

Detecting signals of disproportionate reporting from Singapore's spontaneous adverse event reporting system – an application of the Sequential Probability Ratio Test

Cheng Leng Chan^{1, 2}, Sowmya Rudrappa^{1, 3}, Pei San Ang¹, Shu Chuen Li², Stephen JW Evans⁴

¹ Health Products Regulation Group, Health Sciences Authority, Singapore

² School of Biomedical Sciences & Pharmacy, The University of Newcastle, Callaghan, Australia

³ Genome Institute of Singapore (A-Star), Singapore

⁴ Department of Medical Statistics, London School of Hygiene and Tropical Medicine, University of London, London, United Kingdom

Correspondence:

Ms Cheng Leng Chan,
Group Director,
Health Products Regulation Group,
Health Sciences Authority,
11 Biopolis Way #11-01 Helios Singapore 138667
Email: chan_cheng_leng@hsa.gov.sg
Tel: +65-68663528
Fax: +65-64789032

Keywords:

Adverse drug reactions
Data mining
Drug safety
Pharmacovigilance
Sequential Probability Ratio Test
Signal detection

Key points:

1. Sequential Probability Ratio Tests (SPRT) may have a role in detecting signals from spontaneous reports of suspected adverse drug reactions. They have somewhat different properties to other commonly used statistical methods for that purpose.
2. Using a combination of variable hypothesised relative risks (hRRs) allows for the detection of different types of adverse events (AEs). For drug-AE pairs that are rare with low expected counts, we need to apply a higher hRR for the SPRT method to pick up signals of disproportionate reporting of concern. On the other hand, a lower hRR will be useful for the more common drug-AE pairs.

Detecting signals of disproportionate reporting from Singapore's spontaneous adverse event reporting system – an application of the Sequential Probability Ratio Test

Abstract

Introduction:

The ability to detect safety concerns from spontaneous adverse drug reaction reports in a timely and efficient manner remains important in public health.

Objective:

This paper explores the behaviour of the Sequential Probability Ratio Test (SPRT) and ability to detect signals of disproportionate reporting (SDRs) in the Singapore context.

Methods:

We used SPRT with a combination of two hypothesised relative risks (hRRs) of 2 and 4.1 to detect signals of both common and rare adverse events in our small database. We compared SPRT with other methods in terms of number of signals detected and whether labelled adverse drug reactions were detected or the reaction terms were considered serious. The other methods used were reporting odds ratio (ROR), Bayesian Confidence Propagation Neural Network (BCPNN) and Gamma Poisson Shrinker (GPS).

Results:

The SPRT produced 2187 signals in common with all methods, 268 unique signals, 70 signals in common with at least one other method, and did not produce signals in 178 cases where two other methods detected them, and 403 signals unique to one of the other methods. In terms of sensitivity, ROR performed better than other methods, but the SPRT method found more new signals. The performances of the methods were similar for negative predictive value and specificity.

Conclusions:

Using a combination of hRRs for SPRT could be a useful screening tool for regulatory agencies, and more detailed investigation of the medical utility of the system is merited.

(238 words)

Detecting signals of disproportionate reporting from Singapore's spontaneous adverse event reporting system – an application of the Sequential Probability Ratio Test

1. INTRODUCTION

Post-marketing surveillance of drugs and vaccines is important to minimise risks with marketed drugs. In practice, spontaneous suspected adverse drug reaction (ADR) reporting remains the main source of information for regulators in this monitoring [1]. Pharmacovigilance distinguishes true ADRs from “adverse events” (AEs) that are not caused by the drug in question. A signal is a potential safety concern that a drug may be associated with a previously unrecognised hazard requiring further investigation [2].

Since 1993, the Vigilance and Compliance Branch of the Singapore Health Sciences Authority (HSA) receives spontaneous local ADR reports from healthcare professionals (e.g. 83.2% from clinicians, 12.2% from pharmacists, 2.7% from other healthcare professionals and 1.9% from pharmaceutical companies via facsimile, mail or online). Its Spontaneous Reports System (SRS) database has, in recent years, also received reports from the public healthcare institutions in real time via the Critical Medical Information Store (CMIS), and as a result, there has been a major increase of 40-fold in the number of ADR reports received regularly [3]. The numbers of reports received annually is now about 20,000.

Statistical “data mining” methods emerged in the late 1990s to complement the traditional manual review, and these are commonly called ‘disproportionality analysis’ [4, 5]. Some examples of frequentist statistical methods are the proportional reporting ratio (PRR), reporting odds ratio (ROR) [6-8]. Examples of Bayesian methods are the Bayesian Confidence Propagation Neural Network (BCPNN) [7, 9], Gamma Poisson Shrinker (GPS) [10] and Multi-item Gamma Poisson Shrinker (MGPS). The MGPS is now used by the US Food and Drug Administration (FDA) and the UK Medicines and Healthcare products Regulatory Agency (MHRA) [11, 12]. All these methods are based around the ratio of observed-to-expected counts of reports to obtain signals, and many studies have shown that no single signal detection algorithm (SDA) provides uniformly better performance [13]. The commonly used methods do not allow for multiple looks at the accumulating data over time, which can result in large numbers of false positive findings [12].

One method, the Sequential Probability Ratio Test (SPRT), has less concern associated with multiple testing over time; it is specifically designed to make allowance for multiple looks at data over time.

SPRT was developed by Wald in the 1940s [14, 15] and has mainly been used in process monitoring. A review of the literature showed that the SPRT may offer advantages over the other methods to overcome multiple testing problems [16-18]. It has been used in the context of scanning electronic health records, but has not had extensive evaluation for spontaneous reports.

Specifically, SPRT compares two hypotheses based on the likelihood of observing the data given those hypotheses [17, 19]. Unlike the other methods, SPRT is based

on the difference between (rather than the ratio of) the observed and expected values. However, there is limited research conducted on SPRT to evaluate its performance in an SRS database similar to Singapore's, and a previous evaluation of SPRT used a single alternative hypothesis on simulated data [20]. On theoretical grounds, at least, the methods that do not allow for this form of multiple testing over time will have a higher rate of false positive findings – false signals.

This paper explores how SPRT behaves and reviews its utility and applicability to pick up signals of disproportionate reporting (SDRs) for potential drug safety signals in the Singapore context. We also compare the performance of SPRT with three other SDA, namely ROR, BCPNN and GPS.

2. METHODS

The spontaneous reporting database in Singapore has been described by Ang et al [21]. In the database, each valid report has at least one product and at least one suspected ADR term included. Products are coded using standardised drug names, and adverse reaction terms are coded using the World Health Organisation Adverse Reaction Terminology (WHO-ART) (version 151) [22].

The SPRT method requires that specific hypotheses regarding a relative risk to be detected are set out, and in the context of signals of ADRs, some arise from relatively frequently occurring AEs where small relative risks are nevertheless potentially important, while others are from rare events where only higher relative risks can be detected. The details of the SPRT method are described in the 'Appendix', together with a brief description of the other methods used here,

including the criteria that determine whether the counts for particular drug-AE pairs constitute a signal.

We analysed signals using all data from 1993 to 2013, and in additional analyses, also reviewed the data as they were up to 2011 and examined the new signals that arose in 2012 and 2013, mirroring what is done in practice with accumulating data. We also classified the drug-AE pairs in terms of seriousness of AE and whether the AE was labelled for that drug or not.

We evaluated the methods, as most other comparisons have done, using sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV), considering factors such as whether the pair is a significant SDR and whether the drug-AE pairs are labelled. The drug-AE pairs were reviewed by a senior pharmacist and considered as labelled if they were mentioned in the Micromedex® [23] or current regulatory agency-approved drug labels in Singapore. It could be exact word-to-word matching or synonymously labelled (same meaning). The AE terms were considered as serious if they exist in the WHO critical terms list or medically significant suspected serious ADRs by the US FDA or in the Important Medical Event Terms (IME) list developed by the EudraVigilance Expert Working Group [22, 24, 25]. All the analyses were performed using R software, version 3.3.1 [26], including a signal detection package PhViD [27].

3. RESULTS

The HSA received a total of 151,180 AE reports from 1993 to 2013; these reports involved 23,183 unique drug-AE pairs. There were 1569 different suspected drug

substances and 1014 different AE terms. If every drug had reports for every possible AE, then all possible combinations of drug (1569 drugs) with AE term (1014 terms) would mean that there were potentially 1,590,966 possible pairs (1569 x 1014). Of those possible combinations, only 23,183 (1.5%) unique pairs actually occurred. The counts in each of these 23,183 cells had each of the methods applied to them.

The effect of the two hypothesised relative risks (hRRs) used for detecting signals using SPRT with hRR=2 or hRR=4.1, is shown in Table 1 where we show the distribution of observed and expected counts where signals are detected, giving medians and 5th and 95th centiles. We analysed the SDRs in terms of seriousness of AE and whether the AE is labelled or not; for the period from 1993 to 2013, a total of 137 unique serious and non-labelled drug-AE pairs were signalled by SPRT. Of the 137 drug-AE pairs, 88 drug-AE pairs were not picked up by hRR=2 as the number of observed counts was less than five.

The ROR and SPRT methods detected more SDR compared to BCPNN and GPS. Fig. 1 is the Venn diagram that shows the overlap of significant pairs detected by each method for the complete data from 1993 to 2013. The SDRs detected by the BCPNN method are a subset of the other methods. It is clear from this that the large majority (70%) of signals were detected by all methods (2187/3106). SPRT detected 268 signals ($N < 3$) that were not detected by other methods, while ROR detected 400 signals that were not detected by the other methods.

Comparisons were done by reviewing the numbers of new signals based on drug-AE combinations that had some reports in the database prior to that year but were not signals previously, and totally new signals, where the combination had reports for the

first time in the relevant period. Table 2 gives the number of new significant SDRs for each method for different quarters from 2012 to 2013. In general, SPRT tends to generate a higher percentage of new significant pairs compared to the other methods.

To evaluate which methods performed better than the others, we reviewed the significant SDRs in terms of seriousness of AE, whether the AE is labelled or not labelled, PPV, NPV, sensitivity and specificity. In this analysis, the SPRT method detected more not labelled drug-AE pairs. In terms of PPV, ROR, BCPNN and GPS performed better than SPRT. In terms of sensitivity, ROR performed better than other methods. The performances of the methods were similar for NPV and specificity (see Table 3).

4. DISCUSSION

We have shown that the SPRT method has some different properties to the other methods and that it can be ‘tuned’ to detect signals for rare events as well as more frequent ones. The fact that the hRR has to be pre-specified, while apparently a disadvantage, can be utilised to obtain signals in different circumstances. This method may be suitable for databases with smaller total numbers of reports and where a signal would be detected even with smaller numbers, compared with databases containing many millions of reports, such as those of the FDA and the EU.

More work may need to be done to investigate its use in practice and whether it should be an additional or an alternative method for use in the context of smaller

databases. There is no general 'gold standard' to define which of the drug-AE pairs are really true ADRs and which are not. Methods like ROR and GPS that have been used in the past may have led to labelling, but it is not certain that all such associations are true ADRs. Individual regulatory authorities may need to examine the characteristics of the signals detected and not detected by the different methods in their own data.

Singapore, although having a high reporting rate based on number of AE reports received per million inhabitants, is a small country and, therefore, its total number of reports is not that high. In this situation, there are drug-AE pairs of interest with small numbers of reports, and detecting SDRs using SPRT only with a small value of hRR will be problematic.

For more rare events, it could be useful to adopt a higher hRR for early signalling purposes. For example, dabigatran (anticoagulant) and cerebral infarction was signalled earlier by hRR=4.1 when the number of observed counts was two. Using hRR=2 only gave a signal 9 months later when the number of observed counts reached five. ROR, BCPNN and GPS signalled it 3 months later than SPRT. Diltiazem (antihypertensive / anti-anginal) and vestibular disorders was signalled with hRR=4.1 in 2011 but not by hRR=2, ROR, BCPNN or GPS. Diltiazem and vestibular disorders is not included in the product label, but studies have shown migraine-related dizziness or vertigo have been reported in 7% of patients [28]. Letrozole (anticancer) and epidermal necrolysis was signalled with hRR=4.1, ROR, BCPNN and GPS in 2011, but not by hRR=2. Letrozole and epidermal necrolysis is included in the product labels as either uncommon or rare. Vancomycin (antibiotic) and acute generalised exanthematous pustulosis was signalled using hRR=4.1, ROR and GPS

in 2011, and 6 months later by BCPNN. These findings suggest that SPRT could have a useful role, but it is not clearly superior to other methods.

There are some signals probably resulting from confounding by indication, such as clozapine (antipsychotic) and neurosis, dapson-pyrimethamine (combination of antibiotic and antimalarial) and infection, dasatinib (anticancer) and malignant neoplasm, hepatitis B immunoglobulin and viral hepatitis, pentamidine (antimicrobial) and pneumonia, and rivaroxaban (anticoagulant) and melaena. They may also be markers of the drug being ineffective, but deciding which is true is difficult, if not impossible, from spontaneous reports.

While the SPRT method is intended to allow for multiple looks at accumulating data, it does not explicitly address other issues of multiplicity. There are over 20,000 drug-AE pairs that are tested, and none of the methods make explicit allowance for this form of multiplicity. Here, reports are of suspected adverse drug reactions, so the possibility that they are all chance effects is not tenable, and the application of Bonferroni types of correction would be too extreme and lead to a notable loss of power. False discovery type methods, as described by Gould [12] and Ahmed et al. [29], do address these forms of multiplicity, and they could be applied to SPRT methods as well.

Therefore, it is not possible to choose the hRR to be detected based solely on statistical grounds. Furthermore, it should be noted that the SPRT is a sequential test, and applying it to an existing database is not the most appropriate approach to its evaluation. The most appropriate approach would be to look at the newly arrived data and see what SDRs are detected. It is also clear that the actual hRR that is

most likely to be a real effect is of relevance. It is very likely that very high hRRs for reasonably common effects will have been detected in randomised trials used for licensing. However, extremely rare reactions would not be detected, and spontaneous AE reports are the best tool for detecting them.

Having a large value of hRR will generate SDRs at very low observed or expected counts, but at high observed or expected counts, the signals might be missed. Hence, a detailed analysis of the effect of different hRR values on the database is necessary, and we have tried having different thresholds or hRR values for different observed counts or expected counts, but this does not seem practically sensible. However, having different thresholds for different types of AEs, depending on their rarity, may well be sensible. This may not be easy to define, but is worth exploring in the future.

5. CONCLUSIONS

It appears that SPRT may have some applications in the Singapore's SRS. For AEs that are rare and thus expected to have low expected counts, applying a higher hRR for the SPRT method may pick up SDRs of concern. On the other hand, AEs that are more common need a lower hRR to weed out false positives. To appreciate the value of SPRT in the Singapore database, more in-depth analysis comparing the value of the signals picked up by varying the hRRs would be a useful next step of investigation. Other countries, especially with smaller databases, may find this simple method of SPRT can be applied very easily to their databases and may provide signal detection for some rare events of significance to them. Assuming they have a database and the ability to produce the counts of the pairs, then it is easy to

apply the method, and this could be done using any spreadsheet or statistical software.

Compliance with Ethical Standards

Ethical approval

All data are anonymised, and ethical approval was not required.

Funding

The SAPHIRE (Surveillance and Pharmacogenomics Initiative for Adverse Drug Reactions) Project was funded by a Strategic Positioning Funds grant from the Biomedical Research Council of the Agency for Science, Technology and Research of Singapore (SPF2014/001).

Conflict of interest

Cheng Leng Chan, Sowmya Rudrappa, Pei San Ang, Shu Chuen Li and Stephen JW Evans have no conflicts of interest that are directly relevant to the content of this study.

Appendix

A1 Introduction

The SPRT was developed by A. Wald in the 1940s [14,15] and was mainly used in process monitoring. A review of the pharmacovigilance literature showed that the SPRT may have useful properties not shared by other methods [16-18]. It has been used in the context of scanning electronic health records, particularly for vaccine studies, but has not had extensive evaluation for spontaneous reports. Specifically, SPRT compares two hypotheses based on the likelihood of observing the data given those hypotheses [19]. Unlike the other methods, SPRT is based on the difference between (rather than the ratio of) the observed and expected values. However, there is limited research conducted on SPRT to evaluate its performance in a Spontaneous Reports System (SRS) database similar to Singapore's. A previous evaluation of SPRT used a single alternative hypothesis on simulated data [20]. On theoretical grounds, at least, the methods that do not allow for this form of multiple testing will have a higher rate of false positive findings – false signals. The method is essentially applied to a 2 x 2 table of data in the same way as other methods and where the emphasis is on obtaining an expected count assuming no association between the drug and the adverse event (AE). The method is applied to a single drug and adverse event and a determination of whether there is evidence of an association is made. The process is repeated, doing the same calculations for the next possible pairing and so on, for every drug-AE combination that occurs in the database.

A1.1 The methods of use for SPRT in signal detection

For any drug-AE pair, the analysis of quantitative data mining methods can be based on 2 x 2 contingency tables (Table 4). We focus on one particular drug-AE pair and include the totals of reports of other drugs and AEs.

The number of reports listing drug and the AE are the observed values, and the expected value for each drug-AE pair is calculated (just as for a chi-square test), assuming independence, by the formula:

$$E = \frac{((a + c) * (a + b))}{(a + b + c + d)} \quad A1.1$$

We assume the continuously accumulating data in SRS is described by the Poisson distribution [17] which is widely used in pharmacoepidemiology and pharmacovigilance.

The SPRT uses the log likelihood ratio (LLR) for the Poisson distribution calculated [6] by the formula:

$$LLR = O \times \log_e(hRR) - E \times (hRR - 1) \quad A1.2$$

Where for any drug-AE pair,

- O is the observed number of reports, which is equal to cell a in Table 4.
- hRR is the hypothesised relative risk of interest to constitute a signal.
- E is the expected count as calculated at *A1.1*.

This is a very simple calculation which uses the observed and expected counts multiplied by factors related to the hypothesised relative risk.

A1.2. Varying thresholds in signal detection

A and B are two thresholds in the SPRT method that determine, in this context, whether a signal has been detected:

The signalling rules for SPRT are:

- If $LLR > B$, then reject H_0 and accept H_1 .
- If $LLR < A$, then reject H_1 and accept H_0 .

In the conventional use of SPRT in process control one can stop monitoring if there is convincing evidence that there is no effect. However, in the case of post marketing surveillance, monitoring continues indefinitely, as signals can still occur subsequently due to various factors such as change in pattern of drug use and prescribing, so the threshold A is irrelevant in this context.

The threshold is calculated using α and β ; α (type I error), the probability of incorrect rejection of a true null hypothesis, also called a "false positive". β (type II error) is the probability of accepting a false null hypothesis, also called a "false negative". $1 - \beta$ gives the power of the test.

The values A and B are obtained using the formulae:

$$A \approx \log_e \{\beta / (1 - \alpha)\} \quad A1.3$$

$$B \approx \log_e \{(1 - \beta) / \alpha\} \quad A1.4$$

To minimize the false positives, the value of α should be made small. To lessen the false negatives, the value of β should also be made small. We have used conventional α and β values at 0.05 and 0.20 respectively. This gives the approximate threshold value to be 2.77 for B. Different threshold values based on different choices of α and β can also be derived (see Table 5).

A1.3. Observed and expected counts and their effect on LLR

In order to understand how SPRT works, it is helpful to see what happens to the LLR when its parameters change. A graph of LLR is shown in Fig. 2 with varying values of hRR (in the range of 1 – 6) with a fixed observed count of 3 and an expected count of 1. This graph shows that the most likely value of hRR, (its maximum LLR) is at hRR of 3, and the likelihood of other values of hRR fall off the further hRR is from 3.

A1.4. Observed and expected counts and the relation to hRR

In order to detect a signal, with a threshold value of 2.77 (i.e. α at 0.05 and β values at 0.20) and expected count of 1, using an alternative hypothesis hRR of 2 the minimum observed count is 6 (see Fig. 3). If the hRR is increased to 3, the LLR will cross the threshold when the observed count is greater than or equal to 5.

If we reduce the expected value to zero (though we cannot ever have exactly zero in practice, but it can be very close to zero), then we see that the LLR formula in A1.2 would approximately be equal to $LLR = O \times \log_e (hRR)$. Hence, we can make LLR

arbitrarily large with any observed count (and an extremely small expected count), including a low one, cross the threshold by having a sufficiently high value of the alternative hypothesis for hRR. Although this at first seems counter-intuitive, a very high hRR becomes more likely than the null even with a low observed count with an extremely small expected count. This is because the SPRT is a test comparing two simple hypotheses. Table A3 gives the hRR for the alternative hypothesis with a corresponding minimum observed count needed to cross a threshold of 2.77. It was observed that when SPRT was applied with a low hRR of 2 as the alternative hypothesis to our SRS database, it resulted in low numbers of SDRs. This could be explained as more than 90% of the drug-AE pairs in our SRS had less than 10 reports (see Fig. 4).

In all databases of spontaneous reports, low cell counts occur very frequently (and zero counts are the most frequent of all as noted above), and this is a particular issue with a relatively small database in Singapore compared to the US FDA Adverse Event Reporting System or the European EudraVigilance database. On the other hand, if the alternative hypothesis for hRR is sufficiently high then signals can be obtained even with a single observed count. However, this means that with higher observed and expected values, but with lower values of hRR, then signals become more difficult to detect.

A1.5. Use of two hypothesised values

Initially SPRT with $hRR = 2$ was applied as has been used by others in references cited above. However, rare events with relatively small values of the observed and small values of the expected count, did not produce signals. Fig. 5 shows that the numbers of reports rise when the value of hRR reaches 4.1. This is because the minimum value for the observed count to be a signal with an expected close to zero, becomes 2 and so potentially many more cells have a count of 2 than a count of 3. Rarer AEs, such that although they only occur twice, but have a very small expected value and hence were not detected as a signal using SPRT with $hRR=2$ could be signalled more easily. Fig. 6 shows the overlap of significant pairs detected by SPRT with $hRR=2$ and $hRR=4.1$.

Signals with a notable excess based on a low number of reports with very low expected values are of medical interest as well as those with larger numbers of reports, so we decided to include all the SDRs generated by SPRT with $hRR=2$ or with $hRR=4.1$, if either crossed the threshold of 2.77.

All these calculations are repeated using each drug-AE combination that exists. The calculations are quite simple; the complexity is in doing them for every combination and classifying each pair as giving rise to a signal or not.

A1.6. Comparison of SPRT with other methods

In our study, three methods in common use i.e. reporting odds ratio (ROR), Bayesian Confidence Propagation Neural Network (BCPNN) and Gamma Poisson Shrinker (GPS) were compared with SPRT. Virtually all those using these methods use a criterion based on a minimum count of drug-AE pairs ($N \geq 3$) so this was also imposed here. The reason is that the analyses of ROR and GPS methods with small cell counts with one or two reports can result in statistical instability associated with disproportionalities. If the criterion of $N \geq 3$ was not applied, ROR and GPS methods would generate many SDRs which are false positives. When we did not specify $N \geq 3$ for GPS, the method signalled several drug-AE pairs with two counts. As explained in the study by Ang et al [21], using a stricter threshold criterion for these methods will not give a better performance. Table 7 gives the formulae and thresholds for all the four methods compared in this study. Table 8 shows the formulae for calculating the sensitivity, specificity, positive predictive value and negative predictive value.

6. REFERENCES

1. Waller PC, Evans SJW. A model for the future conduct of pharmacovigilance. *Pharmacoepidemiol Drug Saf.* 2003;12(1):17-29. doi:10.1002/pds.773.
2. Waller P. An introduction to pharmacovigilance. United Kingdom: Wiley-Blackwell; 2010.

3. Koh Y LA, Tan L, Ang PS, Tan SH, Toh D, Chan CL. Pharmacovigilance in Singapore - harnessing IT and genomics to detect safety signals. *Regul Aff J - Pharma*. 2012:13-5.
4. Hauben M, Bate A. Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov Today*. 2009;14(7–8):343-57. doi:10.1016/j.drudis.2008.12.012.
5. Zink RC, Huang Q, Zhang L-Y, Bao W-J. Statistical and graphical approaches for disproportionality analysis of spontaneously-reported adverse events in pharmacovigilance. *Chin J of Nat Med*. 2013;11(3):314-20. doi:10.1016/S1875-5364(13)60035-7.
6. 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. doi:10.1002/pds.677.
7. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf*. 2009;18(6):427-36. doi:10.1002/pds.1742.
8. Egberts AC, Meyboom RH, van Puijenbroek EP. Use of measures of disproportionality in pharmacovigilance: three Dutch examples. *Drug Saf*. 2002;25(6):453-8. doi:10.2165/00002018-200225060-00010.
9. 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. doi:10.1007/s002280050466.
10. Dumouchel W. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. *Am Stat*. 1999;53(3):177-90. doi:10.1080/00031305.1999.10474456.
11. 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 Safety*. 2002;25(6):381-92. doi:10.2165/00002018-200225060-00001.
12. Gould AL. Detecting Potential Safety Issues in Clinical Trials by Bayesian Screening. *Biom J*. 2008;50(5):837-51. doi:10.1002/bimj.200710469.
13. Candore G, Juhlin K, Manlik K, Thakrar B, Quarcoo N, Seabroke S et al. Comparison of Statistical Signal Detection Methods Within and Across Spontaneous Reporting Databases. *Drug Saf*. 2015;38(6):577-87. doi:10.1007/s40264-015-0289-5.
14. Wald A. Sequential Tests of Statistical Hypotheses. *Ann Math Stat*. 1945.

15. Wald A. Sequential Analysis. New York: John Wiley & Sons, Inc. 1947.
16. Evans S. Sequential probability ratio tests applied to public health problems. *Control Clin Trials*. 2003. **24** 67S
17. Talbot J, Aronson JK. Stephens' detection and evaluation of adverse drug reactions: principles and practice. 6th ed. Oxford. Wiley-Blackwell; 2012.
18. Leite A, Andrews NJ, Thomas SL. Near real-time vaccine safety surveillance using electronic health records-a systematic review of the application of statistical methods. *Pharmacoepidemiol Drug Saf*. 2016 Mar;25(3):225-37. doi:10.1002/pds.3966.
19. Spiegelhalter D, Grigg O, Kinsman R, Treasure T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *Int J Qual Health Care*. 2003;15(1):7-13.
20. 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. doi:10.1109/TITB.2005.855566A.
21. Ang PS, Chen Z, Chan CL, Tai BC. Data mining spontaneous adverse drug event reports for safety signals in Singapore - a comparison of three different disproportionality measures. *Expert Opin Drug Saf*. 2016;15(5):583-90. doi:10.1517/14740338.2016.1167184.
22. Stahl M, Lindquist M, Edwards IR, Brown EG. Introducing triage logic as a new strategy for the detection of signals in the WHO Drug Monitoring Database. *Pharmacoepidemiol Drug Saf*. 2004;13(6):355-63. doi:10.1002/pds.894.
23. Micromedex® 2.0 (electronic version). Truven Health Analytics, Greenwood Village, Colorado, USA. Available at: <http://www.micromedexsolutions.com/>. Accessed Sep 19 2015.
24. Safety Reporting Requirements for Human Drug and Biological Products; Federal Register. Department of Health and Human Services. US Food and Drug Administration. 14 March 2003. Available at: <http://www.fda.gov/OHRMS/DOCKETS/98fr/03-5204.pdf>. Accessed Sep 19 2015.
25. Important Medical Event Terms list (based on MedDRA version 18.1). EudraVigilance Expert Working Group. Available at: <https://eudravigilance.ema.europa.eu/human/textforIME.asp>. Accessed Sep 19 2015.

26. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2014. Available at: <http://www.R-project.org>. Accessed Sep 23 2016.
27. Ahmed I, Poncet A. Package 'PhViD'. Version:1.0.6. 2013. Available at: <https://cran.r-project.org/web/packages/PhViD/PhViD.pdf> Accessed Sep 19 2015.
28. Bisdorff AR. Management of vestibular migraine. Ther Adv Neurol Disord. 2011;4(3):183-91. doi:10.1177/1756285611401647.
29. Ahmed I, Thiessard F, Miremont-Salame G, Haramburu F, Kreft-Jais C, Be'gaud B et al. Early Detection of Pharmacovigilance Signals with Automated Methods Based on False Discovery Rates. Drug Saf. 2012;35(6):495-506. doi:10.2165/11597180-000000000-00000.
30. Noren 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. doi:10.1177/ 0962280211403604.

Tables and figures

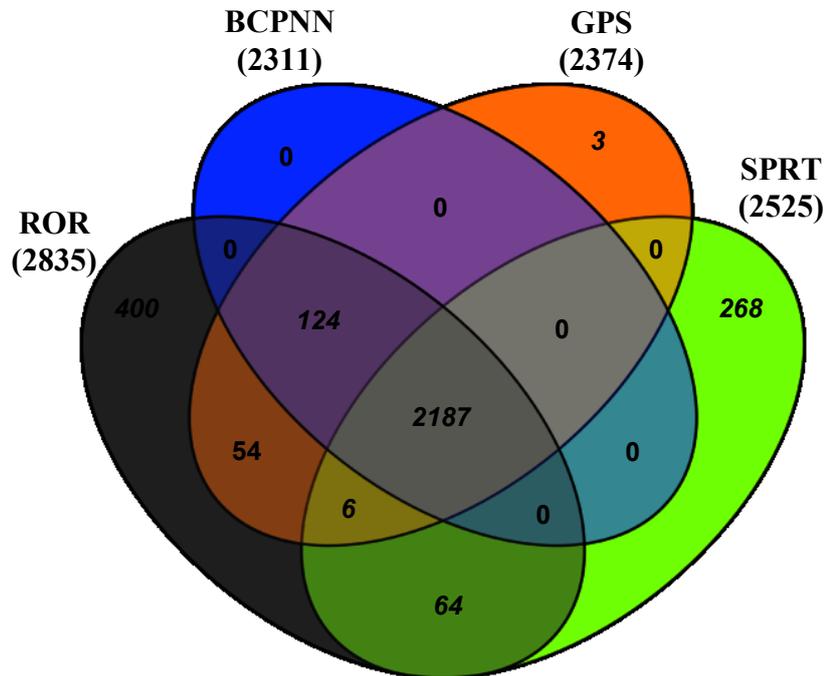
Table 1. Comparing SPRT with hRR=2 and hRR=4.1 in terms of seriousness of AE, whether the AE is labelled or not labelled, observed and expected values.

Year	SPRT	No. of SDRs	Not labelled		Labelled		Observed			Expected		
			Serious	Not serious	Serious	Not serious	Median	5 th Percentile	95 th Percentile	Median	5 th Percentile	95 th Percentile
1993 – 2013	hRR=2 & hRR=4.1 (overlap)	1232	47	71	418	696	11	5	141.8	1	0.02	22.4
	hRR=2 only	200	2	6	44	148	32	10	716	16.9	4	388.2
	hRR=4.1 only	1093	88	129	382	494	3	2	5	0.1	0.001	1

hRR = risk of interest/relative risk in the alternative; SDR = signals of disproportionate reporting; SPRT = Sequential Probability Ratio Test; 5th Percentile is first 5% of data values; 95th Percentile is first 95% of data values.

s. It measures the veracity of SDR on labelledness.

Figure 1: Venn diagram for data from 1993 – 2013 to illustrate significant SDRs by the four methods and their inter-relationships (shapes not drawn to proportion).



ROR = reporting odds ratio; BCPNN = Bayesian Confidence Propagation Neural Networks; GPS = Gamma Poisson Shrinkage; SPRT = Sequential Probability Ratio Test; SDR = signals of disproportionate reporting

Table 2. Breakdown of no. of SDRs generated by each method according to different quarters of years

Cumulative (from 1993)	No. of unique pairs	ROR		BCPNN		GPS		SPRT (hRR=2 OR hRR=4.1)	
		Total no. of SDRs	New Signals #	Total no. of SDRs	New Signals #	Total no. of SDRs	New Signals #	Total no. of SDRs	New Signals #
2012-Q1	19,577	2361	120	1897	120	1945	109	2057	115
2012-Q2	20,155	2440	123	1984	109	2029	97	2145	119
2012-Q3	20,761	2519	125	2042	94	2104	98	2214	104
2012-Q4	21,243	2579	91	2099	72	2151	74	2275	95
2013-Q1	21,720	2656	108	2156	86	2220	94	2327	84
2013-Q2	22,181	2720	94	2209	80	2281	80	2392	91
2013-Q3	22,732	2792	110	2261	77	2327	81	2462	100
2013-Q4	23,183	2835	77	2311	66	2374	68	2525	98

No. of new significant pairs which changed from not significant (No) to significant (Yes) and new significant pairs which were not present in the previous period.

ROR = reporting odds ratio; BCPNN = Bayesian Confidence Propagation Neural Networks; GPS = Gamma Poisson Shrinkage; SPRT = Sequential Probability Ratio Test; SDR = signals of disproportionate reporting;

Table 3: Comparisons of methods in terms of seriousness of AE, whether the AE is labelled or not labelled, PPV, NPV, sensitivity and specificity (1993 – 2013)

Methods	Total no. of SDRs generated	Not Labelled		Labelled		PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
		Serious (%)	Not serious (%)	Serious (%)	Not serious (%)					
ROR (LB>1 & N>=3)	2835	105 (3.7)	186 (6.7)	885 (31.2)	1659 (58.5)	89.7	32.8	15.7	95.8	39.7
BCPNN	2311	86 (3.7)	146 (6.3)	758 (32.8)	1321 (57.2)	90.0	32.2	12.8	96.7	38.0
GPS (EB05>1 & N>=3)	2374	88 (3.7)	150 (6.3)	774 (32.6)	1362 (57.3)	90.0	32.3	13.2	96.6	38.2
SPRT (hRR=2 OR hRR=4.1)	2525	137 (5.4)	206 (8.2)	844 (33.4)	1338 (53.0)	86.4	32.0	13.5	95.1	38.0

ROR = reporting odds ratio; BCPNN = Bayesian Confidence Propagation Neural Networks; GPS = Gamma Poisson Shrinkage; SPRT = Sequential Probability Ratio Test; SDR = signals of disproportionate reporting; NPV = negative predictive value; PPV = positive predictive value

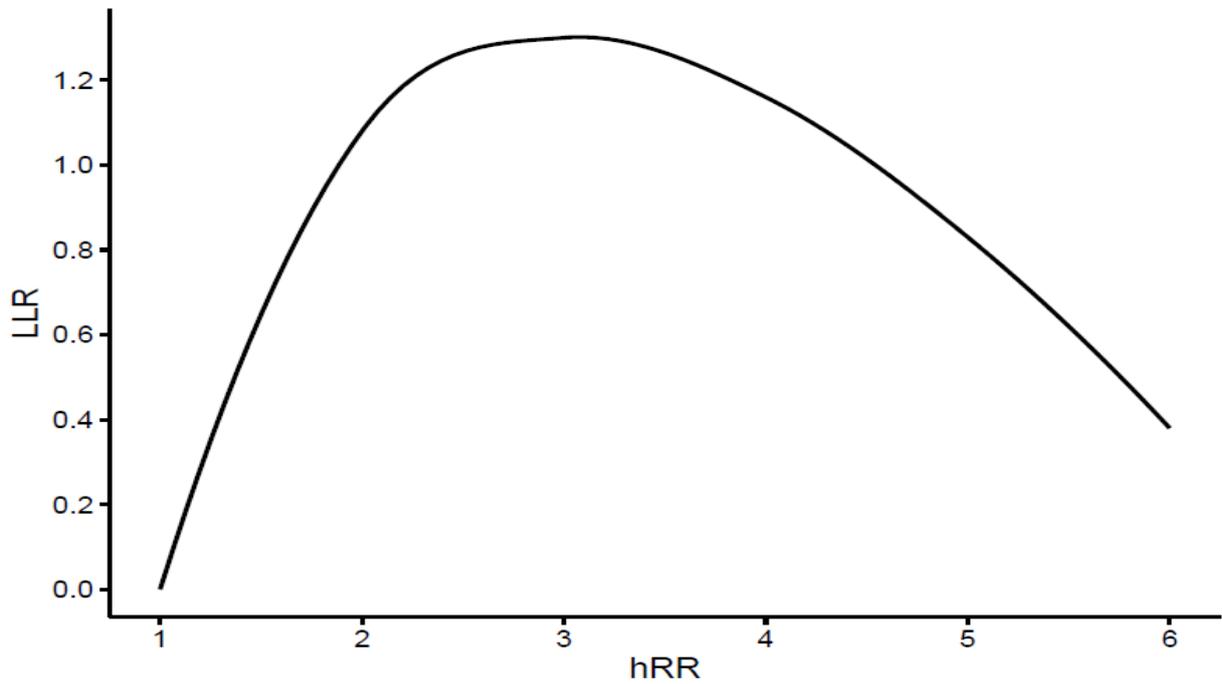
Table 4. 2 x 2 contingency tables of AE and drugs.

	No. of reports listing AE	No. of reports not listing AE	Total
No. of reports listing drug	a	b	a + b
No. of reports not listing drug	c	d	c + d
Total	a + c	b + d	a + b + c + d

Table 5. Different threshold values based on different choices of α and β .

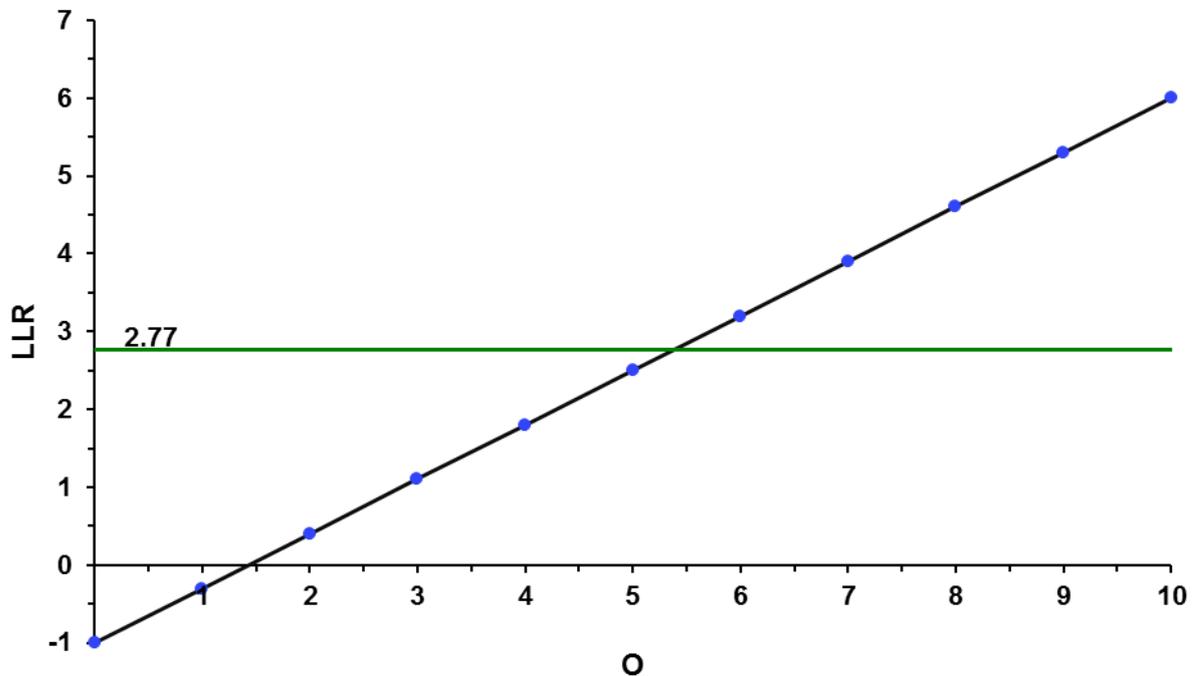
α	β	Threshold A	Threshold B
0.01	0.01	-4.61	4.61
0.05	0.05	-2.95	2.94
0.05	0.1	-2.26	2.89
0.05	0.2	-1.56	2.77
0.1	0.1	-2.20	2.20
0.1	0.2	-1.51	2.08

Figure 2: Graph of LLR with varying values of hRR for fixed O=3 and E=1.



LLR: log likelihood ratio; O: observed number of reports; E: expected values for each drug-AE pair

Figure 3: LLR vs observed values for fixed hRR=2 and E=1



LLR: log likelihood ratio; O: observed number of reports; hRR: hypothesised relative risk (risk of interest/relative risk in the alternative); E: expected values for each drug-AE pair. Threshold=2.77

Table 6. Varying values of hRR with corresponding qualifying SDRs.

Cut-off of threshold (LLR)	Expected value	hRR	Minimum observed value
2.77	1	2	6 and above
		3	5 and above
		4	5 and above
		8	5 and above
2.77	~0	2	4 and above
		3	3 and above
		4	2 and above
		8	2 and above

Figure 4: Distribution of number of reports received for drug-AE pairs (1993 – 2013)

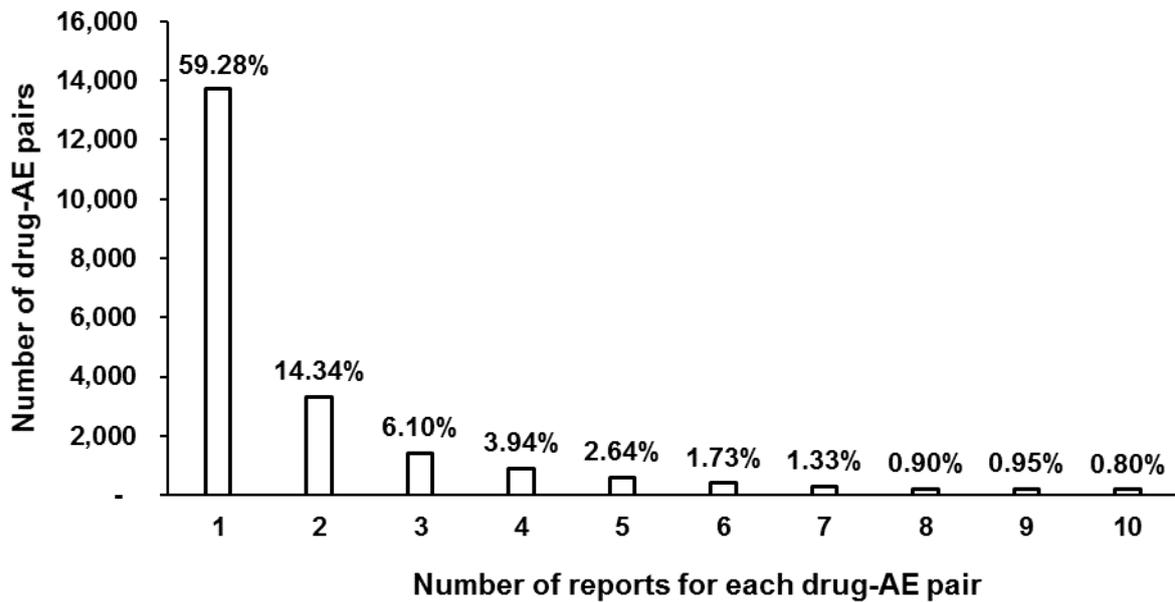
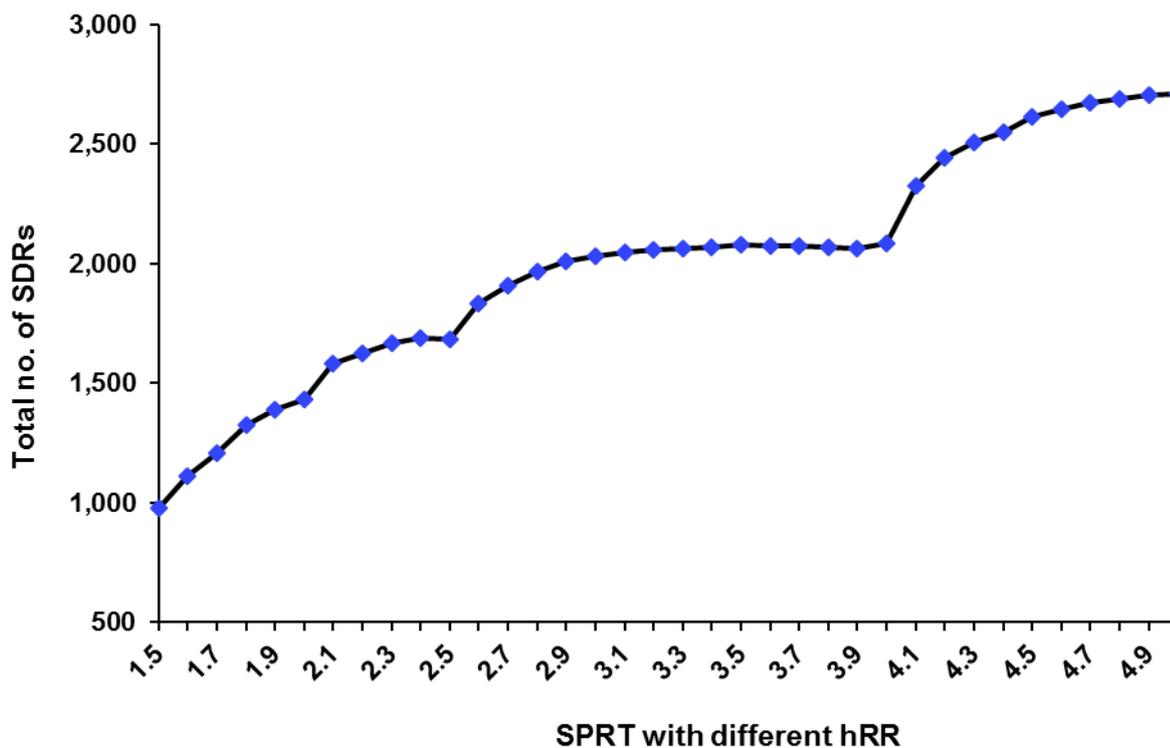
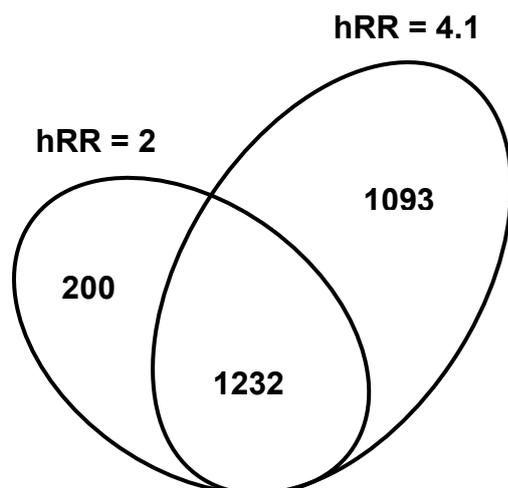


Figure 5: Total number of SDRs with increasing hRR for data from 1993 to 2013



hRR = hypothesised relative risk (risk of interest/relative risk in the alternative); SDR = signals of disproportionate reporting; SPRT = Sequential Probability Ratio Test

Figure 6: Venn diagrams showing overlap of SDRs detected by SPRT with hRR=2 and hRR =4.1 (1993 – 2013)



hRR = hypothesised relative risk (risk of interest/relative risk in the alternative); SDR = signals of disproportionate reporting; SPRT = Sequential Probability Ratio Test

Table 7. Formulas and thresholds of methods used

Methods	Formula	Threshold
SPRT	$LLR = O \times \log_e(hRR) - E \times (hRR - 1)$	2.77 [#]
ROR	$ROR = (a \times d)/(b \times c)$	$ROR - 1.96SE > 1, N \geq 3$
BCPNN	$IC = \log_2(O/E)$	$IC - 1.96SD > 0$
GPS	$EBGM = e^{(EBln)}$	$EB05^{**} > 1, N \geq 3$

$\alpha=0.05$ and $\beta=0.20$, SPRT = Sequential Probability Ratio Test; ROR = reporting odds ratio; BCPNN = Bayesian Confidence Propagation Neural Networks [30]; GPS = Gamma Poisson Shrinkage [2]; ** EB05 denotes fifth percentile of confidence limits of empirical Bayes geometric mean (EBGM). Calculations were done using R software and a signal detection package – PhVID was also used [26,27].

Table 8. Concepts of sensitivity, specificity, positive predictive value, and negative predictive value.

Result/signal	Gold standard (true causality)		Total	
	Labelled	Not labelled		
Yes	TP (true positive)	FP (false positive)	TP + FP	PPV = TP/(TP + FP)
No	FN (false negative)	TN (true negative)	FN + TN	NPV = TN/(FN + TN)
Total	TP+FN	FP + TN	N=TP+FP+FN+TN	Accuracy = (TP+TN)/N
	Sensitivity = TP/(TP + FN)	Specificity = TN/(FP+TN)		

Sensitivity is the proportion of true positive signals based on those drug–AE pairs which were labeled.

Specificity is the proportion of true negative signals that did not map to labeled drug–AE pairs. A high specificity indicates low false positive signals.

Positive predictive value (PPV) is the proportion of the drug–AE pairs that were labeled and were signaled as statistically significant drug–AE pairs.

Negative predictive value (NPV) is the proportion of the drug–AE pairs that were not labeled and were not signaled drug–AE pairs.

Accuracy is the proportion of true signals (both positive and negative) based on total drug-AE pairs. It measures the veracity of SDR on labelledness.