Drug Saf (2015) 38:577–587
DOI 10.1007/s40264-015-0289-5

ORIGINAL RESEARCH ARTICLE

Comparison of Statistical Signal Detection Methods Within and Across Spontaneous Reporting Databases

Gianmario Candore[1] · Kristina Juhlin[2] · Katrin Manlik[3] · Bharat Thakrar[4] · Naashika Quarcoo[5] · Suzie Seabroke[6] · Antoni Wisniewski[7] · Jim Slattery[1]

Published online: 22 April 2015

Abstract
Background Most pharmacovigilance departments maintain a system to identify adverse drug reactions (ADRs) through analysis of spontaneous reports. The signal detection algorithms (SDAs) and the nature of the reporting databases vary between operators and it is unclear whether any algorithm can be expected to provide good performance in a wide range of environments.
Objective The objective of this study was to compare the performance of commonly used algorithms across spontaneous reporting databases operated by pharmaceutical companies and national and international pharmacovigilance organisations.
Methods 220 products were chosen and a reference set of ADRs was compiled. Within four company, one national and two international databases, 15 SDAs based on five disproportionality methods were tested. Signals of disproportionate reporting (SDRs) were calculated at monthly intervals and classified by comparison with the reference set. These results were summarised as sensitivity and precision for each algorithm in each database.
Results Different algorithms performed differently between databases but no method dominated all others. Performance was strongly dependent on the thresholds used to define a statistical signal. However, the different disproportionality statistics did not influence the achievable performance. The relative performance of two algorithms was similar in different databases. Over the lifetime of a product there is a reduction in precision for any method.
Conclusions In designing signal detection systems, careful consideration should be given to the criteria that are used to define an SDR. The choice of disproportionality statistic does not appreciably affect the achievable range of signal detection performance and so this can primarily be based on ease of implementation, interpretation and minimisation of computing resources. The changes in sensitivity and precision obtainable by replacing one algorithm with another are predictable. However, the absolute performance of a method is specific to the database and is best assessed directly on that database. New methods may be required to gain appreciable improvements.

Key Points

The performance of a range of signal detection algorithms has been compared within and across seven spontaneous reporting databases.

No method of signal detection was found to be uniformly better than the others, but each could be modified to improve the effectiveness, within limits, by choice of suitable signalling criteria.

Findings in one database will not necessarily generalize to another and some direct assessment of performance is desirable at a local level.

✉ Jim Slattery
Jim.Slattery@ema.europa.eu;
http://www.ema.europa.eu

[1] European Medicines Agency, 7 Westferry Circus, Canary Wharf, London E14 4HB, UK

[2] Uppsala Monitoring Centre, Uppsala, Sweden

[3] Bayer Pharma AG, Berlin, Germany

[4] Roche, Basel, switzerland

[5] GlaxoSmithKline, London, UK

[6] UK Medicines and Healthcare Products Regulatory Agency, London, UK

[7] AstraZeneca, Alderley Park, UK

1 Introduction

The widespread application of statistical approaches to detecting safety signals in databases containing spontaneously reported adverse drug reactions (ADRs) from marketed use of pharmaceutical products emerged in the late 1990s [1–3]. Since then, manufacturers of medicines and vaccines, regulatory agencies and independent drug safety monitoring organisations have adopted various signal detection algorithms (SDAs) with the goal of improving the sensitivity, objectivity and timeliness of their signal detection processes [4–9]. Furthermore, with organisations receiving increasing volumes of ADR reports year on year, there is a need to appropriately prioritise the work of safety departments, and statistical approaches have the potential to support this objective when used with specified criteria to trigger further evaluation of potential safety concerns [8, 10, 11].

The foundation of the majority of statistical signal detection methods is a 2 × 2 contingency table that relates the observed count for an adverse event of interest and a drug of interest with all other events and drugs in the database that together constitute a background from which an expected count is derived: the principal difference being the method by which the expected value is calculated [12]. Given this basic similarity between the statistics, the relative performance of SDAs in generating what have been termed signals[1] of disproportionate reporting (SDRs) [13, 14] is likely to be strongly driven by the choice of signal threshold [15–17] and the nature of the database background [12, 18–20]. There is a substantial corpus of published studies in which the relative performance of various SDA methods have been examined [11, 21–28]; however, these studies have invariably been limited in the number and diversity of algorithms included and the generalisability of the findings to other spontaneous adverse event databases [11, 23, 24, 29, 30]. To complicate matters further, the availability of a robust reference dataset of ‘true’ ADRs against which the sensitivity of SDAs can be assessed is essential in establishing the overall utility and benefit versus burden of implementing these SDAs into routine pharmacovigilance processes [31].

The aim of this study was to test a number of current SDAs within a number of different databases to establish whether the signal detection they provided was truly different and, if so, if any method was superior. The reason for replicating the work in a number of different datasets was twofold: firstly, to see whether comparisons of methods within one database would be replicated within others—possibly of very different size and with very different sets of products—and, secondly, to see if detection of signals for the same product using the same method might vary between the databases. The latter question will be addressed in a subsequent paper.

The study was carried out as part of the Innovative Medicines Initiative Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (IMI PROTECT) project, a multinational consortium of 35 partners including academics, regulators, small and medium-sized enterprises (SMEs) and European Federation of Pharmaceutical Industries and Associations (EFPIA) companies carrying out a programme of research to address limitations of current methods in the field of pharmacoepidemiology and pharmacovigilance (http://www.imi-protect.eu/). The partners contributing data and analyses to this specific project are the WHO Uppsala Monitoring Centre (UMC), the European Medicines Agency (EMA), the UK Medicines and Healthcare products Regulatory Agency (MHRA), AstraZeneca, Bayer Healthcare, GlaxoSmithKline and Roche.

2 Methods

2.1 Outcome Measures for Quantitative Signal Detection

A good signal detection method would detect ADRs at the earliest opportunity and produce a low number of false positives. Although the timing element is very important to pharmacovigilance, for this direct comparison of methods we have concentrated primarily on the simpler measures of whether a true SDR is ever produced for a given ADR and how many false positives are produced over the same time period. A drug–event combination (DEC) is considered to have a statistical signal if a SDR occurs at any point of time within the study period. Therefore, an SDR does not have to be manifested at the end date. This definition is adopted to simulate a prospective signal detection system in which SDRs would be investigated as they arose.

The summary measures calculated for each quantitative signal detection method are the sensitivity (the proportion of known ADRs that signalled) and the precision (the proportion of SDRs that correspond to known ADRs, also called positive predictive value). Summary information on timing is reported as average delays between product authorisation and signalling.

Moreover, since quantitative signal detection performance at a point in time allows comparison of methods but does not characterise any temporal characteristics in their performance, change in precision over time on the market for a product was measured. This is simply the proportion

[1] To avoid confusion it should be noted that an SDR does not necessarily fulfil the requirements of a signal as defined in pharmacovigilance.

of new SDRs raised in each 6-month time window that are true positives.

2.2 Statistical Methods

A challenge in this study was to standardise analyses across datasets with different characteristics. All the databases code ADRs with MedDRA®[2] but each has a different drug coding dictionary. Hence, an initial step was to construct a map of all of the products in the study. Mapping of simpler variables and comparable outputs from the analyses were ensured by the use of identical macros within each database written in SAS® code (SAS® v. 9.3, SAS Institute, Cary, NC, USA) at the EMA. Outputs from these macros were returned to the EMA for quality checking, collation and comparison.

The SAS® macros first generated standard input data tables accommodating the different coding conventions used in the datasets and then calculated SDRs at monthly intervals as they evolved over calendar time in each database. The period covered by the analysis was from January 1995 to December 2011. A second macro applied the signalling criteria specific to each SDA and hence determined the earliest point at which an SDR arose for each DEC within each database. Lastly, a third macro summarised and collated the results from the different databases and algorithms.

Outcome measures were calculated for all SDAs in each database. Since databases vary in size, age and the type of products for which they receive reports, variation in outcome may be attributable to the SDA or to the characteristics of the database. The effects of the SDA and database were separated using a simple analysis of variance (ANOVA) model applied to the logarithmic transformed sensitivity and precision. This model was used to allow the effect of the individual SDAs to be estimated averaged over databases and not to make statistical inferences concerning them.

2.3 Reference Standard

In order to classify a signal as true positive or a false positive/unknown, a reference standard is required. In an initial attempt to define known ADRs the EMA database that maps section 4.8 of the summary of product characteristics (SPC) for centrally authorised products (CAPs) to MedDRA® preferred terms was used. For products that were not CAPs, a dedicated mapping exercise was carried out. For products in company databases, additional terms from the company reference safety information were used as these should reflect the most current knowledge. For the purpose of the comparisons in this study all SDRs that did not highlight a term in the reference database were defined as false positives.

For the CAPs, it was possible to identify which ADRs had been added to the SPC after the marketing authorisation was granted and a field with this information was added to the reference dataset and used to check the robustness of study conclusions.

2.4 Choice of Products for Study

For this project the initial list of products was based on those previously used in the proportional reporting ratio (PRR) validation study carried out at the EMA [31]. In order to have sufficient numbers of products for analysis in company databases some additional products were included, this decision being made without reference to the known product ADRs. Recently launched products with limited post-marketing exposure, and those that were subject to regulatory, contractual or legal consideration were omitted. The final selection included 220 products, and these, together with the reference set of ADRs, can be accessed online [32].

2.5 Databases Included

The study analysis was carried out separately in a number of different databases of safety reports. Although these databases collect a range of report types, this study analysis was restricted to those reports routinely used at the collaborating organisations. These analyses were executed at the owner sites and only computer code, the logs and summarised results generated by the code were shared between sites. Table 1 summarises some of the main features of the databases and the reports used in this analysis. These databases were as follows:

1. EudraVigilance: Set up by the EMA in 2001, this database contains individual case safety reports (ICSRs) from pharmaceutical companies and European regulatory agencies. It collects post-authorisation and clinical trial reports on all products marketed in the EU and is retrospectively populated to 1995. Medicines are coded against a custom-built dictionary.
2. VigiBase®: The WHO global ICSRs database. It is the repository of the WHO Programme for International Drug Monitoring with more than a 100 member countries across the world. VigiBase® is maintained

[2] MedDRA® (the Medical Dictionary for Regulatory Activities) terminology is the international medical terminology developed under the auspices of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). MedDRA® trademark is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA) on behalf of ICH.

Table 1 Databases in the study

Organisation (abbreviation)	Affiliation (scope)	Database name (millions of spontaneous reports)	Reports used in signal detection	Number of products included in study
European Medicines Agency (EMA)	Competent authority (pan-European)	EudraVigilance (2.4)	Spontaneous reports only	220
Medicines and Healthcare products Regulatory Agency (MHRA)	Competent authority (UK)	Sentinel (0.6)	Spontaneous reports only	207
Uppsala Monitoring Centre (UMC)	Drug safety monitoring centre (international)	VigiBase® (7.0)	All reports	220
Bayer Healthcare Products (Bayer)	Pharmaceutical company (international)	ARGUS (0.6)	Spontaneous and literature	6
GlaxoSmithKline (GSK)	Pharmaceutical company (international)	OCEANS (1.4)	Spontaneous and literature	21
Roche	Pharmaceutical company (international)	ARISg™ (1.1)	All reports	15
AstraZeneca (AZ)	Pharmaceutical company (international)	Sapphire (0.5)	Spontaneous and literature	11

and analysed by the Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring. Reports in VigiBase® are encoded in the WHO Drug Dictionary as well as in MedDRA® and WHO-ART (WHO Adverse Reaction Terminology) (in parallel).

3. Sentinel: The UK MHRA spontaneous ADR database. It contains reports from healthcare professionals and patients (the latter from 2005). Drugs are encoded using an in-house drug dictionary. The MHRA's Sentinel database was set up in 2006 but also includes data from legacy databases with spontaneous case reports dating back to 1964.
4. Sapphire: AstraZeneca's worldwide safety database containing spontaneous adverse event reports from worldwide sources for AstraZeneca's marketed products, as well as serious adverse event reports from clinical trials of AstraZeneca compounds in development. Drugs are mapped to an in-house drug dictionary. Sapphire was implemented in 2008 but also includes data from legacy databases with spontaneous case reports dating back to the 1960s.
5. ARGUS: Bayer's worldwide safety database. It contains spontaneous adverse event reports from worldwide sources for Bayer's Pharmaceutical and Consumer Care marketed products, as well as serious adverse event reports from clinical trials of Bayer's compounds in development. Drugs are mapped to an in-house drug dictionary. ARGUS was set up in 2001 but also includes data from legacy databases with spontaneous case reports dating back to the 1960s.
6. OCEANS: GlaxoSmithKline's worldwide safety database. OCEANS contains spontaneous adverse event reports from worldwide sources for GlaxoSmithKline's marketed products, as well as serious adverse event reports from clinical trials of GlaxoSmithKline compounds in development. Drugs are mapped to an in-house drug dictionary. OCEANS was set up in 1999 but includes data from legacy databases with case reports dating back to the 1960s.
7. ARISg™: Roche's worldwide safety database. ARISg™ contains spontaneous adverse event reports from worldwide sources for Roche's marketed products, as well as serious adverse event reports from clinical trials of Roche's compounds in development. Drugs are mapped to an in-house drug dictionary. ARISg™ was set up in September 2011 but contains reports from legacy databases dating back to the late 1950s.

2.6 Measures of Disproportionality and Definitions of Signal of Disproportionate Reporting

A number of different statistical SDAs were tested. Most are in current use within pharmacovigilance departments with spontaneous reporting datasets. These include algorithms based on the reporting odds ratio (ROR), PRR, the information component and the Empirical Bayes Geometric Mean (EBGM). Another method, the Urn model, based on Fisher's exact test was included as it is a simple algorithm that has been advocated in the literature [16]. A review of these methods was published by Clark et al. [33]. The common feature of these algorithms is that the calculated measure is a comparison of observed counts of drug–adverse event pairs and the expected number based on other drugs in the database [6]. To fully describe an algorithm, it is necessary to specify not only the measure but also the conditions that must be met by the measure to indicate a positive signal.

Table 2 Signal detection methods in the study

Statistical method	Signal detection algorithm	Current use	Conditions for SDR
EBGM	EB05 (1.8, 3, 2.5)	MHRA	EB05 $\geq$1.8 and $n \geq 3$ and EBGM $\geq$2.5
	EB05 (1.8, trend)	AZ	EB05 $\geq$1.8 or positive trend flag[a]
	EB05 (2.0, trend)	GSK	EB05 >2.0 or positive trend flag[b]
IC	IC	UMC	IC lower bound 95 % CI >0
PRR	PRR025 (1.0, 3)	EMA	PRR lower bound 95 % CI $\geq$1 and $n \geq 3$
	PRR025 (1.0, 5)	EMA	PRR lower bound 95 % CI $\geq$1 and $n \geq 5$
	PRR (3.0, 3, 4.0)	No	PRR $\geq$3 and $\chi^2 \geq 4$ and $n \geq 3$[c]
	PRR (2.0, 3, 4.0)	Bayer	PRR $\geq$2 and $\chi^2 \geq 4$ and $n \geq 3$[d]
	PRR (2.0, 3, 3.84)	Roche	PRR $\geq$2 and p(χ^2) $\leq$ 0.05 and $n \geq 3$
ROR	ROR025 (1.0, SHR)	UMC	ROR with shrinkage, lower bound 95 % CI >1[e]
	ROR025 (2.0, 5)	MEB	ROR lower bound 95 % CI >2 and $n \geq 5$
	ROR025 (1.0, 3)	No	ROR lower bound 95 % CI $\geq$1 and $n \geq 3$
	ROR025 (1.0, 5)	No	ROR lower bound 95 % CI $\geq$1 and $n \geq 5$
Urn	Urn1	No	Reporting ratio >1 and unexpectedness >1/0.05
	Urn500	No	Reporting ratio >1 and unexpectedness >500/0.05

χ^2 Chi-squared, *AZ* AstraZeneca, *EB* Empirical Bayes, *EBGM* Empirical Bayes Geometric Mean, *EMA* European Medicines Agency, *GSK* GlaxoSmithKline, *IC* information component, *MED* Medicines Evaluation Board, *MRHA* Medicines and Healthcare products Regulatory Agency, *PRR* proportional reporting ratio, *ROR* reporting odds ratio, *SDR* signal of disproportionate reporting, *SHR* shrinkage has been applied to the estimator, *UMC* Uppsala Monitoring Centre

[a] AZ: increasing trend defined as (1) current EB05 is >EB95 52 weeks ago; or (2) a 50 % increase in EBGM compared with the EBGM 26 weeks ago

[b] GSK: increasing trend defined as either (1) the current EBGM has increased by $\geq$50 % compared with the EBGM value 6 months ago and current EB05 is >EB95 value 6 months ago; or (2) the current EBGM value has increased by $\geq$50 % compared with the EBGM value 6 months ago and current EB05 is at least 1.5

[c] Historic use at MHRA

[d] Bayer: this same definition was used at AZ until 2009

[e] UMC: treats the odds ratio as an observed-to-expected ratio and applies statistical shrinkage similar to that for the IC. For details see Norén et al. [34]

The SDAs that were compared are shown in Table 2. In general, the methods are those used by the project partners and so the conditions for a signal typify those in current practice. However, since the Urn method is not used by any partner, two implementations have been chosen based on a study by Hochberg et al. [16] and, since the ROR is used only by UMC in its shrinkage version, the implementation as used by The Netherlands Medicines Evaluation Board has been added. Moreover, calculations for the ROR using the same signal condition as some PRR implementations have been performed to test whether these methods provide the same results. The implementation of EBGM uses the Multi-item Gamma Poisson Shrinker based on DuMouchel and Pregibon [35].

Finally, we also calculated the entire envelope of achievable performance for SDAs based on the PRR with various values of thresholds on two parameters: the lower end of the 95 % confidence interval and the number of reports related to the DEC.

3 Results

3.1 Overall Comparison Within and Between Databases

The performance of each of the 15 SDAs within each of the seven databases is presented in Fig. 1. The sensitivity and precision in this figure are calculated at the study end date of December 2011. The algorithms are each plotted with a different symbol and the individual datasets are labelled using different colours. The spread of results across databases can be seen from the diversity of results for each algorithm.

As every algorithm was used in each dataset, the differences in performance between datasets can be estimated simply by averaging over the algorithms. Figure 2 shows the performance achieved in the different datasets.

The effect of each algorithm can be much more clearly seen by removing the component of variation attributable to the databases. This is done by subtracting the

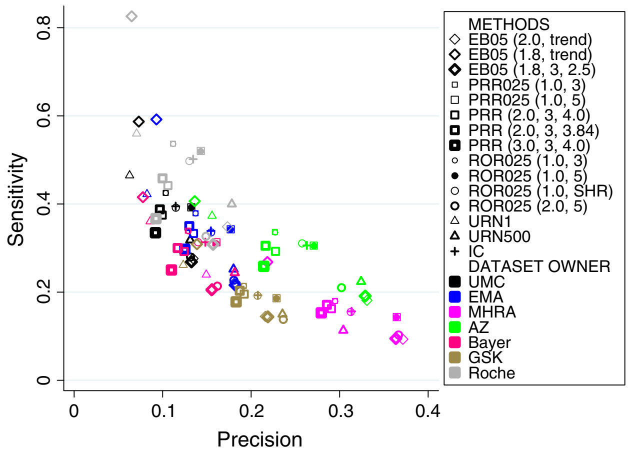

Fig. 1 Sensitivity and precision for each method in each database. *AZ* AstraZeneca, *EB* Empirical Bayes, *EMA* European Medicines Agency, *GSK* GlaxoSmithKline, *IC* information component, *MHRA* Medicines and Healthcare products Regulatory Agency, *PRR* proportional reporting ratio, *ROR* reporting odds ratio, *UMC* Uppsala Monitoring Centre

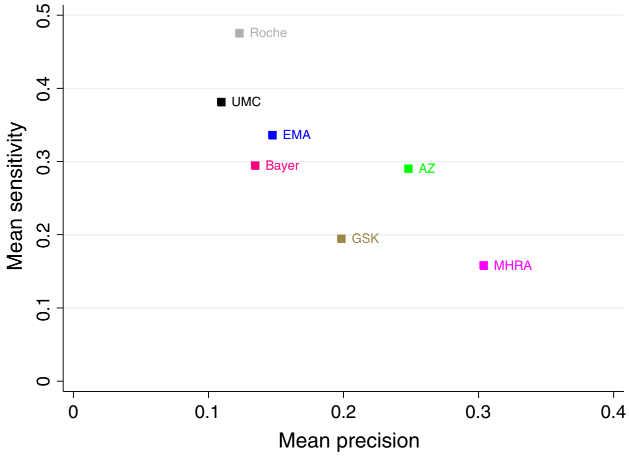

Fig. 2 Performance within each database averaged over signal detection algorithm methods. *AZ* AstraZeneca, *EMA* European Medicines Agency, *GSK* GlaxoSmithKline, *MHRA* Medicines and Healthcare products Regulatory Agency, *UMC* Uppsala Monitoring Centre

contribution attributed to the database in the ANOVA model. Figure 3 shows the results of this calculation. The grouping of results in this figure reflects the fact that comparative figures for sensitivity and precision of methods are similar in all databases and Fig. 4 and Table 3 give the figures averaged over databases to facilitate comparisons.

The observed differences in performance between the databases when using the same SDA may be attributable to a number of different causes. A potential cause is the different background distribution of adverse events. Another potential cause that was investigated was that the set of

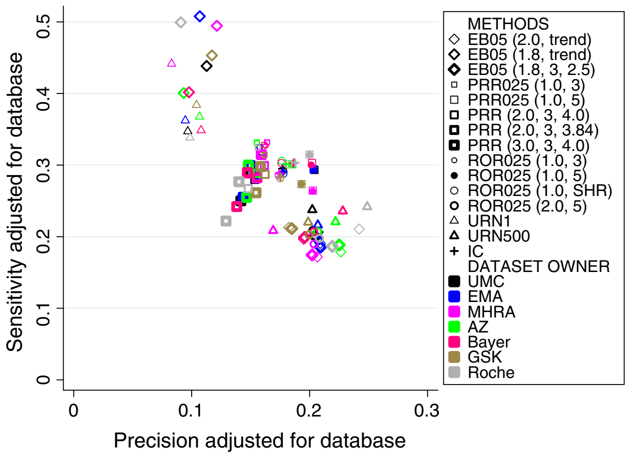

Fig. 3 Precision and sensitivity adjusted for database effects. *AZ* AstraZeneca, *EB* Empirical Bayes, *EMA* European Medicines Agency, *GSK* GlaxoSmithKline, *IC* information component, *MHRA* Medicines and Healthcare products Regulatory Agency, *PRR* proportional reporting ratio, *ROR* reporting odds ratio, *UMC* Uppsala Monitoring Centre

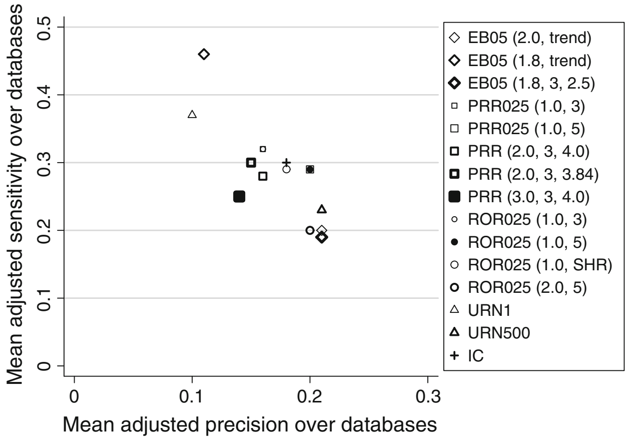

Fig. 4 Mean precision and sensitivity over databases. Note that PRR (1.0, 3) and PRR (1.0, 5) give identical performance to ROR (1.0, 3) and ROR (1.0, 5). *EB* Empirical Bayes, *IC* information component, *PRR* proportional reporting ratio, *ROR* reporting odds ratio

study products represented in each dataset varied; the larger datasets such as EudraVigilance, VigiBase® and Sentinel contain some records for each product, but the industry datasets concentrated on company products. Figure 5 shows the performance measures for EudraVigilance and VigiBase® for all study products and also restricted to products from the four pharmaceutical companies and this is compared with the performance in the company database.

For analyses restricted to AstraZeneca, Bayer and Roche products it is clear that company results resemble those from EudraVigilance and VigiBase®. An anomaly is seen with VigiBase® results for GlaxoSmithKline products,

Table 3 Adjusted sensitivity and precision for all methods

Signal detection algorithm	Sensitivity	Precision
EB05 (1.8, 3, 2.5)	0.19	0.21
EB05 (1.8, trend)	0.46	0.11
EB05 (2.0, trend)	0.20	0.21
IC	0.30	0.18
PRR025 (1.0, 3)	0.32	0.16
PRR025 (1.0, 5)	0.29	0.20
PRR (3.0, 3, 4.0)	0.25	0.14
PRR (2.0, 3, 4.0)	0.28	0.16
PRR (2.0, 3, 3.84)	0.30	0.15
ROR025 (1.0, SHR)	0.29	0.18
ROR025 (2.0, 5)	0.20	0.20
ROR025 (1.0, 3)	0.32	0.16
ROR025 (1.0, 5)	0.29	0.20
Urn1	0.37	0.10
Urn500	0.23	0.21

EB Empirical Bayes, *IC* information component, *PRR* proportional reporting ratio, *ROR* reporting odds ratio, *SHR* shrinkage has been applied to the estimator

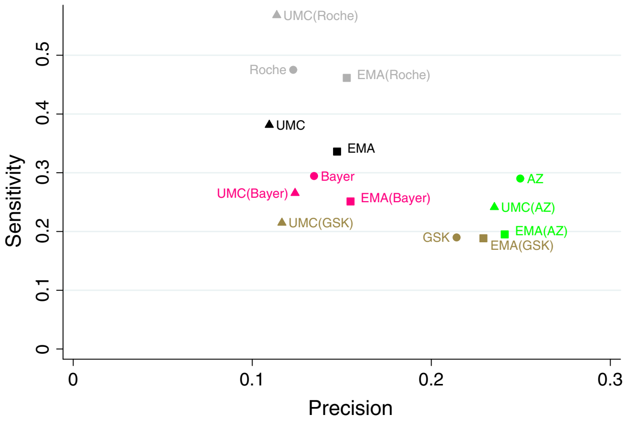

Fig. 5 Differences in performance restricted to common set of products. *AZ* AstraZeneca, *EMA* European Medicines Agency, *GSK* GlaxoSmithKline, *MHRA* Medicines and Healthcare products Regulatory Agency, *UMC* Uppsala Monitoring Centre

which show a lower precision than either GlaxoSmithKline or EudraVigilance results. Closer examination shows that this might be driven by results for vaccines, which account for 12 of the 21 GlaxoSmithKline products in the study; further investigation is ongoing.

3.2 Achievable Performance Figures

The variation in performance that can be obtained with any disproportionality statistic between standard SDAs suggests that a more systematic exploration of the possible thresholds might be helpful. As an example, Fig. 6 shows the range of performance that can be achieved using the PRR025 method in EudraVigilance over a range of thresholds. The PRR was selected for this exercise as the SDA is characterised by just two thresholds and it is a simple and fast calculation. ROR could also have been used as it has similar characteristics. In fact, it gives identical performance to the PRR.

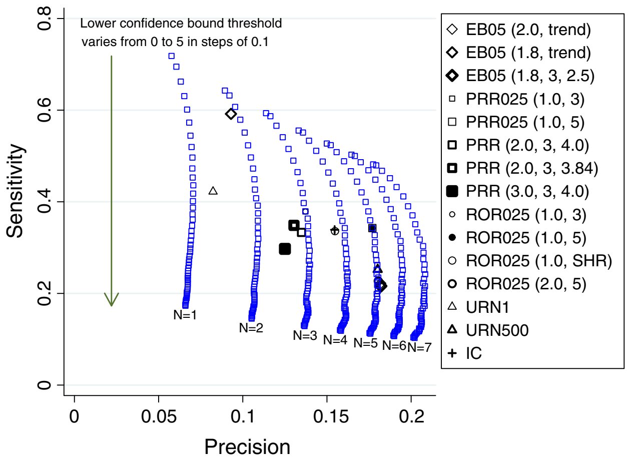

Fig. 6 Envelope of sensitivity and precision achievable with the proportional reporting ratio in EudraVigilance. *EB* Empirical Bayes, *IC* information component, *PRR* proportional reporting ratio, *ROR* reporting odds ratio

The performance of all of the methods considered in this study calculated in EudraVigilance are also shown in Fig. 6, and it is clear that they can be approximated by appropriate choice of thresholds for the PRR. Potential improvements in sensitivity and precision compared with the current PRR thresholds may also be possible but, as discussed in Slattery et al. [17], increases in the threshold for *N* involved some delay in signalling. The average number of months from first report received for a product to the SDR has been calculated and the comparison between the tested methods and the PRR025 envelope methods with the nearest performance is shown in Fig. 7. As some products predated the period covered by the analysis, this calculation excluded products with reports received earlier than January 1995. When the performance of an SDA tested corresponded exactly to one set of thresholds for PRR025, only those values were used, but when, more commonly, the performance fell between two thresholds of n, a weighted average of times is given for the PRR025 methods.

In general, these times are very similar. However, some increase in time to signalling may be incurred by using the Urn500 and the EB05 (1.8, trend) method, relative to the equivalent PRR025 algorithm. The latter does not appear to be inherent in the EBGM, but may be related to the use of trend thresholds in the signalling algorithm.

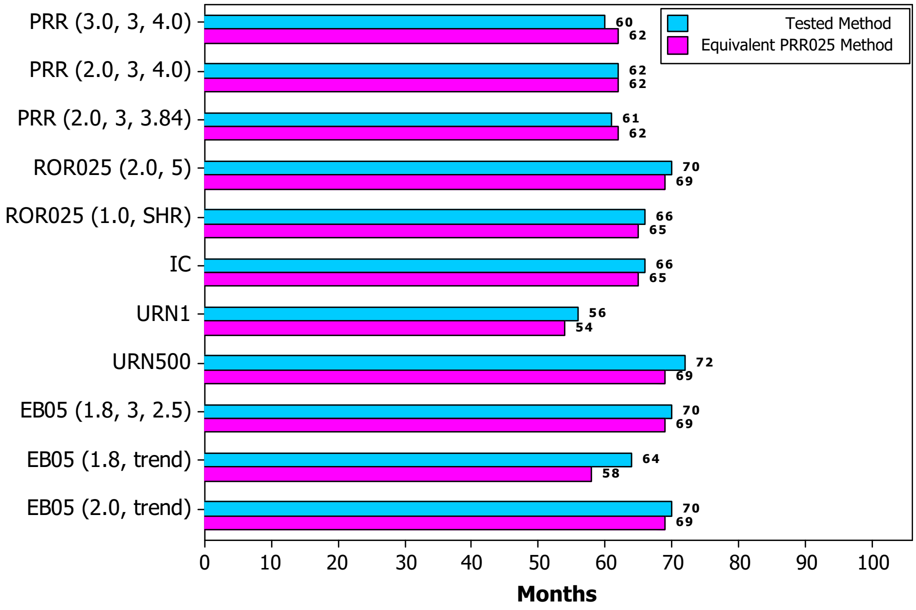

Fig. 7 Average time to detect adverse drug reaction. *EB* Empirical Bayes, *IC* information component, *PRR* proportional reporting ratio, *ROR* reporting odds ratio

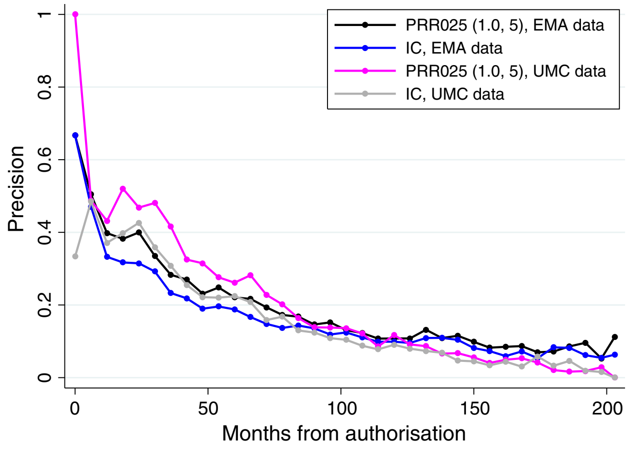

Fig. 8 Change in precision with time from authorisation. *EMA* European Medicines Agency, *IC* information component, *PRR* proportional reporting ratio, *UMC* Uppsala Monitoring Centre

3.3 Changes in Precision with Time on Market

The precision (or positive predictive value) is the proportion of SDRs that turn out to be ADRs. This can be calculated over all SDRs or over subsets defined by attributes of the product or SDR. One attribute of interest is the timing of SDRs in the lifetime of the product. To investigate this, the precision was calculated for SDRs occurring in 6-month windows starting from an origin at the time of product authorisation in the EU. As shown in Fig. 8, a fall in precision over time on market was observed. In case this fall in precision was influenced by early signalling of effects that were already known, the calculation was repeated excluding the ADRs identified prior to authorisation. This did not change the form of the curve.

4 Discussion

Interpretation of the sensitivity and precision values from the different SDAs shown in Fig. 1 is not entirely straightforward. It is clear that variation in these figures may be due to a number of causes, but that the largest alterations in performance are produced by changes in the thresholds that define an SDR rather than by the choice of the disproportionality statistic. The distribution of results in Fig. 1 reveals a roughly linear trade-off between sensitivity and precision, but it is interesting that figures based on the EBGM lie at both extremes of the distribution. The EB05 (1.8, trend) produces a higher sensitivity and a low precision, whilst the EB05 (2.0, trend) and EB05 (1.8, 3, 2.5) algorithms do the opposite. It is likely that performance levels close to any point in this distribution could be achieved by any disproportionality statistic by suitable choice of thresholds (as shown by Fig. 6 for the PRR025 model). The more desirable performance levels are those points falling towards the upper-right side of the scatter of points but the best trade-off of sensitivity and precision will depend on other factors, e.g. the resources available for further evaluation of SDRs.

Sensitivity and precision also vary in a systematic way with the database in which the SDAs are used. This variation was seen in Fig. 2, and the calculations presented in Fig. 4 reveal that a large part of that variation is attributable to the particular set of products on which the performance is evaluated. This is an important result as identification of the sources of variation may eventually allow better prediction of performance in new databases. It is notable that the set of products providing the denominator of the disproportionality statistics remains the entire set contained in each database. Hence, the effect of different background distributions of adverse events from very limited product sets, such as might be found in small company databases, has not been investigated. This would require more extensive identification of the other products in the databases and is beyond the scope of this study.

An interesting finding is that the standard PRR-based algorithms appear to lose performance when a threshold for absolute value of PRR is introduced. The PRR025 (1.0, 3) and PRR025 (1.0, 5), with no such threshold, appear to work slightly better that the PRR (2.0, 3, 4.0) and PRR (2.0, 3, 3.84), with a threshold of 2, which themselves outperform the PRR (3.0, 3, 4.0), with a threshold of 3. Figure 4, with the effect due to the database removed, shows this much more clearly. A further analysis, not reported here, shows the same effect for SDAs based on the EBGM. Thus, although the results for different thresholds generally shift the performance along the line, there are also indications that some sets of rules may dominate others in that they move the performance towards the upper side of the distribution. This suggests the choice of algorithm is not purely concerned with selecting the right trade-off of sensitivity and precision. Care must also be taken not to produce a general decrease in performance.

Plots of precision and sensitivity versus the time (calculated as the average number of months from first report received for a product to the generation of a SDR) show an interesting relationship: on average, SDAs with higher precision take longer to produce a true SDR while SDAs with higher sensitivity are faster in identifying a true SDR. This seems to suggest that the trade-off between sensitivity and precision involves timing as well: the more precise a method is, the lower the sensitivity and the slower the production of true SDRs; the less precise a method is, the greater the sensitivity and the faster the production of true SDRs.

The substantively different average results for different databases shown in Fig. 2 demonstrate a difficulty in selecting optimal methods. The results of testing on one database will not necessarily generalise to another. Thus, it is likely to be necessary to test methods specifically for each database. A useful finding, however, is that an algorithm with relatively (compared to the other algorithms) good performance in one database seems to have relatively good performance in other databases. The clustering of results from different databases for each algorithm seen in Fig. 3 illustrates this point. Hence, the most promising algorithms can be selected for testing on the basis of prior evaluations.

The sources for the variation in performance across databases are probably multifactorial. The effect of different types of product appear to be one important factor but, in this study, local practice was followed at each centre concerning which types of report to include in the calculations and whether only reports where causality was suspected were included. Standardisation of these practices might reduce variation and allow performance results from one database to generalize to others. However, this requires further investigation.

The apparent reduction in precision over time shown in Fig. 8 was mirrored in all methods in all databases. The two databases and methods shown are simply illustrative. The reasons for this reduction are not clear but several possibilities exist. Initially it was noted that the reference set of ADRs used in the study included those identified prior to authorisation. It might be that these are preferentially reported immediately after authorisation, artificially inflating the precision. However, a further analysis including only ADRs identified post-authorisation showed the same reduction in precision with time. It is also possible that the later SDRs only appear to be false positives either because the nature of the adverse event makes the ADR difficult to validate or because insufficient time has elapsed to validate them. To investigate this latter idea we restricted our calculations so that products were excluded 2 years before the end of follow-up. Thus, at least 2 years were available for validation of any signal. This did not change the reduction in precision and thus this explanation seems unlikely. It is even possible that the reduction may not be related to changes in information on particular drugs, but could be a function of the total number of reports in the database. This would be consistent with the moderate overall precision observed for larger databases in our study. An alternative explanation for the decline in precision may be that drug labels preferentially include more widely used ADR terms, and that as the number of reports increases, statistical associations begin to emerge that involve more abstruse ADRs that are less likely to be listed on the SPCs; even if true causal effects, such ADRs may be considered to be covered by broader related ADRs for labelling purposes. Perhaps the simplest explanation is that the majority of safety problems with a product are found in the early years and thus there is less left to find as time progresses. Whatever the explanation, it appears clear that there is some genuine drop in return from statistical signal detection with time on the market and the implications of this for the use of these methods requires evaluation.

Limitations of this work include that overall counts of true positives do not entirely capture the relative benefits of a statistical signal detection process. It is also reasonable to ask whether two methods capture the same adverse reactions and at the same time. Such a calculation is being considered using these study data and will be reported at a later stage.

Lack of a gold standard is a problem in any assessment of signal detection methods. In this study, SDRs that do not correspond to known ADRs are classified as false positives although they may later turn out to be true. This strategy will overestimate the number of false positives but should not undermine the comparison of methods since all methods are compared under the same reference standard. In terms of generalisation to routine pharmacovigilance, the use of all known ADRs creates difficulties. It is better to use the approach of Alvarez et al. [31] and restrict to ADRs that were unknown at the time of the study, but this requires information that was not available to us for the non-CAP products and hence could not be done here. It is possible that knowledge of an ADR may change the reporting practice and alter the chance of an SDR arising. To investigate this we ran a sensitivity analysis that restricted the reference dataset to ADRs identified in the post-authorisation phase, and this did not substantively change the findings of this study.

Although the study included a range of databases, these were generally of moderate or large size. It is possible that databases with very few products or reports might show different properties. It should also be borne in mind, when interpreting these results for a practical pharmacovigilance system, that such a system will include other signal detection processes that may differ between organisations. Hence, the statistical signal detection is an element, but not the sole determinant, of the system.

5 Conclusions

The choice of signalling criteria to define an SDR has a greater impact on signal detection performance in terms of sensitivity, precision and time to signal than the choice of disproportionality methods. Hence, signalling criteria must be selected carefully on the basis of empirical evidence, and choice of a disproportionality measure for signal detection can primarily be based on ease of implementation, interpretation and minimisation of resources.

The changes in sensitivity and precision obtainable by replacing one SDA with another are predictable in the moderately large databases studied. However, the absolute performance of a method is specific to the database and is best assessed directly on that database.

There appears to be a reduction in precision with time and, hence, it may be more productive to place additional effort into evaluation of signals from newer products.

The limits of performance of the current disproportionality statistics are similar and new methods, involving substantially different approaches, may be required to gain appreciable improvements in quantitative signal detection. In performing this study, we have provided a framework that allows new signal detection methods to be directly and easily compared with existing methods, thus providing a benchmark for evaluation.

Acknowledgments We would like to thank Jenny Wong, Jeffery Painter, Ramin Arani, Niklas Noren, Phil Tregunno, Michael Kayser and Magnus Lerch for their part in the study and constructive comments. The views expressed in this paper are those of the authors only and do not reflect the official policy or position of the IMI JU (Innovative Medicines Initiative Joint Undertaking), the European Union, European Federation of Pharmaceutical Industries and Associations (EFPIA) or the Medicines and Healthcare products Regulatory Agency.

Conflicts of interest The PROTECT project has received support from the Innovative Medicine Initiative Joint Undertaking (http://www.imi.europa.eu) under Grant Agreement No. 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution. Gianmario Candore, Kristina Juhlin, Katrin Manlik, Bharat Thakrar, Naashika Quarcoo, Suzie Seabroke, Antoni Wisniewski and Jim Slattery have no financial interest in any commercial signal detection software. Katrin Manlik is an employee of Bayer Pharma AG, Bharat Thakrar is an employee of Roche and holds shares in both Roche and GlaxoSmithKline, Naashika Quarcoo is an employee of and holds shares in GlaxoSmithKline, and Antoni Wisniewski is an employee of and holds shares in AstraZeneca. Products from these companies were among those used to test the methodologies in this research.

References

1. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54(4):315–21.
2. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10(6):483–6.
3. Lindquist M, Stahl M, Bate A, Edwards IR, Meyboom RHB. 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.
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. Van Puijenbroek EP, Diemont WL, Van Grootheest K. Application of quantitative signal detection in the Dutch spontaneous reporting system for adverse drug reactions. Drug Saf. 2003;26(5):293–301.
6. Almenoff J, Tonning JM, Gould AL, Szarfman A, Hauben M, Ouellet-Hellstrom R, et al. Perspectives on the use of data mining in pharmacovigilance. Drug Saf. 2005;28(11):981–1007.
7. Roux E, Thiessard F, Fourrier A, Begaud B, Tubert-Bitter P. Evaluation of statistical association measures for the automatic

signal generation in pharmacovigilance. IEEE Trans Inf Technol Biomed. 2005;9(4):518–27.

8. Czarnecki A, Voss S. Safety signals using proportional reporting ratios from company and regulatory authority databases. Drug Inf J 2008. 2008;42(3):205–210.
9. Hauben M, Horn S, Reich L. Potential use of data-mining algorithms for the detection of 'surprise' adverse drug reactions. Drug Saf. 2007;30(2):143–55.
10. Bate A, Lindquist M, Edwards IR. The application of knowledge discovery in databases to post-marketing drug safety: example of the WHO database. Fundam Clin Pharmacol. 2008;22(2):127–40.
11. Banks D, Woo EJ, Burwen DR, Perucci P, Braun MM, Ball R. Comparing data mining methods on the VAERS database. Pharmacoepidemiol Drug Saf. 2005;14(9):601–9.
12. Gipson G. A shrinkage-based comparative assessment of observed-to-expected disproportionality measures. Pharmacoepidemiol Drug Saf. 2012;21(6):589–96.
13. Hauben M, Aronson JK. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. Drug Saf. 2009;32(2):99–110.
14. Finney DJ. Statistical logic in the monitoring of reactions to therapeutic drugs. Methods Inf Med. 1971;10(4):237–45.
15. Declerck G, Bousquet C, Jaulent MC. Automatic generation of MedDRA terms groupings using an ontology. Stud Health Technol Inform. 2012;180:73–7.
16. Hochberg AM, Hauben M, Pearson RK, OHara DJ, Reisinger SJ, Goldsmith DI, et al. An evaluation of three signal-detection algorithms using a highly inclusive reference event database. Drug Saf. 2009;32(6):509–25.
17. Slattery J, Alvarez Y, Hidalgo A. Choosing thresholds for statistical signal detection with the proportional reporting ratio. Drug Saf. 2013;36(8):687–92.
18. Gogolak VV. The effect of backgrounds in safety analysis: The impact of comparison cases on what you see. Pharmacoepidemiol Drug Saf. 2003;12(3):249–52.
19. Hammond IW, Gibbs TG, Seifert HA, Rich DS. Database size and power to detect safety signals in pharmacovigilance. Expert Opin Drug Saf. 2007;6(6):713–21.
20. Hammond IW, Rich DS, Gibbs TG. Effect of consumer reporting on signal detection: Using disproportionality analysis. Expert Opin Drug Saf. 2007;6(6):705–12.
21. Almenoff JS, LaCroix KK, Yuen NA, Fram D, DuMouchel W. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. Drug Saf. 2006;29(10):875–87.
22. Brown JS, Petronis K, Bate A, Zhang F, Dashevsky I, Kulldorff M, et al. Comparing two methods for detecting adverse event signals in observational data: empirical Bayes gamma poisson shrinker vs. tree-based scan statistic. Pharmacoepidemiol Drug Saf. 2011;20:S144.
23. Bunchuailua W, Zuckerman I, Kulsomboon V, Suwankesawong W, Singhasivanon P, Kaewkungwal J. A comparison of signal detection performance between reporting ODDS ratio and Bayesian confidence propagation neural network methods on adverse drug reaction spontaneous reporting database of the Thai FDA. Value Health. 2010;13(7):A508.
24. Chen Y, Guo JJ, Steinbuch M, Lin X, Buncher CR, Patel NC. Comparison of sensitivity and timing of early signal detection of four frequently used signal detection methods: An empirical study based on the US FDA adverse event reporting system database. Pharm Med. 2008;22(6):359–65.
25. Harpaz R, Dumouchel W, Lependu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clin Pharmacol Ther. 2013;93(6):539–46.
26. Van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection is spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11(1):3–10.
27. Caster O, Noren G, Niklas, Madigan D, Bate A. Large-scale regression-based pattern discovery: the example of screening the WHO global drug safety database. Stat Anal Data Min. 2010;3(4):197–208.
28. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med. 2012;4(125):125ra31.
29. Kubota K, Koide D, Hirai T. Comparison of data mining methodologies using Japanese spontaneous reports. Pharmacoepidemiol Drug Saf. 2004;13(6):387–94.
30. Kurz X, Slattery J, Addis A, Durand J, Segec A, Skibicka I, et al. The EudraVigilance database of spontaneous adverse reactions as a tool for H1N1 vaccine safety monitoring. Pharmacoepidemiol Drug Saf. 2010;19:S330–1.
31. Alvarez Y, Hidalgo A, Maignen F, Slattery J. Validation of statistical signal detection procedures in EudraVigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling. Drug Saf. 2010;33(6):475–87.
32. IMI PROTECT. ADR database. http://www.imi-protect.eu/methodsRep.shtml. Accessed 17 Mar 2014.
33. Clark JA, Klincewicz SL, Stang PE. Spontaneous adverse event signaling methods: classification and use with health care treatment products. Epidemiol Rev. 2001;23(2):191–210.
34. Norén GN, Hopstadius J, Bate A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. Stat Methods Med Res. 2013;22(1):57–69.
35. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco: ACM Press; 2001. pp. 67–76.