BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 23 2012, pages 3123–3130
doi:10.1093/bioinformatics/bts576

Data and text mining

Advance Access publication October 12, 2012

Multinomial modeling and an evaluation of common data-mining algorithms for identifying signals of disproportionate reporting in pharmacovigilance databases

Kjell Johnson[1,*], Cen Guo[2], Mark Gosink[3,*], Vicky Wang[4] and Manfred Hauben[5]

[1]Arbor Analytics LLC, [2]Department of Statistics, University of Michigan, Ann Arbor, MI, USA, [3]Pfizer Global Research and Development, Groton, CT, USA, [4]Pfizer Global Research and Development, Cambridge, MA, USA and [5]Pfizer Worldwide Safety Strategy, New York, NY, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: A principal objective of pharmacovigilance is to detect adverse drug reactions that are unknown or novel in terms of their clinical severity or frequency. One method is through inspection of spontaneous reporting system databases, which consist of millions of reports of patients experiencing adverse effects while taking one or more drugs. For such large databases, there is an increasing need for quantitative and automated screening tools to assist drug safety professionals in identifying drug–event combinations (DECs) worthy of further investigation. Existing algorithms can effectively identify problematic DECs when the frequencies are high. However these algorithms perform differently for low-frequency DECs.

Results: In this work, we provide a method based on the multinomial distribution that identifies signals of disproportionate reporting, especially for low-frequency combinations. In addition, we comprehensively compare the performance of commonly used algorithms with the new approach. Simulation results demonstrate the advantages of the proposed method, and analysis of the Adverse Event Reporting System data shows that the proposed method can help detect interesting signals. Furthermore, we suggest that these methods be used to identify DECs that occur significantly less frequently than expected, thus identifying potential alternative indications for these drugs. We provide an empirical example that demonstrates the importance of exploring underexpected DECs.

Availability: Code to implement the proposed method is available in R on request from the corresponding authors.

Contact: kjell@arboranalytics.com or Mark.M.Gosink@Pfizer.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 20, 2012; revised on September 7, 2012; accepted on September 18, 2012

1 INTRODUCTION

Over the past 50 years, pharmacovigilance has evolved from a small limited-scale data collection and evaluation process involving scientific rationale and debate (Finney, 2003) to a large world-wide systematic data collection process, which now additionally includes extensive statistical analysis (Bate and Evans, 2009). As data collection and statistical evaluation of adverse events have become more systematic, the ultimate goal has remained the same: to identify, as soon as possible, drug–adverse event combinations (DECs) that pose a significant risk to the population (Hauben *et al.*, 2005; Hauben and Reich, 2005).

Although there has been a manyfold improvement in collecting spontaneous reports, the penultimate goal is extremely difficult to achieve. For one, spontaneous reporting systems (SRSs) are far from perfect and contain both error and bias. Known sources of error include incorrect association between drugs and events, over-reporting (multiple reports for the same incident) and under-reporting (events that are never reported) (DuMouchel, 1999; Hauben *et al.*, 2007). In addition, SRS databases are subject to reporting bias [over-reporting of DECs on publicity of a suspected association (Hauben *et al.*, 2005)]. Furthermore, SRS databases lack exposure information, thus implying that the event reporting rates derived from databases can only be considered as relative.

Yet, despite the noise and bias present in SRS databases, it is still possible to use this data to identify many unusually large reporting frequencies, which are also known as signals of disproportionate reporting (SDRs) (Hauben *et al.*, 2004). However, it is extremely important to realize that these SDRs, at best, indicate associations that are potentially causal and that should be scientifically evaluated. Moreover, an SDR represents a numerical output devoid of clinical context and is not equivalent to a signal of suspected causality (Hauben and Aronson, 2009). In addition, SRS data are more akin to a census, rather than an unbiased sample from an underlying true population of adverse event reports. Therefore 'estimates' and corresponding 'confidence intervals' of SDRs stemming from any data-mining algorithm (DMA) should really be viewed as 'pseudo-estimates' and 'pseudo-intervals'.

Many statistical approaches, also known as DMAs, have been applied and developed to find SDRs, especially in the less-than-desirable conditions that SRS databases present. DMAs fall into two categories: traditional (or frequentist) methods and Bayesian methods (Hauben and Zhou, 2003). Each of these approaches provides distinct advantages and inherent disadvantages (Hauben and Reich, 2005). The primary advantages of frequentist approaches are that they are simple to compute, are easy to interpret and have higher sensitivity than current Bayesian methods when comparing common implementations. However, the underlying model assumptions fail for low-count drug–adverse

*To whom correspondence should be addressed.

event combinations, which represent a majority of DECs in SRS databases. Under these conditions, frequentist signal detection methods can become unstable (i.e. the increased detection of signals is accompanied by an increased detection of noise) and unreliable. Bayesian methods, in contrast, attempt to stabilize the resulting ratio metrics for low-count DECs via shrinkage. However, these approaches have been shown to be less sensitive for detecting SDRs for low-count combinations, thus implying that they can overshrink (Hauben and Reich, 2005). Bayesian approaches are also less intuitive and more computationally intensive than frequentist approaches.

Regardless of approach, no one method has been shown to be superior to others at identifying unusual DECs (Bate and Evans, 2009; Hauben *et al.*, 2005), and the lack of both uniformly accepted gold standards of causality and a calculus of costs and utilities associated with correct and incorrect classifications in pharmacovigilance complicates head-to-head comparisons. This should not be surprising, given the inherent messiness of the data. Furthermore, it has been shown that frequentist and Bayesian approaches produce similar results for higher-count DECs (Bate and Evans, 2009). Finally, neither frequentist nor Bayesian approaches optimally handle low-count DECs, which is where pharmacovigilance desires to first identify signal.

In this work, we introduce a method based on a multinomial model for estimating the degree of interaction between a drug and an adverse event. The resulting score is standardized using a nonparametric approach, which avoids the asymptotic pitfalls that frequentist parametric methods must assume in low-count cases. Furthermore, we compare the multinomial approach and common DMAs on the metrics of shrinkage and scoring of DECs. These comprehensive results can enable the practitioner to better assess the relevance of DECs. Last, we show that these methods can be used to identify DECs that occur significantly less frequently than expected. Although traditional pharmacovigilance ignores this direction of the test, we advocate that this direction should not be ignored because these results may suggest potential alternative indications of medicines. Specifically, large negative scores could imply that a drug may be beneficial for a specific event. This possibility is accommodated by a recently proposed definition of signal (Hauben and Aronson, 2009).

Our work is organized as follows: in Section 2, we review several DMAs and explain the caveats of each. In addition, we propose an empirical approach based on a multinomial model and illustrate how this approach avoids the pitfalls of frequentist approaches. In Section 3, we compare each method with the proposed method on a subset of the Adverse Event Reporting System (AERS) data. Then in Section 3.2, we show how these signal detection methods can be used to identify potential alternative indications for individual drugs or classes of drugs. Finally, we summarize and discuss these results in Section 4.

2 METHODS

Many authors have used one or more DMAs to identify signal in SRS databases (Evans *et al.*, 2001; Hauben, 2004; Hochberg *et al.*, 2007), whereas others have contrasted performance of DMAs at identifying known signal (Bate and Evans, 2009; Hauben and Reich, 2005; Hauben *et al.*, 2007; Hochberg *et al.*, 2009). van Puijenbroek *et al.* (2002) provided an extensive comparison of the performance of several DMAs on identifying signal and concluded that common DMAs perform similarly at identifying DEC signal when there are at least four reports for the combination. In this section, we will further explore why common DMAs perform similarly when there are sufficient numbers of reports. However, this analysis did not include one of the commonly used algorithms, the Multi-item Gamma Poisson Shrinker (GPS). In a later section, we will extend on the results presented by van Puijenbroek *et al.* (2002), including the results from our proposed approach.

2.1 Existing signal detection methods

To begin, we will outline the general problem and corresponding notation, and we will define common DMAs and provide references for more in-depth information. First, given a specific drug and adverse effect combination, let a represent the number of reports for the drug and adverse event combination, let b be the number of reports for individuals taking the drug but having other adverse events, let c be the number of reports for individuals with the adverse event but not taking the drug and let d be the number of reports excluding the drug and adverse event. Then the particular drug–adverse event combination can be represented by the following 2×2 table:

		Adverse event	
		Yes	No
Drug	Yes	a	b
	No	c	d

Existing algorithms for detecting signal are based on the association between drugs and events of the above 2×2 table. Common frequentist methods include the reporting odds ratio (ROR) and the proportional reporting ratio (PRR) (Evans *et al.*, 2001), and are often viewed on the log scale

$$\log(\mathrm{ROR}) = \log\frac{ad}{bc}$$

$$\log(\mathrm{PRR}) = \log\frac{a(c+d)}{c(a+b)}$$

Although these two methods are easy to use and interpret, there are some practical constraints that occur for many DECs. Specifically, for ROR to be defined, b and c must be greater than zero; similarly, for PRR to be defined, $a+b$ and c must be greater than zero. To determine the significance of the magnitude of any of the above measures of association, each signal should be evaluated relative to its standard error. The corresponding asymptotic estimates of variability for these methods are

$$SE(\log(\mathrm{ROR})) \approx \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$SE(\log(\mathrm{PRR})) \approx \sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}$$

and depend on sufficient cell counts to be reliable (Stokes *et al.*, 2000). It is easy to see that a, b, c and d must be greater than zero for $SE(\log(\mathrm{ROR}))$ to be defined; a and c must be greater than zero for $SE(\log(\mathrm{PRR}))$ to be defined.

Bayesian methods do not suffer from these kinds of constraints and include the GPS and Bayesian Confidence Propagation Neural Network (BCPNN) (Bate *et al.*, 1998), and are based on estimating the information criterion

$$\mathrm{IC} = \log\frac{a(a+b+c+d)}{(a+b)(a+c)}$$

The posterior expectation of the IC via BCPNN is

$$E(\mathrm{IC}) = \log\frac{(a+1)(n+2)^2}{(n+\delta)(a+b+1)(a+c+1)}$$

$$\text{where } \delta = \frac{n+2}{a+b+1}\cdot\frac{n+2}{a+c+1},$$

and the corresponding estimate of variability is

$$SE(\mathrm{IC}) \propto \sqrt{\frac{\frac{n-a+\delta-1}{(a+1)(1+n+\delta)}}{+\frac{n-(a+b)+1}{(a+b+1)(n+3)}} + \frac{n-(a+c)+1}{(a+c+1)(n+3)}}$$

One can see from the above formulas that E(IC) and SE(IC) are defined when a, b, c or d are zero, which is a distinct advantage of this method, especially when trying to detect signal in sparse combinations.

When log(ROR) and log(PRR) are defined, we can use the above formulas to compare the estimates of variability and their corresponding estimates of variability under common scenarios. For a majority of DECs in SRS databases, $a \ll b, c \ll d \approx n$. Under this scenario, $a+b \approx b$, $a+c \approx c$ and $n+x \approx n$, where x is a small integer. Applying these approximations, one can show the following:

$$\log(\mathrm{ROR}) \approx \log(\mathrm{PRR}) \approx \log\frac{a}{\frac{(a+b)(a+c)}{n}} = \log\frac{a}{E(a)} \tag{1}$$

and

$$\begin{aligned} E(\mathrm{IC}) &= \log\frac{(a+1)(n+2)^2}{n(a+b+1)(a+c+1)+(n+2)^2} \\ &\approx \log\frac{a+1}{\frac{(a+b)(a+c)}{n}+1} \\ &= \log\frac{a+1}{E(a)+1} \end{aligned} \tag{2}$$

Although (1) and (2) appear to be similar in value, they are not when $a \ll b, c \ll d$. In fact, as $E(a)$ decreases, E(IC) decreases more rapidly than log(ROR) or log(PRR) owing to the additional 1 in the denominator of (2). When a is small, $E(a)$ is usually much smaller than 1, which produces large values of log(ROR) and log(PRR) and a much smaller value of E(IC). Therefore, the BCPNN method is less sensitive to pharmacovigilance signal for low-count DECs, which often occurs for more recently approved drugs or for drugs that have low exposure to the population—two key scenarios where pharmacovigilance desires to identify signal.

When $a \ll b, c \ll d \approx n$ and using the above approximations, one can show that the estimates of variability are

$$SE(\log(\mathrm{ROR})) \approx SE(\log(\mathrm{PRR})) \approx SE(\mathrm{IC}) \approx \sqrt{\frac{1}{a}+\epsilon}$$

where ϵ is on the order of $\max\left(b^{-1}, c^{-1}\right)$. Hence, the estimates of variability are similar under this common DEC scenario and depend on the magnitude of a. In a small number of cases, $a>b$, and b is small (possibly close to 0). Under this condition,

$$SE(\log(\mathrm{ROR})) \approx SE(\mathrm{IC}) \approx \sqrt{\frac{1}{a}+\frac{1}{b}+\eta}$$

where η is on the order of c^{-1}. In the small number of cases when $a>b$, the standard errors depend more strongly on the magnitude of b. Alternatively, when $a>b$,

$$SE(\log(\mathrm{PRR})) \approx \sqrt{\frac{1}{\frac{a^2}{b}+a}+\eta}$$

The error for PRR thus depends on the magnitude of a^2. Because a is greater than b, SE(log(PRR)) will be smaller than SE(log(ROR)) and SE(IC). This implies that the standardized PRR will be large when $a>b$ and b approaches zero.

The mathematical setup of ROR, PRR and BCPNN make it easy to identify scenarios that cause these methods to yield different signal and noise values. GPS, in contrast, does not nicely fit into a form that allows for the same kind of succinct mathematical comparison. Below, we provide the forms of the signal and error for GPS, which we will then use to make an empirical comparison across these methods in Section 3.

The GPS method assumes a Poisson distribution for the observed counts for the i-th drug and the j-th adverse event, $a_{ij} \sim \mathrm{Poisson}(\lambda_{ij} * \mathbb{E}_{ij})$, where $\mathbb{E}_{ij} = \frac{d_{ij}}{b_{ij} * c_{ij}}$ is the expected counts for drug i and event j if there is no association. The parameter λ_{ij} characterizes the association between drug i and adverse event j. λ_{ij} larger than 1 means there is positive association, and λ_{ij} smaller than 1 means there is negative association. The GPS method takes a Bayesian framework, which models the parameter λ_{ij} as a random variable, with prior distribution $\pi(\lambda_{ij})$ as a mixed gamma distribution

$$\begin{aligned} \mathbf{P}_{\mathrm{prior}}(\lambda_{ij}) &= \pi(\lambda_{ij}|\alpha_1,\beta_1,\alpha_2,\beta_2,Q) \\ &= Q * \mathrm{Ga}(\lambda_{ij}|\alpha_1,\beta_1) + (1-Q)\mathrm{Ga}(\lambda_{ij}|\alpha_2,\beta_2) \end{aligned}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where $\mathrm{Ga}(\lambda|\alpha,\beta) = \frac{\beta^\alpha \lambda^{\alpha-1}}{e^{\beta\lambda}\Gamma(\alpha)}$ is the gamma distribution with parameter α and β. Given the distribution of the observed counts and the prior distribution of the parameter, we can compute the posterior distribution for the parameter λ_{ij}

$$\begin{aligned} &\mathbf{P}_{\mathrm{posterior}}(\lambda_{ij}) \\ &= \pi(\lambda_{ij}|\alpha_1+a_{ij},\beta_1+\mathbb{E}_{ij},\alpha_2+a_{ij},\beta_2+\mathbb{E}_{ij},\tilde{Q}) \\ &= \tilde{Q} * \mathrm{Ga}(\lambda_{ij}|\alpha_1+a_{ij},\beta_1+\mathbb{E}_{ij}) \\ &+ (1-\tilde{Q})\mathrm{Ga}(\lambda_{ij}|\alpha_2+a_{ij},\beta_2\mathbb{E}_{ij}) \end{aligned}$$

where $\tilde{Q}$ is the posterior probability that the parameter λ_{ij} follows the first gamma distribution. We have

$$\tilde{Q} = \frac{Q * \mathrm{NB}(a_{ij}|\alpha_1, 1/(1+\beta_1/\mathbb{E}_{ij}))}{Q * \mathrm{NB}(a_{ij}|\alpha_1, 1/(1+\beta_1/\mathbb{E}_{ij})) + (1-Q)\mathrm{NB}(a_{ij}|\alpha_2, 1/(1+\beta_2/\mathbb{E}_{ij}))}$$

where $\mathrm{NB}(a|\alpha,p)$ is the negative binomial distribution $\mathrm{NB}(a|\alpha,p) = C_n^{\alpha+n-1}p^n(1-p)^{\alpha-1}$

Then the GPS score is the mean value of λ_{ij} divided by the standard error of λ_{ij}:

$$E(\lambda_{ij}) = E_{\mathbf{P}_{\mathrm{posterior}}}(\lambda_{ij})$$

$$SE(\lambda_{ij}) = SE_{\mathbf{P}_{\mathrm{posterior}}}(\lambda_{ij})$$

To summarize, although ROR and PRR are easy to compute and interpret, their usefulness is somewhat limited. Specifically, cell counts must be sufficiently large for these estimates of independence and corresponding variability to be stable for rare DECs. BCPNN does not suffer from these constraints, but shrinks the expected IC more harshly as the expected value of a decreases. This means that BCPNN will not be able to detect signal for drugs that have low exposure to the population. In the following section, we propose an empirical nonparametric method for estimating variability that can yield more stable results than the above historical methods. Furthermore, it retains the advantages of each of the above methods while avoiding the disadvantages.

2.2 Multinomial modeling of SRS databases

For notational simplicity, let

$$P_{i,j} = \frac{a}{n},\ P_{i,+} = \frac{a+b}{n} \text{and } P_{+,j} = \frac{a+c}{n}$$

Then the probability of the ij^{th} DEC can be represented using the saturated model:

$$\log P_{i,j} = \alpha_i + \beta_j + \gamma_{i,j} \tag{3}$$

where $\alpha_i = \log P_{i,+}$, $\beta_j = \log P_{+,j}$, and $\gamma_{i,j} = \log \frac{P_{i,j}}{P_{i,+} * P_{+,j}} = IC$. In (3), $\gamma_{i,j}$ represents the effect not accounted for by the i-th drug or j-th event. This residual information can also be viewed as the effect due to the interaction between the i-th drug and j-th event. If the ij-th DEC does not interact, then $\gamma_{i,j}$=0; however, if $\gamma_{i,j} \neq 0$, the combination of drug i and event j has a synergistic or antagonistic effect on the probability of the ij-th DEC.

To understand the significance of the magnitude of $\gamma_{i,j}$, we must first understand the variation of this measure. Assume that drugs (rows) and adverse events (columns) are independent and that the counts from an $I \times J$ table follow a multinomial distribution:

$$N_{1,1}, \ldots, N_{i,j}, \ldots, N_{I,J} \sim \text{Multinomial}(P_{1,1}, \ldots, P_{i,j}, \ldots, P_{I,J}, N)$$

whare $P_{i,j} = P_{i,+} * P_{+,j}$

Then it can be shown that given the marginal probability $P_{i,+}$ and $P_{+,j}$, the standard error of $\gamma_{i,j}$ is approximately

$$\text{SE}(\gamma_{i,j} | P_{i,+}, P_{+,j}) \approx \sqrt{\frac{(1 - P_{i,+}) * (1 - P_{+,j})}{N * P_{i,+} * P_{+,j}}} \quad (4)$$

(see Supplementary Material for derivation), and increases as the marginal probabilities of either drug or event decrease, which usually occurs for extremely low-count DECs. Although this approximation works well when all expected cell counts are large, it overestimates the true variation when expected cell counts become small because it is conditioned on the expected counts. Specifically, as $P_{i,+}$ and/or $P_{+,j}$ become small, the standard error increases rapidly.

A better estimate of variation when cell counts are small can be obtained from a simulated independent table by conditioning on observed counts. Let $\{N^{Ind}_{1,1}, \ldots, N^{Ind}_{i,j}, \ldots, N^{Ind}_{I,J}\}$ be the counts from an independent multinomial distribution.

$$\{N^{Ind}_{1,1}, \ldots, N^{Ind}_{i,j}, \ldots, N^{Ind}_{I,J}\} \sim \text{Multinomial}(P_{1,1}, \ldots, P_{i,j}, \ldots, P_{I,J}, N)$$

whare $P_{i,j} = P_{i,+} * P_{+,j}$

The saturated model (3) is then applied to the simulated independent table, and $\gamma_{i,j}$ is computed for all i and j. The $\gamma_{i,j}$s estimated from this simulated data provide a null distribution or reference range under the assumption of no interaction between drug and event, and can be used to better assess relative magnitude of observed $\gamma_{i,j}$s.

Now consider estimating γ_{ij} variation by conditioning on the observed counts from the simulated independent table as follows:

$$\text{SE}_{emp}(\gamma_{i,j} | N_{i,j}) = \text{SE}(\{\gamma^{Ind}_{s,t} : N^{Ind}_{s,t} = N_{i,j}\})$$

This approach avoids using the marginal distribution and instead calculates the standard error based on the observed counts, thus making the standard error estimate more stable, especially in very sparse datasets. As a result of this approach, DECs having the same observed counts will have the same standard error estimate. Figure 1 compares the theoretical and empirical standard errors for a real dataset (see Section 3 for description). For DECs with count greater than five, the standard errors are nearly identical. However, for combinations with five or fewer occurrences, the empirical standard error is less than the theoretical standard error and is a better reflection of the true variation.

2.3 Detecting signal

For any drug and adverse event, consider estimating the relative significance of γ_{ij} by standardizing γ_{ij} with its standard error

$$\gamma^*_{i,j} = \frac{\gamma_{i,j}}{\text{SE}_{emp}(\gamma_{i,j})} \quad (5)$$

The threshold of significance for standardized gamma (SG) is chosen based on the null distribution of $\gamma^*_{i,j}$ estimated from an independent table. Let T_α be the threshold for SG,

$$T_\alpha = \text{quantile}(\{\gamma^{sd,Ind}_{1,1}, \ldots, \gamma^{sd,Ind}_{i,j}, \ldots, \gamma^{sd,Ind}_{I,J}\}, \alpha) \quad (6)$$

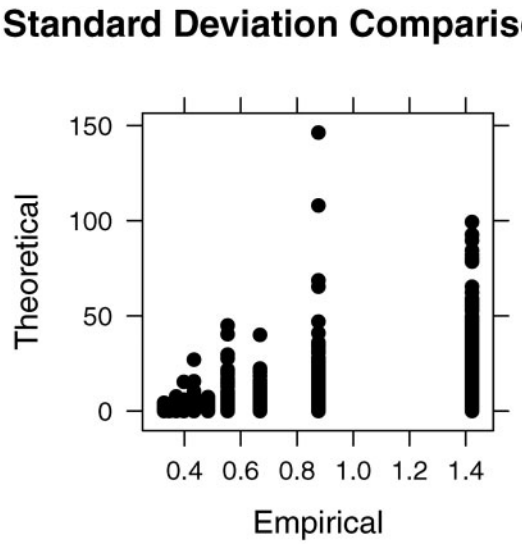

Fig. 1. Scatter plot of the empirical standard deviation versus theoretical deviation from a simulated independent table. From right to left are the standard error values corresponding to the observed counts from 1 to 10 (e.g. the dots on the far right column are for DECs that have $a = 1$). This figure illustrates that for small counts, the theoretical standard errors are larger than the empirical standard errors

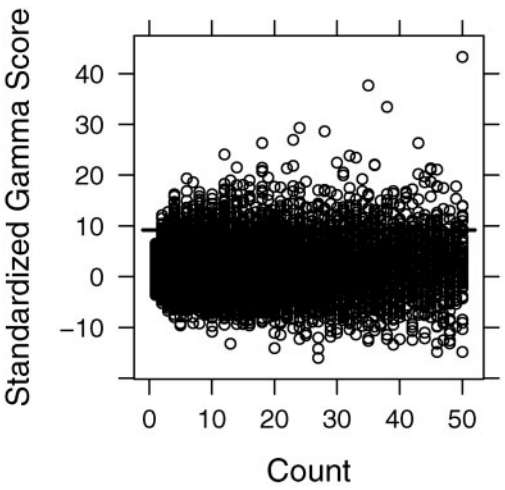

Fig. 2. Scatter plot of the observed counts versus SG using the empirical standard deviation. The horizontal line is the top 1% threshold

where α is the user-specified level, and $quantile(C, \alpha)$ is the α-th quantile of the vector C. This means under independence that $1 - \alpha$ percentage of all the drug–adverse event combinations will exceed the threshold T_α. Figure 2 highlights the top 1% of DECs for a real dataset (see Section 3 for description).

3 RESULTS

3.1 Method comparison

To better understand how each method performs, standardized PRR, BCPNN, GPS and SG are applied to a real dataset

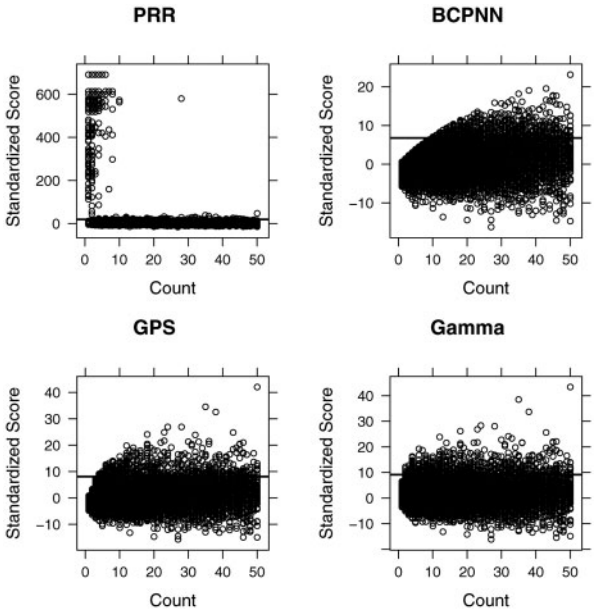

Fig. 3. Scatter plots of standardized PRR, BCPNN, GPS and SG scores for the observed counts between 1 and 50. The horizontal line in each plot is the 99th percentile of the scores corresponding to all the counts between 1 and 50

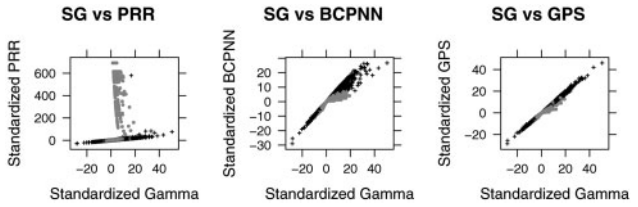

Fig. 4. Scatter plots of SG versus standardized PRR, BCPNN and GPS scores for counts between 1 and 100. Grey dots represent counts between 1 and 10 and black pluses represent counts between 11 and 100

extracted from the Food and Drug Administration's AERS database. We used a commercially available version from Oracle Health Sciences in which the duplicated records of reports are removed and the drug names are standardized. This enables us to calculate accurate counts of reports per DEC. To form the dataset, we extracted all reports with adverse events in the cardiac preferred term category from 1968 to 2009Q2, as cardiac is one of the major categories in drug-induced adverse events. We did not use age, gender or year of report as stratification variables for any of the methods, which is common practice for pharmacovigilance problems. Instead, the comparisons within this text are based on the aggregated data to enable a direct scoring comparison among methods.

Computing was done in R, and the PhViD package was used to compute BCPNN and GPS scores. The BCPNN and GPS implementations in PhViD are close representations of the commercially available versions of these algorithms. Code to compute SG scores is available on request from the corresponding author.

Consider Figure 3, which illustrates the standardized PRR, BCPNN, GPS and SG values for the DECs with counts less than 50. This figure confirms the characteristics noted in Section 2. Specifically, for the DECs with small counts ($a \leq 10$), the standardized PRR method tends to generate a number of large scores, whereas the BCPNN method tends to more strongly shrink scores. In the case where b is 0, the PRR standard errors are small (on the order of $c^{-1/2}$), which inflates the standardized score for PRR (>100). In contrast, when a is small, BCPNN method shrinks the scores toward the prior information that is 0. If we use the 99th quantile as the threshold, many DECs with counts less than 10 will be detected as significant by the standardized PRR method only because those drugs do not have much exposure (e.g. b is small). Any DECs with counts less than 10 will not be detected as significant by the BCPNN method only because they do not have enough observed counts ($a \leq 10$). For GPS and SG, the distribution of the scores is more homogeneous across all observed counts. This enables a universal threshold for identifying interesting DECs regardless of observed counts.

Figure 4 further highlights the contrasts among methods for low-count DECs. Standardized PRR inflates the scores when b equals 0 and has good agreement with SG otherwise. BCPNN shrinks scores toward 0 as a approaches 0, but has better agreement with SG as a increases. SG and GPS have most agreement, but differ slightly for counts between 1 and 4.

We further compared GPS and SG using concordance correlation and found a coefficient of 0.98, indicating that SG should be comparable with one widely used algorithm, GPS, for signal detection in pharmacovigilance. We also looked at the signal identification capabilities of SG in a specific adverse event example of drug-induced immune thrombocytopenia (DITP). DITP can be a serious adverse risk, which has been associated with a number of drugs (Al-Nouri and George, 2012; Reese *et al.*, 2010). To evaluate DITP, SG scores were generated for all DECs at the higher-level term (HLT) level. Count data for all individual reports were extracted from the cumulative AERS 2011Q2 release of Oracle Health Sciences Empirica Signal database. To generate the HLT counts, individual reports mapping to multiple terms were collapsed to single counts at the corresponding HLT-level term. HLT drug–event counts were used to generate SG scores as described above. Using case report data from Reese *et al.*, which provides drug-specific level-of-evidence categories in the Supplementary Table S1 (Reese *et al.*, 2010), 70 drugs with published case reports for thrombocytopenia were identified and an additional 1164 drugs with no case reports or where the case report clinical data did not support evidence for a causal association with thrombocytopenia were also identified. The average SG score for drugs with positive and negative case reports were compared. The average SG score for the positive drugs was 10.87 and for the negative drugs was -0.76, with a Student's t-test P-value of 0.013. Non-parametric analyses were also significant.

3.2 Negative gamma

Traditional pharmacovigilance approaches, as described above, desire to detect adverse events that are a detriment to the population. In addition, these methods can also be used to detect DECs that have a significant negative association (e.g. events

Table 1. Mood-related standardized gamma and cumulative probability for aspirin and diclofenac

Adverse effect	Aspirin	Diclofenac
Anxiety symptoms	−14.48 (2.17E-02)	−12.50 (4.42E-02)
Depressive disorders	−13.89 (1.06E-02)	−10.50 (4.93E-02)
Emotional and mood disturbances NEC	−9.54 (7.79E-02)	−8.06 (1.23E-1)
Mood alterations with depressive symptoms	−6.64 (9.33E-02)	−5.26 (1.60E-1)
Sleep phase rhythm disturbances	−3.49 (6.02E-04)	−3.11 (1.49E-3)

Table 2. Leukemia-related average standardized gamma and statistical significance for dopaminergic compounds

Adverse effect	Average dopaminergic	Average non-dopaminergic	*t*-test
Leukemias NEC	−0.856	0.130	5.17E-06
Leukemias acute NEC	−0.303	0.680	6.72E-05
Leukemias acute lymphocytic	−0.457	0.487	2.42E-05
Leukemias acute myeloid	−1.903	−0.509	5.03E-09
Leukemias chronic NEC	0.955	1.766	9.91E-04
Leukemias chronic T cell	0.774	1.614	6.80E-04
Leukemias chronic lymphocytic	−0.254	0.521	1.09E-04

that occur much less frequently than expected). This information can be used to hypothesize about possible alternative indications for existing single medicines or to identify classes of drugs that have potentially beneficial effects.

To evaluate drugs for potential alternative indications, the above SG scores were used; in addition, *Z*-scores were generated for all drugs in a given event. *Z*-scores were then converted to cumulative probability *P*-values. An examination of the resulting SG scores revealed that two non-steroidal anti-inflammatory drugs have very low scores for mood-related events. For example, aspirin has an SG score of −14.48 for 'Anxiety symptoms' (Table 1). SG scores for aspirin were also very low for a number of other mood-related events. Another Cox-2 inhibitor, diclofenac, had similarly low SGs for these events.

The potential value of Cox-2 inhibitors in the treatment of mood disorders has been evaluated by a number of groups (Ketterer *et al.*, 1996; Lieberman *et al.*, 1987; Mller, 2010). Aspirin inhibits Cox-2, which in turn participates in the metabolism of prostaglandins and arachidonic acids (Santovito *et al.*, 2009). Arachidonic acid is known to play an important part in nerve signal transmission, and alterations in this pathway are thought to be involved in bipolar disorder (Duncan and Bazinet, 2010; Kim *et al.*, 2009). Individuals suffering from bipolar disorder have severe mood swings and often need to take lithium to stabilize their mood (Machado-Vieira *et al.*, 2009). Stolk *et al.* (2010) have reported that aspirin coadministered with lithium can result in a statistically significant reduction in frequency of emotional incidents. Other workers have found that aspirin can shorten the time to onset of effectiveness of antidepressant compounds (Brunello *et al.*, 2006; Mendlewicz *et al.*, 2006). Together, these results demonstrate that aspirin can have dramatic mood effects in some situations.

Recently, dopaminergic compounds have resurfaced as potential therapeutics in the treatment of leukemia (Sachlos *et al.*, 2012; Wick, 1981). Although the mechanism behind this effect has not been worked out, it may be related to dopamine's known modifying effects on prolactin production, which, in turn, has been shown to be linked to leukemia (Braesch-Andersen *et al.*, 1992). To determine whether negative gamma scores could also be used to identify drug-class alternative indications, a set of dopaminergic agents was identified from MESH using the query 'Dopamine Agents [Pharmacological Action]', and drug names with exact matches to the identified compounds were flagged. Of 3056 compounds, 55 were classified as dopaminergic by this method. The average SD scores for the leukemia-related events were calculated and compared for all dopaminergic compounds versus all remaining compounds. For every leukemia event at the HLT level, the average SD score for the dopaminergic agents was statistically lower than that for the remaining compounds. Statistical significance was measured by the Student's *t*-test (Table 2). The results of this analysis demonstrate that negative gammas can be used to identify potential new drug-class indications. We have also used this approach to reconfirm the widely reported melanoma-protective effects of some non-steroidal anti-inflammatory drugs (data not shown).

4 DISCUSSION AND CONCLUSION

The use of the SG score based on a multinomial model and the empirical non-parametric calculation of variance produced results that were similar to those of well-accepted Bayesian shrinkage-based methods, with some potentially important differences. Namely, the magnitude of the shrinkage was less than GPS for cardiovascular DECs with one to four reports in an authentic dataset. As the goal of pharmacovigilance is the detection of novel adverse events in the most expeditious manner, with minimal patient exposure, the observed differences at low reporting frequencies could be significant, especially as some shrinkage-based methods such as GPS may miss signals absolutely or relatively in terms of timing—in other words, some credible signals may be shrunk along with noise. SG scores may therefore provide an option for pharmacovigilance organizations to obtain similar results with a methodology that is arguably more transparent and easily understood by the broader range of drug safety specialists.

Although the historical pharmacovigilance use of SRS databases has focused on finding DECs that occur unusually frequently, this work showcases an alternative non-traditional use of these kinds of databases. As illustrated with the HLT subset of events, the SG approach can identify signals that occur much less frequently than expected. Although this kind of information does not replace results that could be obtained from a controlled clinical study, it does provide a rich source of hypotheses for organizations seeking to explore alternative treatments via currently approved medicines.

Although the mathematical comparisons and contrasts presented in Section 2 are true regardless of application, the research presented in this work does have a few limitations from the pharmacovigilance perspective, and opportunities for future research exist. First, we did not use a gold standard database containing accepted drug–event signal and adverse event terminology. Therefore, we cannot use the results above to make a declarative statement about the superiority of any method at finding this kind of signal. Second, the ability to stratify the analysis by age, gender and year of report is common for standard real-world pharmacovigilance work. The SG approach as presented above could be used in this context by applying the method to stratified versions of the data. Alternatively, a more complex model could be derived accounting for these factors. A third limitation is that our method comparison analysis was limited to the subset of cardiovascular events. In real-world pharmacovigilance, global database screening is performed on the entire database, possibly minus certain sources of reports that are considered not truly spontaneous (e.g. study report). In addition, we used an early implementation of the BCPNN, and the results we obtained may therefore not apply to current implementations that have, in effect, been better calibrated for low-frequency DECs (Hauben and Bate, 2009; Noren *et al.*, 2006). Furthermore, we did not take into account the impact of the adverse event terminology structure used to memorialize adverse events in spontaneous reporting databases (Hauben *et al.*, 2006). The Medical Dictionary for Regulatory Affairs (MedDRA), which is used to code the adverse event data, is a hypergranular hierarchical thesaurus in which a medical concept may be fragmented in the database across multiple conceptually similar but literally distinct preferred terms. We did not know the extent to which this neutralizes some or all of the performance differentials observed. For example, if one method highlighted an association between a drug and the cardiac preferred term cardiomyopathy and another did not, or one highlighted it earlier than another, the clinical significance of this difference would depend on the simultaneous results of both methods with related terms such as myocarditis. If the method that failed to highlight cardiomyopathy did highlight myocarditis, and this represents variability in coding a single concept, then the differences from a practical perspective are not significant. Ultimately, semantic search approaches may lessen these biases by allowing many related event terms to be coalesced into a single event category. Several groups have begun using methods to explore ontology models, and more recently to analyze the MedDRA ontology (Bisgin *et al.*, 2011; Dupuch *et al.*, 2012; Reich *et al.*, 2012). However, a full exploration of this methodology is beyond the scope of this article.

Regardless of these caveats, we believe that the SG approach represents a new and freely accessible methodology for the pharmacovigilance and research communities. The open nature of the SG method offers substantial opportunities for further exploration. We present one such opportunity in the potential utility of using negative SG scores to identify alternative indications for existing drugs and drug classes.

Funding: K.J., M.G., V.W. and M.H. were employees of Pfizer Inc. at the time this work was done. C.G. was funded with a grant from Pfizer.

Conflict of Interest: none declared.

REFERENCES

Al-Nouri,Z.L. and George,J.N. (2012) Drug-induced thrombocytopenia: an updated systematic review. *Drug Saf.*, **35**, 693–694.

Bate,A. *et al.* (1998) A Bayesian neural network method for adverse drug reaction signal generation. *Eur. J. Clin. Pharmacol.*, **54**, 315–321.

Bate,A. and Evans,S.J.W. (2009) Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol. Drug Saf.*, **18**, 427–436.

Bisgin,H. *et al.* (2011) Mining FDA drug labels using an unsupervised learning technique—topic modeling. *BMC Bioinformatics*, **12 (Suppl. 10)**, S11.

Braesch-Andersen,S. *et al.* (1992) Dopamine-induced lymphoma cell death by inhibition of hormone release. *Scand. J. Immunol.*, **36**, 547–553.

Brunello,N. *et al.* (2006) Acetylsalicylic acid accelerates the antidepressant effect of fluoxetine in the chronic escape deficit model of depression. *Int. Clin. Psychopharmacol.*, **21**, 219–225.

DuMouchel,W. (1999) Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am. Stat.*, **53**, 177–190.

Duncan,R.E. and Bazinet,R.P. (2010) Brain arachidonic acid uptake and turnover: implications for signaling and bipolar disorder. *Curr. Opin. Clin. Nutr. Metab. Care*, **13**, 130–138.

Dupuch,M. *et al.* (2012) Grouping the pharmacovigilance terms with a hybrid approach. *Stud. Health Technol. Inform.*, **180**, 235–239.

Evans,S.J.W. *et al.* (2001) Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol. Drug Saf.*, **10**, 483–486.

Finney,D.J. (2003) *From Thalidomide to Pharmacovigilance: A Personal Account. A Worldwide Yearly Survey of New Data and Trends in Side Effects of Drugs.* Vol. 26, Elsevier, Amsterdam, The Netherlands.

Hauben,M. (2004) Early postmarketing drug safety surveillance: data mining points to consider. *Ann. Pharmacother.*, **38**, 625–630.

Hauben,M. and Aronson,J.K. (2009) Defining signal and its subtypes in pharmacovigilance based on a systematic review of previous definitions. *Drug Saf.*, **32**, 99–110.

Hauben,M. and Bate,A. (2009) Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov. Today*, **14**, 343–357.

Hauben,M. and Reich,L. (2005) Potential utility of data-mining algorithms for early detection of potentially fatal/disabling adverse drug reactions: a retrospective evaluation. *J. Clin. Pharmacol.*, **45**, 378–384.

Hauben,M. and Zhou,X. (2003) Quantitative methods in pharmacovigilance: focus on signal detection. *Drug Saf.*, **26**, 159–186.

Hauben,M. *et al.* (2004) Postmarketing surveillance of potentially fatal reactions to oncology drugs: potential utility of two signal-detection algorithms. *Eur. J. Clin. Pharmacol.*, **60**, 747–750.

Hauben,M. *et al.* (2005) The role of data mining in pharmacovigilance. *Expert Opin. Drug Saf.*, **5**, 929–948.

Hauben,M. *et al.* (2006) What counts in data mining? *Drug Saf.*, **29**, 827–832.

Hauben,M. *et al.* (2007) Illusions of objectivity and a recommendation for reporting data mining results. *Eur. J. Clin. Pharmacol.*, **63**, 517–521.

Hochberg,A.M. *et al.* (2007) Using data minging to predict safety actions from FDA Adverse Event Reporting System data. *Drug Inf. J.*, **41**, 633–643.

Hochberg,A.M. *et al.* (2009) An evaluation of three signal-detection algorithms using a highly inclusive reference event database. *Drug Saf.*, **32**, 509–525.

Ketterer,M.W. *et al.* (1996) Is aspirin, as used for antithrombosis, an emotion-modulating agent? *J. Psychosom. Res.*, **40**, 53–58.

Kim,H.-W. *et al.* (2009) Altered arachidonic acid cascade enzymes in postmortem brain from bipolar disorder patients. *Mol. Psychiatry*, **16**, 1–10.

Lieberman,H.R. *et al.* (1987) The effects of caffeine and aspirin on mood and performance. *J. Clin. Psychopharmacol.*, **7**, 315–320.

Machado-Vieira,R. *et al.* (2009) The role of lithium in the treatment of bipolar disorder: convergent evidence for neurotrophic effects as a unifying hypothesis. *Bipolar Disord.*, **11**, 92–109.

Mendlewicz,J. *et al.* (2006) Shortened onset of action of antidepressants in major depression using acetylsalicylic acid augmentation: a pilot open-label study. *Int. Clin. Psychopharmacol.*, **21**, 227–231.

Mller,N. (2010) COX-2 inhibitors as antidepressants and antipsychotics: clinical evidence. *Curr. Opin. Investig. Drugs*, **11**, 31–42.

Noren,N.G. *et al.* (2006) Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Stat. Med.*, **25**, 3740–3757.

Reese,J.A. *et al.* (2010) Identifying drugs that cause acute thrombocytopenia: an analysis using 3 distinct methods. *Blood*, **116**, 2127–2133.

Reich,C. *et al.* (2012) Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *J. Biomed. Inform.*, **45**, 689–696.

Sachlos,E. *et al.* (2012) Identification of drugs including a dopamine receptor antagonist that selectively target cancer stem cells. *Cell*, **149**, 1284–1297.

Santovito,D. *et al.* (2009) Cyclooxygenase and prostaglandin synthases: roles in plaque stability and instability in humans. *Curr. Opin. Lipidol.*, **20**, 402–408.

Stokes,M.E. *et al.* (2000) *Categorical Data Analysis Using the SAS(R) System*. SAS Institute Inc., Cary, NC, USA.

Stolk,P. *et al.* (2010) Is aspirin useful in patients on lithium? A pharmacoepidemiological study related to bipolar disorder. *Prostaglandins Leukot. Essent. Fatty Acids*, **82**, 9–14.

van Puijenbroek,E.P. *et al.* (2002) A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol. Drug Saf.*, **11**, 3–10.

Wick,M.M. (1981) Levodopa and dopamine analogs: dihydroxy and trihydroxybenzylamines as novel quinol antitumor agents in experimental leukemia in vivo. *Cancer Treat. Rep.*, **65**, 861–867.