namely a beta distribution. The only difference will be the updated parameter values of the posterior distribution that are derived from simple addition of the observed data to the parameters of the prior distribution.

Fig. 6. Beta distribution.

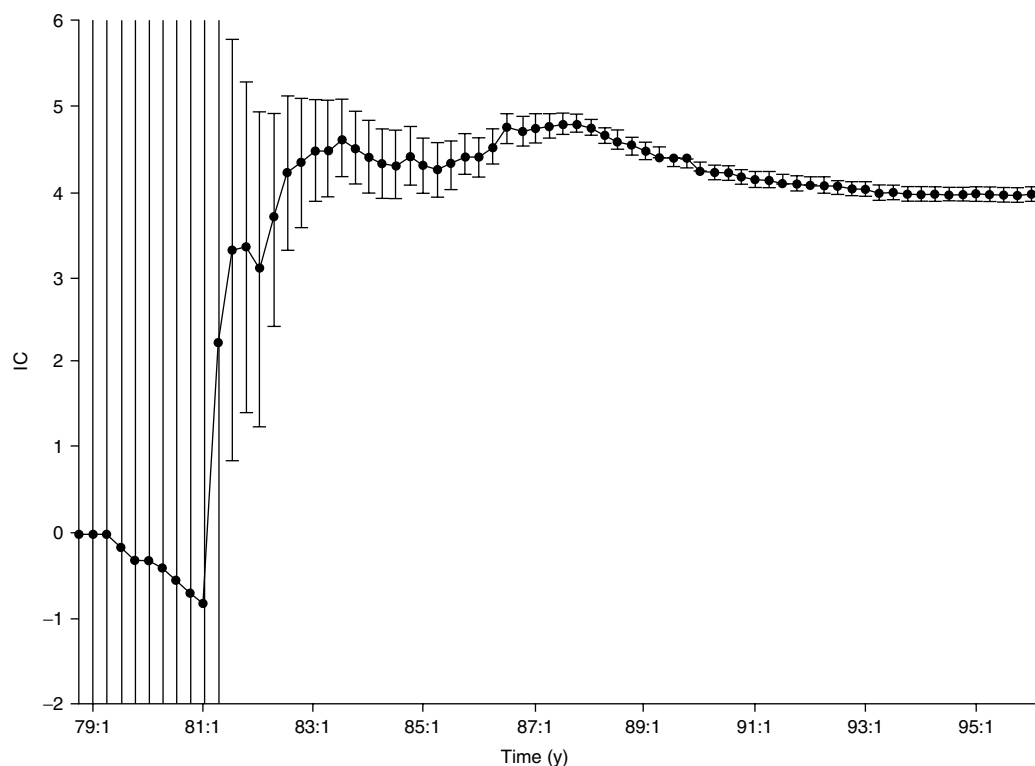


Fig. 7. Change in information component (IC) between 1979 and 1996 for the association of captopril and cough. The IC is plotted at quarterly intervals using data from the WHO database with (95% CI shown).

came positive briefly in 1968 corresponding to the addition of one report of rash with digoxin. The CI for this IC is quite wide, however. More importantly, with subsequent time scans the estimated IC is more precise and clearly negative over time.

The UMC evaluated the performance characteristics of the BCPNN with 'new' drugs (not defined), including a comparison with the former signalling method.^[3] Two established compendia of drug information (Martindale's Extra Pharmacopoeia and the Physicians' Desk Reference [PDR]) were used to assess associations derived from a retrospective (1993) BCPNN quarterly scan of the database. Drug-adverse event combinations for which the lower 95% CI of the IC changed from negative to positive in the first quarter of 1993 were considered a positive association. Drug-

adverse event combinations were classified as negative associations if the lower 95% CI for the IC changed from a positive value to a negative value in the same time period. An association was considered a valid signal if the drug-adverse event combination was listed in the July 2000 online versions of Martindale and the PDR and not listed in the 30th edition (1993) of Martindale. A nonsignal was defined as a drug-adverse event combination that was not listed in the July 2000 on-line versions of Martindale or listed in the 30th edition (1993) of Martindale. The performance of the BCPNN was also assessed against the previous signalling procedure ('level 2' association; see section 2) by looking at all drug-adverse event combinations meeting the 'level 2' threshold in the first quarter of 1993.

A total of 682 drug-event combinations met threshold criteria for a positive association in the time period of interest; 107 of these involved 'new' drugs. Thirty-two drug-adverse event combinations met criteria for negative associations during the same time period, of which 15 involved 'new' drugs. Seventy-one (66%) positive associations with new drugs were not listed in the 30th edition of Martindale. Thirty-six drug-adverse event associations were either compatible with (listed as a WHO high-level term) or specifically referenced in the 30th edition of Martindale. Excluding non-signals (defined as already known drug-event combinations) and cases involving a drug withdrawn during the time period reviewed the positive predictive value was 44% (42/95) and the negative predictive value was 85% (11/13).

Ten drug-adverse event combinations met the level 2 threshold in 1993 and were subsequently

circulated as a signal at various times after meeting threshold criteria. Six of the ten drug-adverse event combinations meeting level 2 threshold were flagged by the BCPNN. Four of the six drug-adverse event combinations met BCPNN signal association criteria prior to being circulated as a signal first detected by the previous signalling procedure.

There are no published comparative evaluations of the BCPNN versus other methods of Bayesian data mining, although there are unpublished data comparing the BCPNN with both PRRs and EBS. A major research focus at the UMC is training the BCPNN in pattern recognition to discover higher level dependencies, including variable combinations never reported on any single adverse event report that could represent previously unrecognised syndromes.

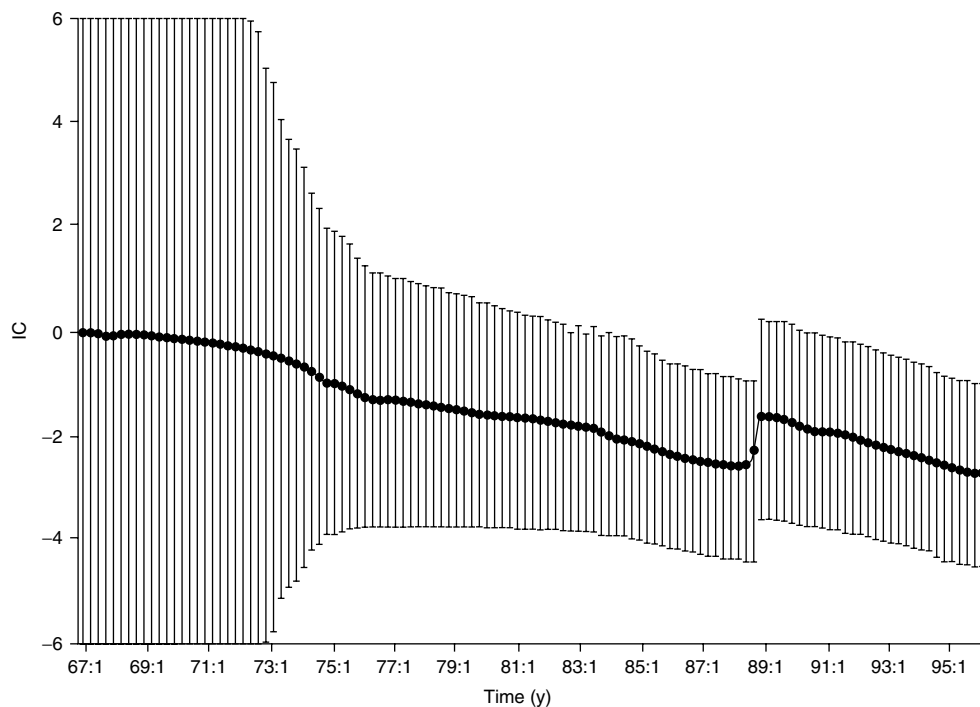


Fig. 8. Change in information component (IC) between 1967 and 1996 for the association of digoxin and acne. The IC is plotted at quarterly intervals using data from the WHO database with (95% CI shown).

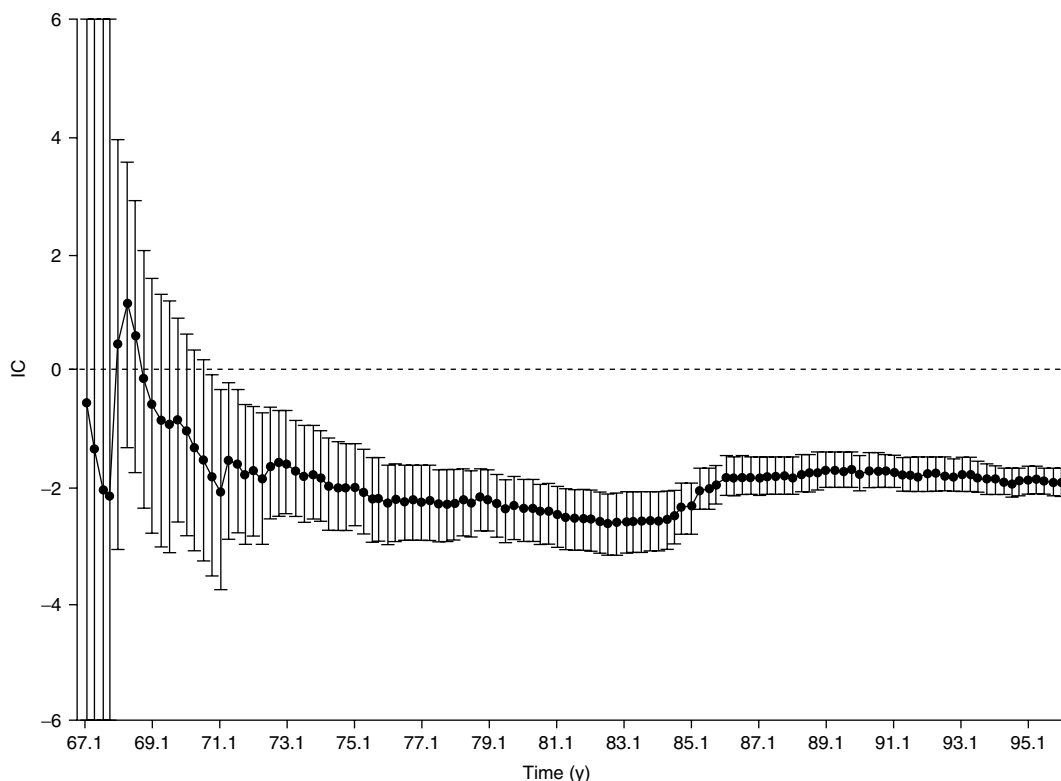


Fig. 9. Change in information component (IC) between 1967 and 1996 for the association of digoxin and rash. The IC is plotted at quarterly intervals using data from the WHO database with (95% CI shown).

The UMC is marketing paid subscriptions to the BCPNN on a per product rate. Subscribers can perform their own signal detection and also receive a copy of the signal report that UMC sends to international regulatory bodies.

Ultimately the utility of a signal detection system is not just determined by the 'goodness of fit' of the model but whether the signal detected results in the discovery of causal associations when individual cases making up the signal are scrutinised by expert clinical reviewers. Although the signal detected for cough and captopril by the BCPNN 'pre-dated' the first published literature reports of this association, this is not in itself informative, since the results of individual causality assessments remain unknown. Unless causality assessments were performed prospectively, the utility of

the first signal cannot be completely evaluated, since association does not prove causation and can have numerous explanations

6.2 Empirical Bayes Screening (EBS)

Another approach to finding 'interestingly large' cell counts is EBS,^[11-15] which was initially developed and applied to the FDA databases of spontaneous reports. This was originally undertaken to explore ways to detect signals of adverse events that were gender-selective.^[28,29] In EBS a baseline frequency or null hypothesis is used that averages cell counts like multiple two-way tables for tests of independence. It is helpful to look at the formula for expected cell counts provided by the originator of this application, DuMouchel,

since it can be expressed in a simple manner for those not used to multiple summation notation. DuMouchel expresses the expected cell count as follows (equation 8):

$$E_{ij} = \sum_k N_{i,k} N_{.jk} / N_{.k}$$

where i = drug, j = event, k = number of strata. The sum of $N_{i,k}$ represents the total number of reports with a given drug (drug i is being summed over all events, j) and $N_{.jk}$ is the total number of reports with the given event (event j summed over all drugs i), and the denominator is the total number of reports (summation over events i and drugs j). Let us examine what this means using a simplified $k(2 \times 2)$ contingency table for drugs and adverse events (table IV). The k represents a third stratification variable. We see that independence between drug and event would give the following formula (equation 9):

$$E_{ij} = \sum_k [(A + B) / (C + D)] \times C$$

If the drug and event are independent the proportional representation of that event for the specified drug should be the same as the proportional representation of that event in the entire database $(A + B) / (C + D)$. Multiplying this by the total number of reports with the given drug gives the expected count for that drug-event combination. Referring to table IV it is easy to see the equivalence of the two ways of expressing expected cell counts. Three distance functions are then used to rank drug-event frequencies according to how ‘interestingly large’ they are:

1. A relative risk measure $RR = N_{ij} / E_{ij}$ (observed/expected cell counts). Although this measure cannot be regarded as a formal relative risk, the concept is intuitive but does not convey the precision of the estimate, so that the same relative risk could have markedly different statistical interpretations depending on the absolute values of numerator and denominator.

Table IV. $k(2 \times 2)$ contingency table for calculating E_{ij}

No. of reports	Given drug (i)	All other drugs	Corresponding notation by DuMouchel
With given adverse event (j)	A	B	$A + B = N_{.jk} = \sum_i N_{ijk}$
Total	C	D	$C = N_{i,k} = \sum_j N_{ijk};$ $C + D = \sum_i \sum_j N_{ijk}$

2. A test of statistical significance, $\log P_{ij}$, which accounts for sampling variability by calculating the probability of obtaining an observed drug-event frequency given the expected or baseline expectation value for a Poisson distribution. The importance of ranking the drug-event frequencies over the absolute value of the signal scores is underscored by the description of the $\log P_{ij}$: ‘The concept behind such a measure is not that the null hypothesis is taken seriously, but only that the test statistic or its degree of significance might be a useful measure for ranking the degree of association among the different cells’.^[11]

3. The $EBGM_{ij} = 2^{EB\log^2_{ij}}$ (the geometric mean of the empirical Bayes posterior distribution of the true relative report ratio).

The last measure requires more explanation. Note that neither of the first two scalar functions incorporates information on the size of the cell plus a corresponding estimate of precision. How do we evaluate drug-event combinations with a given relative reporting rate but differing expected and observed frequencies? A drug-event combination that is observed in 200 reports and has an expected frequency of 20 reports would have the same relative reporting rate of 10 (200/20) as would a drug-event combination reported 20 times with an expected frequency of 2. The smaller numbers of the latter drug-event combination are more unstable statistically.

EBS is able to deal with this problem by calculating the log likelihood function. Assume that the various (equation 10):

$$\lambda s = \mu_{ij} / E_{ji}$$

(where μ_{ij} is an unknown mean of Poisson distribution) for each drug-event (observed count N_{ij}) represent a draw from a common prior distribution of λ s. How can we model the probability distribution of λ ? It turns out that mixtures of Poisson distributions can be effectively modelled using the

family of gamma distributions (figure 10). For enhanced modelling flexibility DuMouchel assumes a mixture of two gamma distributions. This provides greater modelling flexibility by increasing the number of modifiable parameters from two to five (two parameters for each gamma plus a mixing

The gamma distributions may be daunting to those without a background in mathematical statistics. In this figure we try to show why gamma distributions are chosen for modeling uncertainty in Poisson processes in Empirical Bayes Screening.

Gamma distributions constitute a family of two parameter distributions. This means that each member of the family (each curve) is completely specified by two parameters. One of these parameters specifies the shape of the curve and the other specifies the scale of the curve. This family of curves includes many of the most familiar distributions including the exponential and χ^2 distributions.

The formula for the gamma distributions, which we will justify below is

$$P(x; n, \lambda) = [\lambda^n / \Gamma(n)] x^{n-1} e^{-\lambda x}$$

One potential source of confusion to look out for is that different authors use different symbols for the gamma parameters. The most common variation is using α instead of n and $1/\beta$ instead of λ . So another commonly encountered formula for a gamma distribution would be

$$P(x; \alpha, \beta) = [(1/\beta) / \Gamma(\alpha)] (x/\beta)^{\alpha-1} e^{-(1/\beta)x}$$

The standard mechanistic explanation of the gamma distribution is that it describes the probability of observing various waiting times until a certain number of Poisson events take place. So a gamma (3,2) distribution would tell you the probability distribution of the length of time you would have to wait to see 3 adverse events occur given that on average 2 adverse events occur per unit of time. We can gain some insight into the gamma distribution by comparing it with the familiar Poisson distribution.

$$P(x; \mu) = e^{-\mu} \mu^x / x!$$

Poisson Distribution
Discrete distribution

Gives the probability of observing different numbers of rare events in a given period of time.

$$P(x; n, \lambda) = [\lambda^n / \Gamma(n)] x^{n-1} e^{-\lambda x}$$

Gamma Distribution
Continuous distribution

Gives the probability of observing different periods of time for a given number of rare events.

How do we justify the use of a gamma distribution? The simplest way to do this is to consider the event we are modelling (adverse event probabilities) and the shape of the gamma curves. Since by definition and derivation Poisson events are rare (e.g. adverse drug events detected in the postmarketing period) and greater than or equal to zero (e.g. negative rates of occurrence do not make sense) we would want a curve that is bounded on the left by 0 and is skew right (weighted to lower values). This is precisely the shape that can be achieved with certain members of the family of gamma distributions. So it makes sense for rare events with a relatively high variance and for which negative values do not make sense.

Do not be intimidated by the gamma function, Γ , which appears in the denominator of the gamma distribution.

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx$$

Just as factorials of integers appear in the expression of many discrete probability distributions (e.g. $N!$), the gamma function generalises the idea of factorial to non-integer values for continuous probability distributions like gamma.

Another advantage of the gamma distribution is that it forms what is known as a conjugate pair with the Poisson distribution. If we have a hypothesis about the parameter of interest, say that it follows a gamma distribution, we can update our estimate with count data from a Poisson sample and the computation required to revise the prior parameters is limited to simple addition.

Fig. 10. Gamma distributions.

Why is the negative binomial distribution used to model the unconditional observed cell counts? The negative binomial distribution is similar to the binomial but instead of providing the probability of n successes in N trials it describes the probability distribution of the number of trials that are required to achieve n successes given probability of success of q and probability of failure as p .

$$P(n) = [(r+n-1)!(n-1)!r!](q)^n(p)^r$$

Why should the observed drug-event cell counts follow such a distribution? Remember each individual cell represented a Poisson process where the Poisson parameter λ was not single value but had its own probability distribution (gamma distribution). Also keep in mind that a useful diagnostic for a Poisson distribution is whether the mean of the distribution is the same as the variance. It turns out that when there is uncertainty in a Poisson parameter λ , the added uncertainty in the Poisson process itself is reflected as an increase in the variance. The negative binomial distribution is similar to the Poisson distribution but with a greater variance. Intuition tells us that the wide range of cell counts would be over dispersed for a Poisson distribution. The negative binomial distribution would take this increased variance into account. This intuition can be rigorously confirmed by deriving the negative binomial as $\int (\text{Poisson density})(\text{gamma density})$ although we will not work out this problem in integration by parts here.

Fig. 11. Negative binomial distribution.

parameter describing the relative contribution of each gamma). So the unconditional prior distribution of λ is a mixture of two gamma distributions with five parameters as $\alpha_1, \alpha_2, \beta_1, \beta_2$ and P (mixing parameter).

For each individual observed drug-event count N_{ij} , there is a posterior conditional distribution for λ_{ij} . This conditional distribution is also a mixture of two gamma distributions because Poisson and gamma form a conjugate pair. It is a function of the parameters of unconditional gamma distribution as well as the observed and expected counts. In order to evaluate the conditional gamma distribution for each observed count, one needs to first estimate the parameters of the unconditional gamma mixture. First, one would consider the marginal distributions of each N_{ij} , which are negative binomial mixtures. The distributions are derived as a mixture of Poisson distributions, where the Poisson mean has a gamma distribution. Here, intuitively, one may view that the relationship between Poisson and negative binomial distribution is analogous to the relationship between the familiar normal distribution and t-distribution. The likelihood function, which is given by (equation 11):

$$L(\theta) = \Pi_{ij} \{P f(N_{ij}; \alpha_1, \beta_1, E_{ij}) + (1-P) f(N_{ij}; \alpha_2, \beta_2, E_{ij})\}$$

is then used to determine the maximum likelihood estimate of the parameters of the gamma mixture. By maximising this function we are essentially finding which values of the five parameters ($\alpha_1, \beta_1, \alpha_2, \beta_2, P$) would maximise the probability of obtaining the entire observed negative binomial cell counts (figure 11). The initial parameter values chosen to start the MLE (figure 12) process for the gamma distribution are arbitrary and become less important, since the maximum likelihood estimates (figure 12) weighs the observed data heavily. A 'weak rationale' was used to specify the initial parameters of the gamma mixture distributions. It was assumed that most (2 of 3) drug-event combinations have values that cluster around the null hypothesis value of $\lambda = 1$ (i.e. not a signal) and suggests a gamma distribution for this component. The remaining one-third of cells are assumed to the site of drug-event dependencies and are hypothesised to have a gamma with a higher mean and variance.^[11]

Once the parameters of the unconditional gamma mixture are determined, the conditional gamma mixture for each observed cell can be evaluated by updating the unconditional parameters with the observed and expected data for each cell individually. By calculating the expectation value of $\log \lambda$ instead of λ , we get a statistic that is the

Bayesian equivalent of $\log_2 RR$, namely (equation 12):

$$EB\log_{2ij} = E[\log_2(\lambda_{ij}) | N_{ij}] = \\ E[\log(\lambda) | N = N_{ij}] / \log(2)$$

By calculating the expectation value of the log-transformed data we affect our 'shrinkage', since imprecise estimates of λ with wide interval estimates and very small lower bounds from such cells have the lower bound 'shrunk' less by log transformation than the larger upper bounds of these cells (log transformation reduces the variance of the data). This lowers the estimates of cells with low expected cell counts.

A tabular summary of the database scan is created that ranks the cells on all three scores. Working with a commercial vendor, the FDA has developed visual graphical displays using the software CrossGraphs (Belmont Research Inc. 1996) to plot and graph empirical Bayes scores including colour-coded spatial maps of signal scores, fingerprints, chromatographic maps, timeline summaries, and paired female-male signal scores (figures 13 and 14). This system includes interactive drill-down capability.^[11]

In an analysis of the FDA spontaneous reporting system database the technique was applied to about 1.2 million reports listing approximately 4.9 million events. The data were slightly compressed by subsuming trade name drugs under generic drug codes and limiting the analysis to events listed in at least 100 reports.^[11]

Sixty-five combinations were ranked in the top 1000 combinations by all three criteria (RR, $\log P$ [rank], and EBGM [rank]) indicative of the varying performance characteristics of the three criteria. It is useful to compare the characteristics of the drug-adverse event combinations uniquely chosen by different criteria. All cells chosen by only the RR criterion had $N \leq 2$ and $E < 2$. Highly ranked cells by EBGM criteria alone tended to be combinations with moderate N and small, but 'not tiny' expected cell counts ($0.28 < E < 0.63$). Highly ranked cell

by $\log P$ criteria alone tended to have huge observed counts N , large expected cell counts ($E > 193$) and relatively low N/E ratios. Cells with RRs that differ by multiples of hundreds but with varying levels of variance based on cell counts can have the same EBGM due to shrinkage.^[11]

Screening is not limited to drug-adverse event combinations. The strength or ranking of signals has been coupled with logistic regression analysis with predictor variables such as dose and gender. Dependencies between drugs can be used to look for signals of interactions causing a given adverse event.^[30]

The FDA has been attempting to validate the EBS model and its assumptions, noting the need for goodness of fit plus signals that are truly informative of drug-adverse event concerns.^[11] While comprehensive data have not been published, drug-adverse event associations contained in prescribing information and/or identified during previous analysis of the database were evaluated. These known associations were highlighted as 'higher than expected' alerts, and sometimes the alert appeared earlier in chronological time although precise details were not given.^[12] The authors go on to say 'earlier identification of a known association is a confirmation of the model's performance'. However it remains unanswered whether individual case causality assessment would have established a strong enough link to conclude a causal association was present or reasonably likely. Events known not to be a problem with some drugs or drug classes suggest that false-negative signals are not being detected. Details of the validation process that were not provided assume greater importance given the dynamic nature of the database.

An interesting corollary of this work is the finding that within selected drug classes a reasonable correlation (overall correlation >0.63) was observed between the total number of reports naming a drug and independent estimates of drug utilisation obtained from external data sources. The total number of reports with a given drug associated

with a signal score less than 1 may be a surrogate for adjusting relative drug exposures.^[28,29]

EBS is publicly available on the Internet at <ftp://ftp.research.att.com/dist/gps>. The site includes the

gamma Poisson shrinker algorithm program with a user interface, a copy of the paper 'Bayesian data mining in large frequency tables, with an application to the spontaneous reporting system' by

In discussing Bayesian inference we learned that in probability you usually know the parameters of the distribution and try to predict the outcomes. Likelihood is the inverse of this in that we are given observed data and try to figure out which parameter value would be most likely to generate the observed data. Essentially it is the probability density function for the observed data as a function of parameter values. *The maximum likelihood estimate (MLE) is the parameter value(s) that make the observed data most likely which is the peak of the likelihood function.*

This is not the same as probability. A probability distribution returns the probability of observing a given sample of data for a given sample size and parameter. The cumulative probability distribution must sum or integrate to provide a total probability of one. The likelihood returns the probability that different parameter values would produce a single observed sample of data. It does not have to add up to one since there are infinitely many parameter values that could, with varying likelihood, produce the observed sample.

To show how this works we will use an absurdly simple example where we already know the answer and for which we would never have to use the technique to begin with. A simple coin flip experiment can be described as a binomial distribution with parameters p = (probability of obtaining a head) and n = (total number of flips).

$$[n!/(n-h)!]p^h(1-p)^{n-h}$$

Let's say that out of 100 flips the total number of heads (h) is 56. It should be obvious that we already have the answer to the question as to what the MLE is – it is $p=0.56$! But let's see how likely it would be for various values of p to produce the observed data (56 heads and 44 tails).

$$L(p = 0.48|\text{data}) = 100!/56!44! (0.48)^{56} (0.52)^{44} = 0.0222$$

$$L(p = 0.50|\text{data}) = 100!/56!44! (0.50)^{56} (0.50)^{44} = 0.0389$$

$$L(p = 0.52|\text{data}) = 100!/56!44! (0.52)^{56} (0.48)^{44} = 0.0581$$

$$L(p = 0.54|\text{data}) = 100!/56!44! (0.54)^{56} (0.46)^{44} = 0.0739$$

$$L(p = 0.56|\text{data}) = 100!/56!44! (0.56)^{56} (0.44)^{44} = 0.0801$$

$$L(p = 0.58|\text{data}) = 100!/56!44! (0.58)^{56} (0.42)^{44} = 0.0738$$

$$L(p = 0.60|\text{data}) = 100!/56!44! (0.60)^{56} (0.40)^{44} = 0.0576$$

$$L(p = 0.62|\text{data}) = 100!/56!44! (0.62)^{56} (0.38)^{44} = 0.0378$$

The MLE is 0.56 since this value of the parameter makes the observed data most likely. Notice that you can choose to evaluate as many values of the parameter as you like. The number will be determined in part by the minimum 'tolerance' that you are willing to be off by.

With more complex models and parameters, including parameters that are uncertain and have their own probability distribution, it becomes very hard to estimate the MLE. Analytical methods can be used which graph the likelihood as a line or surface and then look for the first two derivatives of the curve to look for the maxima of the curve or surface. If the number of parameters increases (increasing dimensions) the 'parameter space' rapidly enlarges. Computers are well suited to performing numerical MLE in such cases.

How does this relate to Empirical Bayes Screening? We have our initial estimated, subjective parameters for the gamma mixture and we have our likelihood that is the mixture of negative binomials representing the distribution of actual observed values. We 'pool' these distributions and determine which expression for the parameters would be most likely to produce the sample of observed negative binomial counts (determine the MLE). This maximisation involves a search in five-dimensional parameter space $\{\theta: \alpha_1, \alpha_2, \beta_1, \beta_2, P\}$ for the vector that maximises the likelihood as evidenced by the first and second derivatives of the function being zero. The likelihood is

$$L(\theta) = \Pi_{ij} \{P f(N_{ij}; \alpha_1, \beta_1, E_{ij}) + (1-P) f(N_{ij}; \alpha_2, \beta_2, E_{ij})\}$$

This involves millions of calculations. The computational procedures required for these calculations are based on the Newton-Raphson method. This is an old calculus-based technique that was devised to find the roots of an equation (e.g. the values of the independent variable (e.g. x) for which the value of the function (e.g. $f(x)$) equals zero.

To summarise, maximum likelihood estimation involves the following four steps: (i) obtain your data; (ii) specify a model; (iii) compute the likelihoods and; (iv) find the value of the parameters that maximises the likelihood.

Fig. 12. Maximum likelihood estimates.

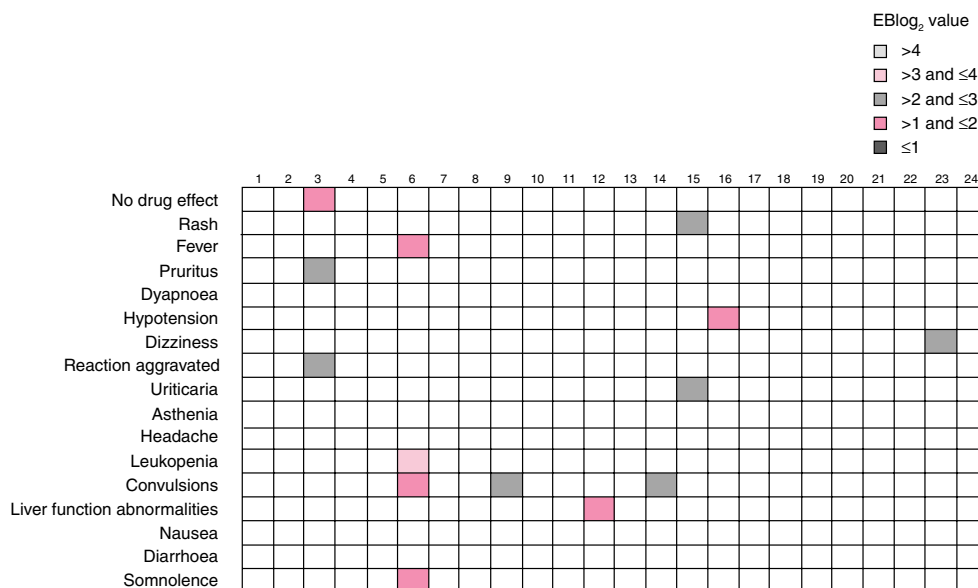


Fig. 13. Spatial maps of signal scores for the most frequently reported events (rows) and drugs (columns) in the database by intensity of the empirical Bayes signal score with the strength of the signals represented by differential shading as displayed in the key above.

William DuMouchel, which describes the theory of GPS, and an accessible help system with instructions.

7. Discussion of PRRs, BCPNN and EBS

PRRs, BCPNN and EBS are currently the most widely studied and used methods of automated signal detection. All are numerator-based and therefore 'self-contained' but the underlying models and computational intensity vary. The question naturally arises as to which is the preferred method for a given database. Many performance characteristics must be considered, including the total number of signals detected and the proportion that are meaningful.

Any institution contemplating automated signal detection needs to give serious consideration to the resource requirements for evaluating detected signals. A crucial performance parameter is therefore the overall volume of signals detected, since this will determine the resource requirements neces-

sary for the timely and effective evaluation of signals detected. Even if the rate of false alarms is not higher than with traditional signal detection methodology, the absolute number of signals and therefore the absolute number of false alarms is likely to increase. Detecting an extremely large volume of signals poses a potential problem from the perspective of triage.

Direct comparisons of the detected signal volume are somewhat imprecise, since these methods are primarily applied to different databases (PRRs/ADROIT/UK Yellow Card database, BCPNN/WHO database, and EBS/FDA Adverse Event Reporting System [AERS] database). However, published data give some indication of the relative signal volume.

As mentioned above (section 5), recently published data from the MCA cited a total of approximately 480 signals detected (10% of preferred terms) over a time interval of 2 years. However, these 481 signals were detected using only the top

15 drugs in terms of number of reports. Since the UK Yellow Card System contains over 400 000 reports describing 600 000 reactions, the number of signals detected from the entire database would be substantially higher.^[8] Without additional signal criteria PRRs may detect a higher signal vol-

ume because of the lack of a shrinkage procedure built into the model. Therefore, to reduce the volume of potential signals that are evaluated, the MCA employs triage criteria including the strength of the signal, the novelty of the event, its clinical importance (severity and seriousness) and

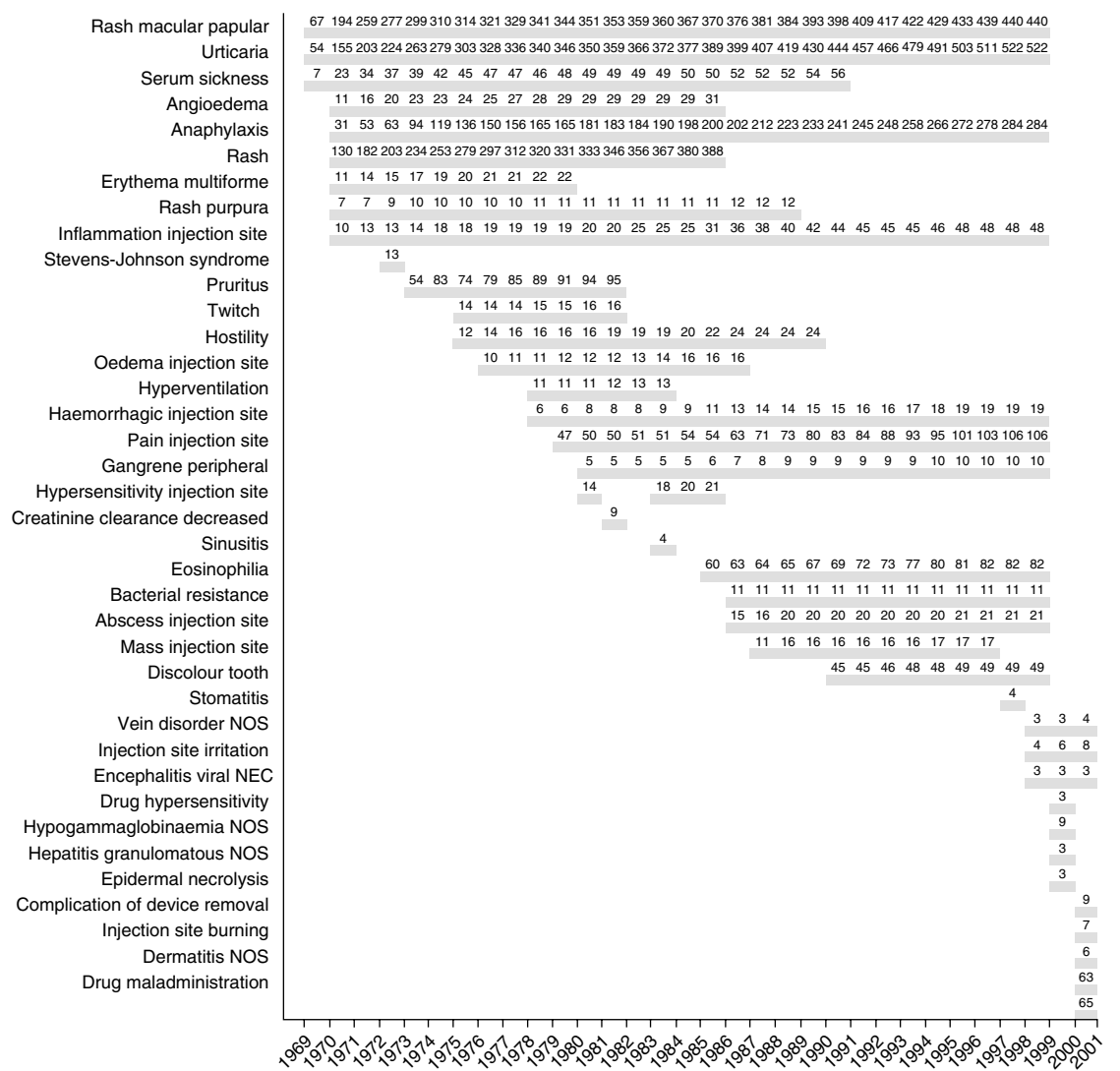


Fig. 14. Time line summary: cumulative reports and number of reports according to the year when the first signal was detected for penicillin. NEC = not elsewhere classified; NOS = not otherwise stated.

its preventability. These triage criteria are referred to by the acronym 'SNIP' (strong, new, important and potentially preventable).^[8]

The problem of signal volume and potential solutions was recently highlighted for the BCPNN. The UMC reported that the BCPNN was detecting more than 2000 potential signals per quarter. To address this, the UMC has implemented additional signal selection algorithms to be used in parallel with the BCPNN to filter down the number of drug-adverse event combinations that are triaged as potentially important. Four triage algorithms are designed to focus on serious and new events, events with a rapid reporting increase, events that are clinically of special interest because they are typically drug related (e.g. rhabdomyolysis, agranulocytosis and Stevens-Johnson syndrome) and 'international signals' that are reported from more than one country.^[30]

Data will soon be published on signal volume detected by EBS applied to the FDA Adverse Event Reporting System (1997–present) and Spontaneous Reporting System (1968–1997) databases. Scanning the total databases that cover a time interval of 34 years, approximately 3.5% of almost 1.2 million drug-adverse event combinations were flagged as potential signals equal to approximately 40 000 signals in 34 years of data. This comes out to less than 2000 signals per year of data.^[31] Unlike PRRs and BCPNN, EBS stratifies by key variables such as age, gender and time and is therefore less prone to false positive signals resulting from Simpson's paradox (confounding).

There are no published head-to-head comparisons of PRRs versus BCPNN versus EBS, but data are starting to appear comparing PRRs and EBS. PRRs are easy to understand and far less computationally intensive than BCPNN and EBS. A pharmaceutical company spontaneous reporting database containing over half a million events was used to compare PRRs and EBS.^[32] The database was scanned by both methods. The threshold criteria for PRRs are unclear, but it appears that a signal threshold consisted of a $PRR > 2$, chi-square $p <$

0.05, and a minimum of two reports of the drug-event combination. Drug-event combinations were ranked and the ranked lists of the two methods were compared. Further evaluation of each method consisted of applying logistic regression to investigating possible dose-response relationships and drug-drug interactions. The investigators found PRRs both easy to implement and accessible to drug safety physicians and has become their preferred method of signal detection. The EBS was equal to PRRs in generating large posterior estimates for known drug reactions. For events with relatively high observed frequencies, both methods generated comparable event rankings. The EBS was 'better' for ranking low frequency events but at the expense of extra programming, sophisticated software and knowledge of Bayesian methods.^[32] The criteria for deciding which ranking was better were not described in detail. It remains to be determined how the automated methods compare with certain intensive nonautomated methods, such as a drug-specific reporting frequency cut-off (e.g. a potential signal is evaluated if the event exceeds a threshold, say 2% of all reports with the drug). It would be interesting to see if all signals detected by automated methods involved events that exceeded such a threshold for each drug.

Published data on the utility of PRRs, EBS and BCPNN have been retrospective in nature. This is a critical gap, since legitimate signals may be detected that remain unconfirmed by individual case causality assessment until long after they first appeared as a signal. Moreover the retrospective nature of these investigations could have introduced bias in the conduct and/or analysis of the data mining.

In almost all instances these methods of automated signal detection have been applied to postapproval safety databases that are global in extent and contain hundreds of thousands to millions of adverse event reports. The smallest database that could be identified to which automated signal detection was studied is The Netherlands Pharmacovigilance Foundation, containing 39 790 ad-

verse event reports involving 17 330 different drug-adverse event combinations. This database was used to examine the level of concordance between various estimates of disproportionality (e.g. PRRs, RORs) with that derived from the IC in the BCPNN. Among the drug-event combinations with at least four reports there was a high level of concordance in terms of the drug-event combinations highlighted as potential signals. Among drug-adverse event combinations with fewer than four reports the other methods tended to highlight more drug-adverse event combinations. As the number of reports per drug-adverse event combination increased the concordance between the various measures was found to increase. However, the differential performance of PRRs, EBS and BCPNN in databases that differ qualitatively or quantitatively is incompletely characterised.^[33]

At present, these three methods appear comparable in performance. PRR is the easiest to understand and implement. It may not be as predictive as the Bayesian methods for low-frequency events, since the Bayesian models have a shrinkage procedure built in that lowers the signal score for drug-adverse event associations with less statistical support (low frequency of reports). It is also the only one of the three with published data on its ability to detect new signals leading to regulatory interventions. The positive predictive value, however, was quite low. The BCPNN seems to detect the largest volume of potential signals from the WHO database, which is an important resource consideration. EBS is publicly available on the Internet and may have more built-in flexibility, since relative rankings rather than absolute signal scores are emphasised. The numerous inherent confounders and limitations cannot be eliminated by any mathematical model. The need for expert clinical reviewers will probably increase because of the increased potential signal volume that these methods may detect.

8. Summary and Conclusions

Those who choose a career in pharmacovigilance know how the complex data presented to us are a source of both fascination and frustration. The fascination is the infinite number of permutations and combinations that are a rich source of hypothesis. The frustration is to try to overcome limitations in the data so that we can use our intellectual curiosity about this complex information to serve the health of the public. Although there have been previous attempts to apply statistical methodology to the analysis of pharmacovigilance data, these techniques have limited or no support in routine signal detection.

Pharmacovigilance is evolving into a more quantitative science in that quantitative statistical techniques are being applied to 'dirty' postmarketing databases traditionally considered resistant to such attacks. Although various techniques and data sources have been used, most activity currently centres on two quantitative, numerator-based methods of signal detection: PRRs and Bayesian data mining. Within Bayesian data mining are two different but related techniques: EBS and the BCPNN.

While data on the comparative performance characteristics of the various numerator-based methods are limited, PRRs and Bayesian data mining appear comparable, with EBS having a possible advantage with drug-event combinations involving very small numbers. Another potential advantage of EBS is that it does not have stand-alone threshold criteria for signalling. It uses the RR , $\log P_{ij}$ and $EBGM_{ij}$ to rank drug-event dependencies. The BCPNN provides a stand-alone threshold for signalling that applies to any cells meeting this combination regardless of ranking. The EBS may therefore provide more flexibility in establishing threshold criteria that can be an important consideration given the high number of false alarms and associated cost in resources to evaluate all these false alarms. PRRs may provide most of the power of EBS without the software, hardware and programming requirements.

Comparative performance criteria should be precise and consistent. The positive predictive value can be defined in various ways. The positive predictive value of a signal of an unlabelled event is probably more meaningful than the positive predictive value of any event detected by these methods.

Any institution contemplating automated signal detection needs to give serious consideration to the resource requirements for evaluating signals detected. Even if the rate of false alarms is not higher than traditional signal detection methodology, the absolute number of signals and therefore the absolute number of false alarms is likely to increase. Detecting an extremely large volume of signals poses a potential problem from the perspective of triage. The problem of signal volume and potential solutions was recently highlighted for the BCPNN as described in section 7.^[30]

EBS provides three different measures. Using combinations of signal scores could mitigate the problem of false positives. Other possible solutions as described in section 7 are to limit analysis to cells with a minimum cell count or adverse medical events that are characteristic of drug-related phenomena. Examining the stability or evolution of the signal over time might improve the performance characteristics.

It is crucial to remember that automated methods do not replace the need for expert clinical case review and interpretation by drug safety professionals experienced in the nuances of pharmacoepidemiology and clinical medicine. An example is the recent application of the BCPNN in which a signal was detected of possible heart-muscle disorder associated with antipsychotics, especially clozapine. The automated signalling procedure does not currently take into account numerous confounding factors, such as the fact that in many countries clozapine is used for schizophrenia unresponsive to three other antipsychotics of different chemical classes and that cardiac muscle changes with clozapine could therefore represent a cumulative effect of previous antipsychotic drug exposure.^[34]

Additional pitfalls can occur, which are related to the size and coding procedures of the database being screened. With all three methods, strong signals with one drug can obscure signals of another drug in the database. With PRRs a strong signal of an event with a given drug inflates the denominator of the PRRs with that drug, thereby lowering the signal score. Similar considerations indicate that a strong signal with a drug or class of drugs reduces the PRR for another drug or drug class.

With PRRs, BCPNN and EBS a signal may be diluted or enhanced in databases that differ quantitatively or qualitatively. A hypothetical example may clarify this concept. Consider a hypothetical database of a pharmaceutical company that does not market ACE inhibitors. Let us imagine that we apply EBS to this company database as well as to a global regulatory database that enters drugs not only from that company but from all others as well, including makers of ACE inhibitors. In EBS the expected cell counts for a given drug-adverse event combination are directly related to the proportional representation of that adverse event in the overall database. Therefore, for some events that are strongly associated with a certain class of drugs, such as ACE inhibitor-induced cough, the expected cell counts for that drug-adverse event combination may be higher in the global regulatory database containing reports with ACE inhibitors. Given that all reports from the hypothetical pharmaceutical company were submitted to the regulatory authority, reports of cough with drugs manufactured by the pharmaceutical company will be associated with a stronger signal score because the expected cell count associated with cough in the company's database is smaller. Such an effect could be defined by the temporal evolution of the signal score, since the proportional representation of cough in the regulatory database should decrease over time. However it would be difficult to predict the effect of compositional differences between databases on a given signal score since the expected cell counts and signal scores would also be a function of the overall size of the database.

The characteristics of the dictionary used can also have profound implications, since a coding variability of a given event could result in dilution of a signal. Coding with effective dictionaries with limited synonymous but separate preferred or higher-level terms would promote more effective automated screening.

As Bayesian data-mining methods are being used more widely and refined, they are being extended to the detection of higher-order multi-item associations. This will allow, for example, the detection of signals involving constellations of events or the detection of drug interactions in the form of drug-drug-adverse event associations. Recent work with the BCPNN has showed that patterns of adverse events associated with a drug that are never reported together on a single adverse event report can be inferred via pattern recognition of interdependencies.^[35] EBS has already been applied to the detection of multi-item associations in other disciplines^[36] and is currently being explored by the US Centers for Disease Control and Prevention to detect signals of vaccine-related medical syndromes in the Vaccine Adverse Event Reporting System (VAERS) database. This program is known as Multi-item Gamma Poisson Shrinker (MGPS) and is publicly available for noncommercial use on the Internet.

The mathematical content of these methods should not intimidate drug safety professionals who are not formally trained in statistics. The extra effort exerted in understanding the theoretical foundations of these methods is not only intrinsically gratifying to the intellect but will also arm the drug safety professional with the knowledge needed to make informed judgements without being overawed by the mathematics. The equations should not obscure the numerous limitations and biases in the data and the methods, including the essentially nonclinical nature of the underlying algorithms.

Collaborative interaction between all stakeholders will promote the establishment of best practices in signal detection.

Post-Publication Note

For additional information on recent developments in signal detection since the writing of this article, we highly recommend *Drug Safety* 2002; 25 (6): 379-471, that contains the proceedings of a meeting held by the Drug Safety Research Unit, Southampton, England, June 25–26, 2001.

Acknowledgements

Dedicated to the memory of my father Richard S. Hauben, MD.

We are grateful for the thoughtful insights and encouragement of Dr Ana Szarfman of the FDA and Andrew Bate of the WHO Collaborating Centre for International Drug Monitoring.

References

1. Meyboom RHB, Egberts ACG, Edwards IR, et al. Principles of signal detection in pharmacovigilance. *Drug Saf* 1997; 16 (6): 355-65
2. Waller PC, Lee EH. Responding to drug safety issues. *Pharmacoepidemiol Drug Saf* 1999; 8: 535-52
3. 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 database. *Drug Saf* 2000; 23 (6): 533-42
4. Kaufman DW, Rosenberg L, Mitchell A. Signal generation and clarification: use of case-control data. *Pharmacoepidemiol Drug Saf* 2001; 10: 197-203
5. Evans SJW. Pharmacovigilance: a science or fielding emergencies? *Stat Med* 2000; 19: 3199-209
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. Evans SJW, Waller P, Davis S. Proportional reporting ratios: the use of epidemiological methods for signal detection [abstract]. *Pharmacoepidemiol Drug Saf* 1998; 7 Suppl. 2: S102
8. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001; 10: 483-6
9. Heeley EL, Wilton LV, Shakir SAW. The use of proportional reporting ratios for signal generation in prescription event monitoring [abstract]. *Pharmacoepidemiol Drug Saf* 2002; 10 (S1): S48
10. Kilgour-Christie J, Czarniecki A, Simmons V. Proportional reporting ratio calculations from a pharmaceutical company database [abstract]. *Pharmacoepidemiol Drug Saf* 2001; 10: S159
11. 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
12. O'Neill RT, Szarfman A. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system [discussion]. *Am Stat* 1999; 53 (3): 190-5
13. Louis TA, Shen W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system [discussion]. *Am Stat* 1999; 53 (3): 196-8
14. Madigan D. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system [discussion]. *Am Stat* 1999; 53 (3): 198-200

15. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system [reply]. *Am Stat* 1999; 53 (3): 201-2
16. Praus M, Schindel F, Fescharek R, et al. Alert systems for post-marketing surveillance of adverse drug reactions. *Stat Med* 1993; 12: 2383-93
17. Shore DL, Quade D. A surveillance system based on a short memory scheme. *Stat Med* 1989; 8: 311-22
18. Norwood PK, Sampson AR. A statistical methodology for post-marketing surveillance of adverse drug reaction reports. *Stat Med* 1988; 7: 1023-30
19. Parker RA. Analysis of surveillance data with Poisson regression: a case study. *Stat Med* 1989; 8: 285-94
20. Tubert P, Begaud B, Haramburu F, et al. Spontaneous reporting: how many cases are required to trigger a warning? *Br J Clin Pharmacol* 1991; 32: 407-8
21. Schroeder DR. Detecting a rare adverse drug reaction using spontaneous reports [statistics]. *Reg Anesth Pain Med* 1998; 2: 183-9
22. Purcell P, Barty S. Statistical techniques for signal generation: the Australian experience. *Drug Saf* 2002; 25 (6): 415-21
23. Heeley E, Wilton L, Shakir A. Automated signal generation in prescription event monitoring. *Drug Saf* 2002; 25 (6): 423-32
24. Evans S. Statistical and epidemiological principles in detecting signals of adverse drug reactions. *Drug Information Association 38th Annual Meeting*; 2002 Jun 16-20; Chicago (IL)
25. Code of Federal Regulations (CFR). Parts 310, 314, and 600 [online]. Available from URL: www.access.gpo.gov/cgi-bin/cfrassemble.cgi?title=200221 [Accessed 2003 Jan 28]
26. Clark JA, Berk RH, Klinecwitz SL. Calculation of the probability of multiplicities in two cell-occupancy models: implications for spontaneous reporting systems. *Drug Inf J* 1999; 33: 1195-203
27. Begaud B, Moride Y, Tubert-Bitter P, et al. False positives in spontaneous reporting: should we worry about them? *Br J Clin Pharmacol* 1994; 38: 401-4
28. Kwong K, Sammon J, Hornbuckle K, et al. A novel method for detecting potential drug-drug interactions using post-marketing adverse event database [abstract]. *Pharmacoepidemiol Drug Saf* 2001; 10 (S1): S39
29. O'Neill RT, Szarfman A. Some US Food and Drug Administration perspectives on data mining for pediatric safety assessment. *Curr Ther Res Clin Exp* 2001; 62: 650-63
30. A new strategy for detection of signals in the WHO database 2002. *Uppsala Reports* 2001 Nov; 17: 8-9
31. Szarfman A, Machado GS, O'Neill RT. Screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in FDA's spontaneous reports database. *Drug Saf* 2002; 25 (6): 381-92
32. Thakrar BT, Blesch KS, Sacks ST, et al. On technical methods of signal detection [abstract]. *Pharmacoepidemiol Drug Saf* 2001; 10 (S1): S98
33. Van Puijtenbroek E, Bate A, Leufkens H, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol Drug Saf* 2002; 11: 3-10
34. Coulter DM, Bate A, Meyboom R, et al. Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study. *BMJ* 2001; 232: 1207-9
35. Bate A, Lindquist M, Edwards IR. Pattern recognition using a recurrent neural network and its application to the WHO database [abstract]. 17th International Society for Pharmacoeconomics; 2001 Aug 23-26; Toronto, Canada
36. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. *Proc. KDD. San Diego (CA): ACM Press*, 2001

Correspondence and offprints: Dr M. Hauben, Safety Evaluation and Epidemiology, Pfizer Inc., 235 East 42nd Street, New York, NY10021, USA.
E-mail: manfred.hauben@Pfizer.com