Drug Safety 2006; 29 (10): 875-887
0114-5916/06/0010-0875/$39.95/0

© 2006 Adis Data Information BV. All rights reserved.

Comparative Performance of Two Quantitative Safety Signalling Methods

Implications for Use in a Pharmacovigilance Department

June S. Almenoff,[1] *Karol K. LaCroix,*[1] *Nancy A. Yuen,*[1] *David Fram*[2] and *William DuMouchel*[2]

1 Global Clinical Safety and Pharmacovigilance, GlaxoSmithKline, Research Triangle Park, North Carolina, USA
2 Lincoln Technologies (a subsidiary of Phase Forward Inc.), Waltham, Massachusetts, USA

Abstract

Background and objectives: There is increasing interest in using disproportionality-based signal detection methods to support postmarketing safety surveillance activities. Two commonly used methods, empirical Bayes multi-item gamma Poisson shrinker (MGPS) and proportional reporting ratio (PRR), perform differently with respect to the number and types of signals detected. The goal of this study was to compare and analyse the performance characteristics of these two methods, to understand why they differ and to consider the practical implications of these differences for a large, industry-based pharmacovigilance department.

Methods: We compared the numbers and types of signals of disproportionate reporting (SDRs) obtained with MGPS and PRR using two postmarketing safety databases and a simulated database. We recorded signal counts and performed a qualitative comparison of the drug-event combinations signalled by the two methods as well as a sensitivity analysis to better understand how the thresholds commonly used for these methods impact their performance.

Results: PRR detected more SDRs than MGPS. We observed that MGPS is less subject to confounding by demographic factors because it employs stratification and is more stable than PRR when report counts are low. Simulation experiments performed using published empirical thresholds demonstrated that PRR detected false-positive signals at a rate of 1.1%, while MGPS did not detect any statistical false positives. In an attempt to separate the effect of choice of signal threshold from more fundamental methodological differences, we performed a series of experiments in which we modified the conventional threshold values for each method so that each method detected the same number of SDRs for the example drugs studied. This analysis, which provided quantitative examples of the relationship between the published thresholds for the two methods, demonstrates that the signalling criterion published for PRR has a higher signalling frequency than that published for MGPS.

Discussion and conclusion: The performance differences between the PRR and MGPS methods are related to (i) greater confounding by demographic factors with

PRR; (ii) a higher tendency of PRR to detect false-positive signals when the number of reports is small; and (iii) the conventional thresholds that have been adapted for each method. PRR tends to be more ‘sensitive’ and less ‘specific’ than MGPS. A high-specificity disproportionality method, when used in conjunction with medical triage and investigation of critical medical events, may provide an efficient and robust approach to applying quantitative methods in routine postmarketing pharmacovigilance.

Background

Detection of safety signals for marketed products is a challenging yet critical aspect of product lifecycle management. Postmarketing adverse event data are difficult to assess quantitatively because they are derived from voluntary reports (introducing uncertainty regarding the extent of under-reporting) and because they are not easily linked to meaningful drug exposure data for the reporting population. Historically, postmarketing signal detection has relied on qualitative methods applied to individual case reports and case series (e.g. global introspection, rule-based methods) supplemented by simple frequency-based methods (e.g. serial reporting rates, clustering). However, none of these methods are satisfactory for efficiently and systematically screening large adverse event databases for safety signals.[1]

Recently, statistical methods based on the disproportionality of adverse event reporting have become more widely used by regulatory agencies and the pharmaceutical industry.[2-7] These ‘data mining’ methods detect drug-adverse event combinations reported with a frequency that is disproportionately high with respect to some computed ‘baseline’. The baseline is typically derived from what amounts to an ‘independence model’ for the reporting of drugs and events. Two of the most commonly used disproportionality methods are the proportional reporting ratio (PRR)[2] and the empirical Bayes multi-item gamma Poisson shrinker (MGPS).[8,9] These methods can be used to screen large adverse event databases systematically for safety signals. When used in conjunction with traditional qualitative methods, disproportionality methods have a strong potential to enhance the efficiency of signal detection, without compromising the critically important medical judgement that is integral to the process.

One concern that has been expressed about disproportionality methods is that routine use could generate large numbers of alerts by chance alone (i.e. ‘false’ alerts). Each false alert would prompt further investigation in order to determine whether a signal exists, potentially compromising the efficiency of the overall process.[10,11] Based on our knowledge of the methodologies and their published thresholds, we expect that PRR will generate a greater number of alerts than MGPS.[12] Several retrospective comparisons have explored the sensitivity of the two methods in the context of selected safety issues.[13-16] However, a systematic, comprehensive evaluation of the operating characteristics of these two methods has not previously been performed in a real-world pharmacovigilance setting. In the present study, we compared the performance of PRR and MGPS when systematically and objectively applied to two large postmarketing safety databases. We then explored the reasons for the observed differences and considered the practical implications of using these methods for routine signal detection in a large pharmacovigilance department.

Methods

Databases

Adverse Event Reporting System Database

Adverse Event Reporting System (AERS) is the US FDA’s postmarketing safety database for non-vaccine drug products marketed in the US. The public-release version of AERS that was used in this study contains >2 million reports involving approxi-

© 2006 Adis Data Information BV. All rights reserved.

mately 5000 marketed drug products (including single-ingredient products and combination products) and covers the period 1968 through December 2002. AERS is a passive surveillance system that relies on voluntary reporting of adverse events to the FDA by healthcare professionals and consumers, as well as required reporting by pharmaceutical manufacturers. The database includes all spontaneous reports from US sources; serious and unlabelled spontaneous reports from non-US sources; and serious, unlabelled and attributable adverse events from clinical trials worldwide. Adverse event terms in the release of AERS used in this study were coded using the Medical Dictionary for Regulatory Activities (MedDRA) version 6.0.

Proprietary Adverse Event Database

The second database used for this analysis was the GlaxoSmithKline proprietary database Operating Company Event Accession and Notification System (OCEANS). Specifically, we used a subset of OCEANS consisting of voluntary, spontaneous reports involving non-vaccine products that were received from healthcare professionals and consumers. At the time that this study was conducted, OCEANS contained approximately 500 000 spontaneous adverse event reports and covered the period 1960 through September 2003. Adverse event terms in OCEANS were coded using MedDRA version 6.0.

Selection of Drugs

Five drugs were selected for study in AERS and five drugs were selected for study in OCEANS. The drugs selected for study had the following attributes: (i) they were each currently marketed in the US; (ii) they were each contained in at least 500 reports for AERS or 300 reports for OCEANS; and (iii) they were each a non-injectable, non-vaccine, prescription drug product of a different therapeutic/pharmacological class than any of the other drugs selected.

We used the following procedure to randomly select five drugs from each database: random numbers were generated using the random number generation tool in the Microsoft Excel 97 spreadsheet software. Each random number was matched to a numbered list of all suspect drugs contained in the database, ordered highest to lowest by report count. If the drug that was picked out by the current random number met the selection criteria, that drug was included in the study and the next random number was matched to the list. If the drug did not meet the selection criteria, that drug was rejected for the study and, again, the next random number was matched to the list. This matching procedure continued until five drugs were selected for each of the two databases.

Data Mining Algorithms

The implementations of the MGPS and PRR disproportionality methods used in this study were those contained in the safety data mining computer system described by Fram et al.[7] (WebVDME, Lincoln Technologies, Inc., Waltham, MA, USA). The WebVDME system provides a Web-based interface for specifying, executing and reviewing data mining analyses of databases of spontaneous adverse drug event reports. The disproportionality methods in theory can be based on counting either reports or drug-event combinations; these two approaches are not identical since an individual report can mention more than one drug and more than one event. The implementations of MGPS and PRR used for this study were based on counting reports as opposed to drug-event combinations.

Multi-item Gamma Poisson Shrinker Algorithm

The MGPS algorithm computed the empirical Bayes geometric mean (EBGM) and corresponding two-sided 90% confidence intervals (EB05, EB95) for each observed drug-event combination in each of the databases studied.[8,9] Expected counts (E) were calculated using a stratified, full-independence model. EBGM values represent the relative reporting rates after Bayesian smoothing (adjusted ratios of observed counts to expected counts, N/E) for the corresponding drug-event combinations. An EBGM value of 10 indicates that a drug-event combination has been reported to the database ten times as frequently as expected if there were no statistical association between reporting of the drug and reporting of the event.

© 2006 Adis Data Information BV. All rights reserved.

Stratification

Drug utilisation and adverse event occurrence are often correlated with patient demographic variables such as age and sex or with secular variables such as the year of report. In order to not confound 'true' drug-event associations with 'spurious' associations arising from such independent correlations with patient or secular characteristics, MGPS uses a Mantel-Haenszel[17] approach to stratify the computation of expected counts. MGPS first computes separate values of expected counts for drug-event combinations within each distinct stratum defined by the chosen stratification variables and then computes an estimate of the relative reporting rate as RR = N/E, where N is the sum of observed counts across strata and E is the sum of expected counts across strata. In this study, MGPS was performed with stratification for age, sex and year of report, unless otherwise specified.

Proportional Reporting Ratio Algorithm

The PRR algorithm computed the PRR value and associated Chi-squared hypothesis-test statistic (with Yates correction) for each observed drug-event combination in each database. To obtain the value for PRR, the proportion of reports for a particular drug that mention a particular event is divided by the proportion of reports not containing that drug that mention that same event.[18] The computation of PRR was performed without stratification, as described by Evans et al.[2]

Analysis

EBGM (EB05, EB95) and PRR (Chi-squared) values were computed for all possible drug-event combinations for each database. All data mining was performed at the level of MedDRA preferred terms (PTs), and all data mining runs were configured to analyse only drugs reported as suspect (i.e. drugs reported as concomitant were not included in the analysis). Results were then filtered to display all the drug-event combinations observed in the database for the selected drugs. Drug-event combinations with EB05 ≥2 were considered 'signalled by MGPS' based on the empiric threshold described by Szarfman et al.[4] Drug-event combinations with a PRR ≥2, Chi-squared ≥4, and N ≥3 were considered 'signalled by PRR' based on the empiric threshold described by Evans et al.[2]

In order to distinguish an alert based on exceeding empiric thresholds using disproportionality methods from signals generated with more traditional methods, the term 'signals of disproportionate reporting' (SDRs) is used in this article to denote those drug-event combinations exceeding an empiric threshold.[19] The percentage of possible SDRs detected was defined as the number of events (MedDRA PTs) signalled for a drug divided by the number of possible SDRs that could occur with that particular drug. The number of possible SDRs for a particular drug equals the number of unique MedDRA PTs reported at least once for that drug. This definition assumes that any event reported at least once for a given drug is a potential SDR.

Medical assessment of event terms signalled by only one of the methods involved a review of the terms only, not a review of the cases, and was conducted in an unblinded fashion, i.e. the assessor had knowledge of which method had signalled the events.

Simulations were performed as described by Rolka et al.[20] using the data simulation function in WebVDME. Simulated or 'artificial' databases of adverse event reports were constructed so that there would be no statistical associations between co-reported drug and event terms. Each artificial report was constructed by incorporating the set of drugs from one randomly selected report in AERS and the set of events from a different randomly selected report in AERS.

Results

The following drugs were selected for study using the procedure described in the methods section: from OCEANS, bupropion, trimethoprim, abacavir, busulfan and nabumetone; from AERS, budesonide, mirtazapine, terazosin, losartan and timolol. All ten drugs selected can be described as 'mature' drugs, with the year of first marketing in the US ranging from 1954 to 1998. No drug selected from AERS is marketed in the US by GlaxoSmithKline.

© 2006 Adis Data Information BV. All rights reserved.

Table I. Comparison of numbers of signals of disproportionate reporting (SDRs) detected using the multi-item gamma Poisson shrinker (MGPS) and proportional reporting ratio (PRR) methods for selected drugs

Drug	Year first marketed in the US	No. of possible SDRs	Possible SDRs detected [n (%)]		PRR : MGPS[a]
			MGPS	PRR	
OCEANS database					
Bupropion	1985	2606	23 (0.9)	160 (6.1)	7
Trimethoprim	1980[b]	279	6 (2.2)	28 (10)	4.7
Abacavir	1998	1293	141 (10.9)	286 (22.1)	2
Busulfan	1954	457	44 (9.6)	75 (16.4)	1.7
Nabumetone	1991	1232	44 (3.6)	137 (11.1)	3.1
AERS database					
Budesonide	1994	370	46 (12.4)	148 (40.0)	3.2
Mirtazapine	1996	441	25 (5.7)	171 (38.8)	6.8
Terazosin	1987	376	20 (5.3)	49 (13.0)	2.5
Losartan	1995	579	24 (4.1)	191 (33.0)	8
Timolol	1978	346	51 (14.7)	74 (21.4)	1.5

a The mean ratio for the results from the OCEANS database was 3.7 and from the AERS database was 4.4.

b Non-GlaxoSmithKline.

AERS = Adverse Event Reporting System; **OCEANS** = Operating Company Event Accession and Notification System.

Table I shows the number and percentage of possible SDRs detected from the OCEANS and AERS databases for the ten (five per database) randomly selected drugs using both MGPS (with stratification) and PRR. With the OCEANS database, for all five drugs, the percentage of possible SDRs detected was markedly higher with PRR than with MGPS. Similar results were observed when MGPS and PRR were applied to the AERS data. On average, PRR identified three to four times as many SDRs as MGPS for each drug.

Table II shows that for seven of the ten drugs studied in the OCEANS and AERS databases, the drug-event combinations signalled by MGPS were a subset of those signalled by PRR (i.e. there were no events signalled by MGPS that were not detected with PRR). For the remaining three drugs (bupropion, terazosin and timolol), MGPS detected 23, 20 and 51 signals, respectively (table I), of which only 1, 1 and 3, respectively, were not detected with PRR. Hence, the large differences in the numbers of signals detected are almost entirely due to events being signalled by PRR but not by MGPS. Medical assessment of the adverse event terms for the signals detected by PRR but not by MGPS revealed that the majority were either (i) closely related to terms signalled by both methods (e.g. ‘orthostatic hypotension’ signalled by both methods for terazosin, but ‘hypotension’ signalled by PRR but not by MGPS) or (ii) likely disease related. (Assessments of disease-relatedness were based on evaluation of the specific event terms in the context of the well established safety profiles of these mature products, as well as their treatment indications.) However, there

Table II. Number of signals of disproportionate reporting (SDRs) detected by either the multi-item gamma Poisson shrinker (MGPS) or the proportional reporting ratio (PRR) method only

Drug	MGPS only	PRR only
OCEANS		
Bupropion	1	138
Trimethoprim	0	22
Abacavir	0	145
Busulfan	0	31
Nabumetone	0	93
AERS		
Budesonide	0	102
Mirtazapine	0	146
Terazosin	1	30
Losartan	0	167
Timolol	3	26

AERS = Adverse Event Reporting System; **OCEANS** = Operating Company Event Accession and Notification System.

© 2006 Adis Data Information BV. All rights reserved.

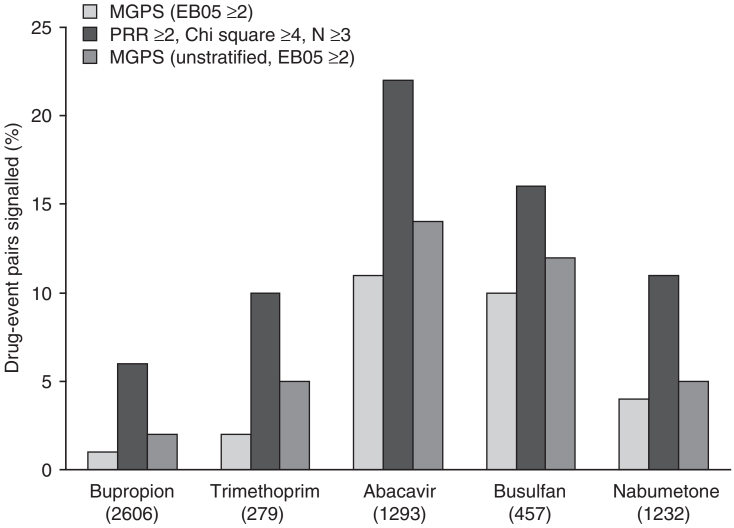

Fig. 1. Comparison of the percentage of signals of disproportionate reporting (SDRs) detected using the multi-item gamma Poisson shrinker (MGPS) and proportional reporting ratio (PRR) methods in GlaxoSmithKline's proprietary Operating Company Event Accession and Notification System (OCEANS) database. The numbers of possible SDRs are given in parentheses below the drug names. **EB05** = lower 5% confidence bound of the empirical Bayes geometric mean; **N** = the number of reports for a given drug-event combination.

were some events signalled by PRR alone that may represent potentially medically significant issues. For this reason, medical triage is also an important component of the signal detection process. The vast majority of the events signalled by PRR alone involved small numbers of reports (usually fewer than ten).

In order to further characterise these results, we examined three factors that we believed might account for the observed differences in signalling frequency between PRR and MGPS: (i) confounding due to demographic variables; (ii) the instability of PRR with small numbers of reports; and (iii) the signalling thresholds conventionally used with each method.

Confounding by Demographic/Temporal Factors

The computation of PRR as described by Evans et al.[2] employs no stratification scheme, and thus many SDRs represent apparent drug-event associations that are actually due to separate associations between a drug and a demographic/secular variable and between an event and the same demographic/secular variable. In contrast, MGPS stratifies the data by age, sex and year of report.[9] To explore this hypothesis, we repeated the previous analysis for both AERS and OCEANS using MGPS without stratification and again compared the percentage of possible SDRs detected. The results, illustrated in figure 1 and figure 2, show that PRR detected more signals in comparison with unstratified MGPS as well as in comparison with stratified MGPS, but the average difference (≈2-fold) was not as great as with stratified MGPS (3- to 4-fold). As with stratified MGPS, unstratified MGPS detected almost no SDRs that were not also detected by PRR; the SDRs not detected by both methods consisted primarily of associations detected by PRR alone. These results suggest that the absence of stratification with PRR accounts for some, but not all, of the differences seen between the two methods. These results complement the results of studies describing the performance characteristics of PRR.[21,22]

One variable used for stratification is patient sex. One of a number of examples of how stratification serves to minimise confounding by demographic factors was seen for the product terazosin, an α-adrenoceptor antagonist. The following adverse

© 2006 Adis Data Information BV. All rights reserved.

event terms were signalled with unstratified but not stratified MGPS: 'Testicular disorder NOS' (EB05 unstratified = 2.78, stratified = 1.27) and 'Penile disorder NOS' (EB05 unstratified = 4.51, stratified = 1.81). This product is used for the treatment of benign prostatic hyperplasia, a disease seen only in men. In this setting, stratification limits the signalling of apparent drug-event associations that are actually due to independent relationships between the drug and a demographic characteristic (in this case, male sex) and an event and the same characteristic.

Performance When the Number of Reports Is Small

PRR is unstable when N is small. Data mining in large databases involves the determination of relative reporting ratios for millions of possible drug-event pairs having a wide range of counts. The variance of the PRR statistic is quite large when N is small. The PRR signalling rule attempts to address this problem by requiring that the Chi-squared 'test' statistic be ≥4, corresponding to a one-sided p-value of about 0.025, i.e. a statistical 'false alarm' rate of 2.5%.[2] However, there is no adjustment for the multiple tests being carried out: this rule still leaves room for thousands of statistically significant but false-positive results due to chance alone.

In contrast, MGPS uses an empirical Bayesian approach that is more robust than PRR in the face of the multiple-comparisons challenge. The Bayesian method tends to shrink the estimated ratio N/E when N or E is small, with the notion that these small numbers are subject to high variance and therefore produce less reliable estimates of N/E than in situations where both N and E are large. The precise amount of shrinkage is determined by an analysis of the entire ensemble of (N, E) pairs; a 'prior distribution' is estimated in a maximum-likelihood fitting process allowing Bayesian statistical theory to determine the shrinkage formula.[8,9] The practical effect is that the ratio N/E is left almost unmodified when N and E are large (e.g. N >20 and E >1), while the ratio can be substantially reduced for much smaller values of N and/or E. Thus, the signalling criterion for small counts is much stricter for MGPS than for PRR. We carried out two investigations to explore the relative instability of PRR and EBGM for small N and to determine how this might contrib-

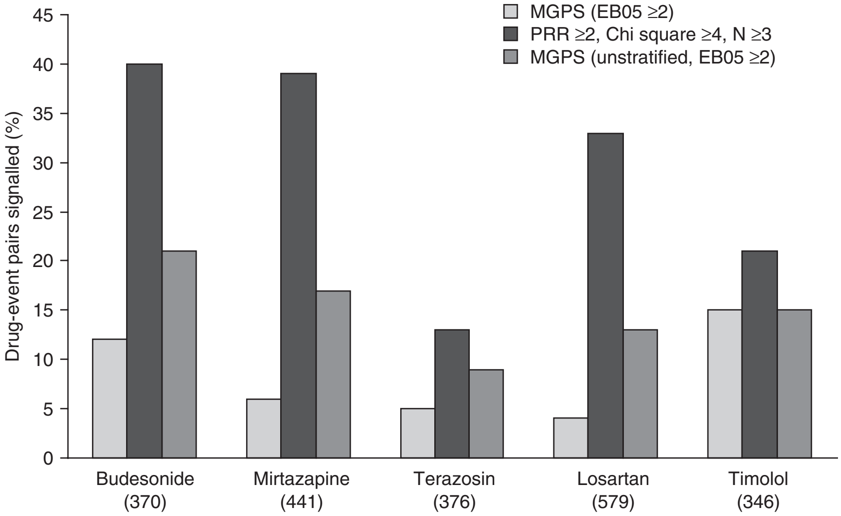

Fig. 2. Comparison of the percentage of signals of disproportionate reporting (SDRs) detected using the multi-item gamma Poisson shrinker (MGPS) and proportional reporting ratio (PRR) methods in the US FDA's Adverse Event Reporting System (AERS) database. The number of possible SDRs is given in parentheses below the drug names. **EB05** = lower 5% confidence bound of the empirical Bayes geometric mean; **N** = the number of reports for a given drug-event combination.

© 2006 Adis Data Information BV. All rights reserved.

Table III. Timolol: cumulative distributions [n (%)] of number of reports (N) for signals of disproportionate reporting (SDRs) detected by the multi-item gamma Poisson shrinker (MGPS) method and the proportional reporting ratio (PRR) method

N	MGPS – stratified (51 SDRs)	PRR (74 SDRs)
≤3	0 (0)	8 (11)
≤4	2 (4)	14 (19)
≤5	2 (4)	15 (20)
≤6	7 (14)	19 (26)
≤7	9 (18)	24 (32)
≤8	10 (20)	26 (35)
≤9	10 (20)	27 (36)
≤10	12 (24)	30 (41)

ute to the observed overall differences in signalled drug-event associations.

To determine whether the aforementioned concepts account for the differences seen between the signalling behaviours of MGPS and PRR, we analysed the distribution of N associated with SDRs for each of the drugs studied, for each of the methods. First, for each of the selected drugs within each database, we calculated the percentage of SDRs as a function of N up to N = 10. A representative example is shown for the drug timolol in table III. For timolol, 19% of PRR SDRs involve N ≤4, compared with 4% of MGPS SDRs. These results support the aforementioned hypotheses, in that a substantial fraction of SDRs that are uniquely detected with PRR have small values of N (≤10).

Next, we compared the performance of MGPS and PRR using simulated or ‘artificial’ databases of adverse event reports that were carefully constructed so that there would be no statistical associations between co-reported drug and event terms. The data simulation function in WebVDME was used to generate three such artificial databases, each containing 15 000 artificial reports. Each artificial report was constructed by incorporating the set of drugs from one randomly selected report in AERS and the set of events from a different randomly selected report in AERS. By combining in each newly constructed report the entire list of drugs from one actual AERS report and the entire list of events from another actual AERS report, any associations among drugs (e.g. commonly co-prescribed drugs) and among events (e.g. commonly co-reported symptoms or syndromes) would be preserved, while associations between drugs and events would be eliminated. Consequently, the reports in the simulated databases contained realistic case elements but, by design, no relationships between the drugs and events. MGPS calculations with these databases were performed without stratification since no attempt was made to incorporate demographic and temporal information into the simulated reports. The signalling criteria for MGPS and for PRR were defined as in the previous analyses (see the Methods section). Because these simulated databases were explicitly constructed so as to contain no drug-event associations, we considered any drug-event combination result exceeding the published threshold to be a false-positive SDR.

The results from these three independent data simulations are shown in table IV. All data mining runs yielded similar results for a given method across the three datasets. Each dataset contained approximately 32 000 drug-event pairs (and thus 32 000 potential SDRs). PRR detected an average of 361 SDRs (1.1% of 32 000) for the three runs, which is consistent with the Type I error that would be predicted for a Chi-squared analysis with 1 degree

Table IV. Data simulations

Simulated database	No. of drug-event pairs (i.e. possible SDRs in the database)	PRR: no. of drug-event pairs meeting signal threshold[a]	MGPS: no. of drug-event pairs where EB05:		
			>1	>1.5	≥2
1	31 945	346	37	3	0
2	32 352	358	72	2	0
3	32 075	379	64	1	0

a N ≥3, PRR ≥2, and Chi-squared ≥4.

EB05 = lower 5% confidence bound of the empirical Bayes geometric mean; **MGPS** = multi-item gamma Poisson shrinker; **PRR** = proportional reporting ratio; **SDR** = signal of disproportionate reporting.

of freedom. A lower proportion than the nominal significance level of 2.5% shows up because of the additional constraints applied (N ≥3, PRR ≥2) beyond the Chi-squared requirement. In contrast, MGPS detected no SDRs (EB05 ≥2) from any of the simulated datasets. The MGPS empirical Bayes analysis of the ensemble of (N, E) pairs in the simulated dataset detects that the overall variation is consistent with the following null hypothesis: every count N is Poisson distributed with its respective mean and variance equal to E, determined from the independence assumption. (The alternative hypothesis is that there is 'extra-Poisson variation' due to associations between drugs and events.) As a consequence, all N/E ratios are subject to severe shrinkage. Table IV shows that none of the lower 5% limits of the relative reporting rate are as high as 2 and only 1–2% of them are even as high as 1.

Signalling Criteria (Thresholds)

It seems apparent that PRR and MGPS are not operating at the same signalling frequency level when the published criteria associated with each method are used. To further investigate the influence of the signalling criteria, we adjusted the thresholds employed with each method so that each method detected the same number of SDRs (specifically, the average of the numbers detected using the two methods), thereby explicitly creating a situation where the two methods were operating at the same signalling frequency level. This enabled us to better understand the degree to which adoption of the conventional, empirically selected thresholds associated with the two methods accounts for the large observed differences in signal (SDR) volume. It also allowed us to gain insight into which types of signals are 'favoured' by each method under conditions of equivalent overall signalling frequency.

Table V shows the adjustment process for the respective drugs used in the AERS and OCEANS database examples. The columns show, respectively, the number of SDRs detected using MGPS (EB05 ≥2) and the number of adverse events signalled using PRR (PRR ≥2, N ≥3, Chi-squared ≥4). The column headed 'average no. of SDRs' is the average of the counts in the previous two columns, rounded down to the nearest integer. The next two columns show what the thresholds for MGPS and

Table V. Number of signals of disproportionate reporting (SDRs) for each method and the thresholds needed to equalise them

Drug	No. of possible SDRs detected		Average no. of SDRs	Adjusted threshold to achieve average		No. alerted by both methods
	EB05 ≥2	PRR ≥2[a]		EB05	PRR[a]	using adjusted thresholds[b]
OCEANS						
Abacavir	141	286	213	1.391	4.012	160
Bupropion	23	160	91	1.362	3.246	60
Busulfan	44	75	59	1.296	4.389	52
Nabumetone	44	137	90	1.257	3.673	61
Trimethoprim	6	28	17	1.102	4.525	10
AERS						
Budesonide	46	148	97	0.980	4.254	76
Losartan	24	191	107	1.217	3.841	68
Mirtazapine	25	171	98	0.932	4.113	50
Terazosin	20	51	35	1.408	3.130	25
Timolol	51	74	62	1.693	2.896	52

[a] Additional standard criteria of N ≥3 and Chi-squared ≥4 were used.

[b] The number of overlapping preferred terms (within the 'average no. of SDRs') that were the same by both methods after adjustment of thresholds.

EB05 = lower 5% confidence bound of the empirical Bayes geometric mean; **AERS** = Adverse Event Reporting System; **OCEANS** = Operating Company Event Accession and Notification System; **PRR** = proportional reporting ratio.

© 2006 Adis Data Information BV. All rights reserved.

PRR would have to be in order to get the average number of SDRs for each drug. For example, with budesonide in the AERS database, using the published signalling criteria, there were 46 SDRs using MGPS and 148 SDRs using PRR. When the signalling criteria are adjusted so that MGPS and PRR each detect 97 signals (the average of 46 and 148), the MGPS criterion is EB05 ≥0.980 and the PRR criterion is PRR ≥4.254, N ≥3 and Chi-squared ≥4. The 97 PTs signalled are not necessarily the same for the two methods; hence, the last column in table V shows how many PTs overlap. For budesonide, 76 of 97 overlap, and each method flags 21 PTs (97 − 76) that the other does not. These data show that, to detect the same number of SDRs, the PRR threshold must be adjusted upward by a factor of 1.5–2 and the MGPS threshold must be adjusted downward by a factor of 2–3. This analysis demonstrates, in a concrete way, that the signalling criteria that have been published for PRR are much less stringent (i.e. resulting in a higher signalling frequency) than those that have been published for MGPS.

Discussion

We sought to understand the operating characteristics of two disproportionality methods, MGPS and PRR, when applied to a large, public adverse event reporting database (AERS), a smaller, proprietary adverse event database (OCEANS) and a simulated database. Our data show that, using published thresholds, PRR detects many more SDRs (up to four times as many) than MGPS for all drugs studied. We investigated the qualitative nature of the differences observed when using the published thresholds and found that many of the excess SDRs of PRR over MGPS are related to the following factors: (i) confounding within the unstratified PRR analysis, such that the existence of simultaneous relationships between a drug and demographic/temporal factors (i.e. age, sex and year of report) and between an event and the same factors result in detection of an apparent, but likely spurious, drug-event association; (ii) the variation inherent in the calculation of PRR with small sample sizes and without adjustment for multiple comparisons; and (iii) a less stringent threshold for PRR as compared with MGPS.

Regarding the first issue (confounding), MGPS employs an adjustment technique, the Mantel-Haenszel method, that is commonly employed in pharmacoepidemiology to reduce confounding by demographic and temporal factors. This adjustment clearly reduces the number of SDRs detected by MGPS and, importantly, the associations not detected were almost certainly related to demographic characteristics of the populations being treated (e.g. elimination of SDRs for menstrual abnormalities in a patient population that commonly experiences these symptoms in the absence of medications). It is generally believed that stratification is an important tool to allow safety evaluators to focus on drug-event combinations that are more likely to be causally related, since they are not just a reflection of constellations of population-related prescribing patterns and population-related symptom susceptibilities and morbidities. The PRR calculations in this study were performed without stratification, as it was our intention to study PRR as described by Evans et al.[2] However, it is possible to perform PRR with stratification.

Regarding the second issue, the Bayesian shrinkage calculations that MGPS uses are more conservative for small values of N and E, resulting in fewer SDRs, and the Bayesian theory provides a basis for negotiating the trade-off between sensitivity (detecting true signals) and specificity (avoiding false signals). Of course, there is no perfect solution to this trade-off, and no doubt some of the many ‘extra’ SDRs generated by PRR compared with MGPS (when all drug-event combinations are considered) might be the initial signs of a true safety issue.

After accounting for possible confounding by demographic/temporal factors, there remain two situations that will result in elevated signal scores in which the drug-event pair identified may not represent a causal association that is relevant from a drug safety perspective. One situation involves an indirect association, such as a co-morbidity in the

© 2006 Adis Data Information BV. All rights reserved.

population or an adverse effect from a product that is frequently co-prescribed with the drug of interest. Such indirect associations may be termed *reliable* in that they are likely to persist as data accumulates, and only medical knowledge can further explain them. The other situation is that of SDRs detected by chance as a result of small counts; these SDRs may be termed *unreliable* in that the associations are likely to disappear or be greatly reduced as more data accumulate. The MGPS algorithm seems less likely to generate a large signal score in the latter situation.

In our simulation studies using artificial random databases of approximately 32 000 drug-event pairs, PRR ≥2 flagged 361 false-positive signals, or a rate of 1.1%. Such a rate may be unacceptably high when applying a signal-detection method to a large database such as AERS, which contains approximately 1 million observed drug-event pairs. In the simulation, MGPS showed no false positives with a signal threshold of EB05 ≥2 and a false-positive rate of approximately 1–2% with a threshold of EB05 >1. The simulation study helps to clarify the statistical signalling behaviour of the two methods in a simple, well characterised environment. The zero false-positive rate in the simulation with EB05 ≥2 is a reflection of the ability of the Bayesian method to adapt to the overall prevalence of associations in the database (or lack thereof). In a real-world database with a variety of causal and non-causal associations, the MGPS false-positive rate is unlikely to be zero, but the Bayesian approach will help suppress transient associations caused by small counts. Roux et al.[23] have published another simulation experiment comparing several disproportionality measures, including PRR and a Bayesian measure similar to EBGM. Their simulated databases differed from ours in that they were not designed to mimic the exact distributions of events or drugs of a real database. However, they did simulate the occurrence of a variety of drug-event associations rather than our simulation of a no-association scenario. The results of Roux et al.[23] agreed with ours in that the empirical Bayesian shrinkage measures showed fewer false positives than PRR and related measures, especially when counts were small.

We also showed that understanding the relative performance of the two methods requires examination of the chosen thresholds. Our analyses provided quantitative examples of the relationship between the published thresholds for the two methods. These results show that the methods would have similar signalling rates if either the PRR threshold was adjusted upward or the MGPS threshold was adjusted downward. Previously published retrospective analyses comparing the apparent sensitivity of PRR and MGPS to detect known safety issues have been described by Hauben and Reich[13,16] and Hauben.[14,15] These analyses give specific examples where PRR apparently detected an issue sooner than MGPS or where MGPS 'failed' to detect an issue that was signalled by PRR. Although these examples highlight that, using the published thresholds, MPGS may be less 'sensitive' than PRR, they do not provide a corresponding analysis of the specificity characteristics of the two methods and they do not calculate overall sensitivity and specificity.[24] Our findings, also based on retrospective analysis, provide further insight into these apparent differences – that MGPS is more robust than PRR with small numbers of reports and that MGPS generates fewer false-positive signals by minimising the artifacts that can occur with demographic confounding and multiple comparisons.

Prospective studies exploring the performance of these methods over a broad range of products and signalling thresholds would provide valuable information for optimising their use in pharmacovigilance departments. Another area for further study involves the gathering of metrics with regard to whether these tools add efficiency over the use of traditional methods alone.

Conclusions

The detection of safety signals in postmarketing adverse event databases is an important activity that helps to monitor the safety of marketed products. Routine pharmacovigilance using these databases typically involves computation of crude reporting

© 2006 Adis Data Information BV. All rights reserved.

rates (number of reports concerning the drug-event pair of interest divided by estimated exposure to drug), medical review of individual case reports and subjective assessments of case series (i.e. global introspection). While these approaches are valuable for evaluating relationships between drugs and events, they are not an efficient or accurate means of determining whether a drug-event pair has been reported with high frequency. Crude reporting rates can be difficult to interpret because postmarketing reporting of adverse events is voluntary (the numerator problem) and because drug exposure can be difficult to estimate accurately and systematically (the denominator problem). Consequently, it can be difficult for a safety evaluator to put crude reporting frequencies in context. For a pharmacovigilance department with many drugs to monitor, a relative measure of reporting frequency (as provided by disproportionality analysis) may be seen as an improvement over crude reporting frequencies and may have the potential to introduce efficiencies for prioritising the review of large volumes of adverse event reports.

Disproportionality methods applied to postmarketing safety databases cannot be relied upon to replace medical judgment in the signal detection process. These methods do not have the sensitivity to detect all possible drug toxicities, particularly those toxicities that occur very infrequently. Recognising this lack of sensitivity and other limitations of postmarketing reporting systems, we believe that a prudent approach is to couple a high-specificity disproportionality method with ongoing medical triage and a focused review of classic drug toxicities (e.g. torsades de pointes, Stevens-Johnson syndrome), of adverse events consistent with the pharmacological properties of a drug and of other medically significant events. The use of a method that has high specificity enhances the detection of signals that are truly occurring at a greater than expected frequency, thereby minimising the diversion of resources to the investigation of signals that are unlikely to be clinically meaningful. In our experience, this approach optimises the use of an automated statistical method while reducing the potential inefficiency of a high 'false alarm' rate.

Acknowledgements

David Fram and William DuMouchel are employees of Lincoln Technologies (a subsidiary of Phase Forward Inc.), a software development company that markets WebVDME. June S. Almenoff, Karol K. LaCroix and Nancy A. Yuen are employees of GlaxoSmithKline and this work was conducted as part of their employment. No sources of funding were used to assist in the preparation of this review.

References

1. Clark JA, Klincewicz SL, Stang PE. Overview: spontaneous signaling. In: Mann RB, Andrews EB, editors. John Wiley & Sons Ltd, 2002: 247-71
2. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology Drug Saf 2001; 10: 483-6
3. O'Neill RT, Szarfman A. Some FDA perspectives on data mining for pediatric safety assessment. Cur Ther Res Clin Exp 2001; 62 (9): 650-63
4. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf 2002; 25 (6): 381-92
5. Lindquist M, Stahl M, Bate A, et al. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. Drug Saf 2000; 23 (6): 533-42
6. Egberts AC, Meyboom RH, van Puijenbroek EP. Use of measures of disproportionality in pharmacovigilance: three Dutch examples. Drug Saf 2002; 25 (6): 453-8
7. Fram DM, Almenoff JS, DuMouchel W. Empirical Bayesian data mining for discovering patterns in post-marketing drug safety. In: Conference on knowledge discovery in data. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2003 Aug 24-27; Washington, DC. New York: ACM Press, 2003: 359-68
8. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System (with discussion). Am Stat 1999; 53 (3): 177-202
9. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. In: Conference on knowledge discovery in data. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2001 Aug 26-29; San Francisco (CA). New York: ACM Press, 2001: 67-76
10. Faich G. Department of Health and Human Services, US FDA. Risk management public workshop: pharmacovigilance practices and pharmacoepidemiologic assessment: April 11, 2003 [online]: 53-65. Available from URL: http://www.fda.gov/cder/meeting/RMtranscript3.doc [Accessed 2004 Dec 23]
11. Lilienfeld DE. A challenge to the data miners. Pharmacoepidemiol Drug Saf 2004 Dec; 13 (12): 881-4
12. Gould AL. Practical pharmacovigilance analysis strategies. Pharmacoepidemiol Drug Saf 2003; 12: 559-74

© 2006 Adis Data Information BV. All rights reserved.

13. Hauben M, Reich L. Drug-induced pancreatitis: lessons in data mining [letter]. Br J Clin Pharmacol 2004; 58: 560-2
14. Hauben M. Application of an empiric Bayesian data mining algorithm to reports of pancreatitis associated with atypical antipsychotics. Pharmacother 2004; 24: 1122-9
15. Hauben M. Trimethoprim-induced hyperkalaemia: lessons in data mining [letter]. Br J Clin Pharmacol 2004; 58: 338-9
16. Hauben M, Reich L. Safety related drug-labelling changes: findings from two data mining algorithms. Drug Saf 2004; 27: 735-44
17. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959; 22: 719-48
18. Finney DJ. The detection of adverse reactions to therapeutic drugs. Stat Med 1982; 1: 153-61
19. Hauben M, Reich L. Communication of findings in pharmacovigilance: use of the term "signal" and the need for precision in its use. Eur J Clin Pharmacol 2005; 61: 479-80
20. Rolka H, Bracy D, Russell C, et al. Using simulation to assess the sensitivity and specificity of a signal detection tool for multidimensional public health surveillance data. Statist Med 2005; 24: 551-62
21. Kubota K, Koise D, Hirai T. Comparison of data mining methodologies using Japanese spontaneous reports. Pharmacoepidemiol Drug Saf 2004; 13 (6): 387-94
22. van Puijenbroek EP, Bate A, Leufkens HGM, 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
23. Roux E, Thiessard F, Fourrier A, et al. Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. IEEE Trans Inf Technol Biomed 2005; 9: 518-27
24. Levine JG, Tonning JM, Szarfman A. Reply: the evaluation of data mining methods for the simultaneous and systematic detection of safety signals in large databases: lessons to be learned. Br J Clin Pharmacol 2006, 113

Correspondence and offprints: Dr *June S. Almenoff*, GlaxoSmithKline, Research Triangle Park, 5 Moore Drive, Mail Stop 5.4214.4C PO Box 13398, NC 27709-3398, USA.
E-mail: june.s.almenoff@gsk.com

© 2006 Adis Data Information BV. All rights reserved.