Adverse drug reaction or innocent bystander? A systematic comparison of statistical discovery methods for spontaneous reporting systems

Louis Dijkstra1 | Marco Garling2 | Ronja Foraita1 | Iris Pigeot1

1Leibniz Institute for Prevention Research & Epidemiology, BIPS, Achterstraße 30, 28359 Bremen, Germany
2Scientific Institute of TK for Benefit & Efficiency in Health Care, WINEG, Bramfelder Straße 140, 22305 Hamburg, Germany

Correspondence
Louis Dijkstra, Achterstraße 30, 28359 Bremen, Germany.
Email: dijkstra@leibniz-bips.de

Funding information
Federal Joint Committee in Germany, Grant/Award Number: 01VSF16020

Abstract
Purpose: Spontaneous reporting systems (SRSs) are used to discover previously unknown relationships between drugs and adverse drug reactions (ADRs). A plethora of statistical methods have been proposed over the years to identify these drug-ADR pairs. The objective of this study is to compare a wide variety of methods in their ability to detect these signals, especially when their detection is complicated by the presence of innocent bystanders (drugs that are mistaken to be associated with the ADR, since they are prescribed together with the drug that is the ADR’s actual cause).

Methods: Twelve methods, 24 measures in total, ranging from simple disproportionality measures (eg, the reporting odds ratio), hypothesis tests (eg, test of the Poisson mean), Bayesian shrinkage estimates (eg, the Bayesian confidence propagation neural network, BCPNN) to sparse regression (LASSO), are compared in their ability to detect drug-ADR pairs in a large number of simulated SRSs with varying numbers of innocent bystanders and effect sizes. The area under the precision-recall curve is used to assess the measures’ performance.

Results: Hypothesis tests (especially the test of the Poisson mean) perform best when the associations are weak and there is little to no confounding by other drugs. When the level of confounding increases and/or the effect sizes become larger, Bayesian shrinkage methods should be preferred. The LASSO proves to be the most robust against the innocent bystander effect.

Conclusions: There is no absolute “winner”. Which method to use for a particular SRS depends on the effect sizes and the level of confounding present in the data.

Keywords
innocent bystander effect, pharmacoepidemiology, pharmacovigilance, side effect, surveillance

1 INTRODUCTION

Despite great efforts in testing drugs for possible side effects, some, in particular rare and/or late-onset, adverse drug reactions (ADRs)
remains undetected until after the launch of the respective drugs.\textsuperscript{1-3} Spontaneous reporting systems (SRSs) have been established over the years\textsuperscript{4-7} in order to detect these unknown ADRs. Medical professionals and patients can send in a report when a drug is suspected to have triggered an adverse event (AE). These reports are collected, stored, and subsequently analyzed by medical experts.

A plethora of statistical methods have been proposed to aid in the detection of drug-ADR pairs. All of these methods involve two steps:

1. For each drug-AE pair, a score is computed that reflects their "association strength";
2. These scores are used to compile a shortlist of drug-AE pairs; any pair with a score that exceeds a predefined threshold is forwarded to the experts for further investigation.

All methods except the LASSO\textsuperscript{8,9} base their scores on the 2 \times 2 contingency tables for each of the drug-AE pairs (see Table 1). The count $a$ denotes the number of reports that contain both the drug and AE of interest, $b$ is the number of reports that mention the drug but not the AE, and so on. Some measures that rely on these tables are, for example, the reporting odds ratio\textsuperscript{10} (ROR = $(ad)/(bc)$) and the proportional relative risk\textsuperscript{11} (PRR = $[a/(a+b)]/[c/(c+d)]$).

A potential downside of these disproportionality measures\textsuperscript{7,12} is that they are vulnerable to the innocent bystander effect.\textsuperscript{2,7,8} This effect refers to a form of confounding where a drug is thought to be associated with a certain ADR, because it is prescribed together with the drug that causes the reaction.

Several studies in the past compared the performance of some of these statistical methods. Van Puijenbroek et al\textsuperscript{10} compared five methods to the output of the Bayesian Confidence Propagation Neural Network (BCPNN).\textsuperscript{1,13} Such a comparison study relies on the quality of the chosen "reference method"; if it performs poorly in certain cases, it is impossible to see whether the other methods might do better. Other comparison studies relied on a reference set; a list of drug-ADR pairs that are thought to be associated and pairs that are not.\textsuperscript{14-17} These true positives and negatives are used to test whether the method is indeed able to pick up on those associations. The Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT) performed the most extensive study based on a reference set.\textsuperscript{15} Fifteen measures with varying thresholds were applied to seven databases. It showed that no measure performed uniformly better than the others.

Other studies rely on simulated data,\textsuperscript{18,19} which has the advantage that (a) the associations are truly known and (b) the methods' performance can be explored in a variety of parameter settings, for example, for varying odds ratios between drugs and ADRs. A disadvantage of simulation studies is that it is unclear to what extent the simulated data reflects reality. The simulated data sets that were used, however, assume independence between drugs,\textsuperscript{18,19} which is unrealistic. It is, therefore, unclear to what extent the measures' performance is affected by the presence of innocent bystanders. Many of the studies used predefined thresholds,\textsuperscript{20,21} except the work by Candore et al.\textsuperscript{15} Using predefined thresholds might skew the results, since the choice of thresholds could be unfavorable for a particular method.

In this paper we perform a simulation-based study with 24 measures, ranging from simple disproportionality measures, hypothesis tests, Bayesian shrinkage estimates, to sparse regression (see Section 2). The simulations are unique in the sense that they contain innocent bystanders (for a description see Section 3). Instead of using thresholds, we use a threshold-free measure, the area under the precision-recall curve for assessing the performance of the presented measures. Our results are presented in Section 4 and critically appraised in the discussion.

## 2 | Statistical Methods for Postmarketing Surveillance

Table 2 lists the 24 measures considered in this article. The column "Category" contains a categorization of the various measures. The column "Measure" contains the notation for each measure, where we try to stick to the notation commonly used in the literature. The year and the publication in which the measure is mentioned and/or used for the first time can be found in the columns "Year" and "Reference(s)."

For some measures, it is common to use the lower endpoint of the confidence/credible interval. This is denoted here by the subscripts "025" and "05." In case of "025," the 95% confidence/credible interval is used; in case of "05" the 90% interval. The "Number of Reports" measure, equal to $a$ in Table 1, is the number of times the drug and AE are reported together. This measure was suggested by Norén et al\textsuperscript{22} as a "placebo" measure; each measure should at least be able to outperform this basic count. There are two versions of the BCPNN, which we denote here with the superscripts "original" and "alternative." The former refers to the version as it was first proposed by Bate et al.\textsuperscript{15} The latter uses a different prior.\textsuperscript{1} Yule's Q is not
considered here, since it is a rescaling of the ROR\textsuperscript{10} and, therefore, performs equally well. All measures are implemented and publicly available as the R package pvm at www.github.com/bips-hb/pvm.

### Table 1: 2 × 2 contingency table for a drug-AE pair

|            | AE of interest | All other AEs | Total |
|------------|----------------|---------------|-------|
| Drug of interest | $a$            | $b$           | $a + b$ |
| All other drugs | $c$            | $d$           | $c + d$ |
| **Total**   | $a + c$        | $b + d$       | $N$   |

### 3 | METHODS

We first introduce the simulation setup, before we describe how the measures’ performance will be assessed. An extensive description of the simulation can be found in Appendix A. Each report to an SRS contains two lists:

1. The AEs the patient experienced, and
2. The drugs suspected to have caused the AEs.

All the reported drugs in the SRS are represented by binary variables, $X_1, X_2, \ldots, X_m$. Similarly, all AEs are represented by $Y_1, Y_2, \ldots, Y_n$. A report can then be represented by the binary vector:

$$\text{Report} = (X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n).$$

where $X_i$ is 1 if the report contains drug $i$, and 0 otherwise. Similarly, $Y_j$ is 1 when the report lists AE $j$, and 0 otherwise. An SRS is a collection of reports:

$$\text{SRS} = (\text{report}_1, \text{report}_2, \ldots, \text{report}_n).$$

where $N$ is the number of reports. In the simulations, there are 500 drugs, 500 AEs, and 50,000 reports.

Some drugs are prescribed more frequently when another drug is prescribed, for example, to suppress side effects. There are two types of drugs, that is, the probability of the drug being listed depends on (a) no other drugs or (b) one other drug. In the first case, the probability of a drug, $X_k$, to be reported is $P(X_k = 1) = \pi_k$. In case a drug, $X_k$, is influenced by drug $X_h$, we specify two conditional probabilities: (a) when drug $X_h$ is not on the report and (b) when drug $X_h$ is on the report, that is,

$$P(X_j = 1 \mid X_h = 0) = \pi_j \text{ and } P(X_j = 1 \mid X_h = 1) = \gamma_j.$$  

Note that $\gamma$ is the same for all $j$. The probabilities, $\pi_i$, are drawn from a beta distribution, which is a common choice in the field, with rate and shape parameters 1 and 20, respectively, to ensure that drugs tend to be listed infrequently. In the simulations, $\gamma$ is .5, .75, or .9.

### Table 2: An overview of all 24 measures considered. The column “Category” provides a rough categorization. The notation for each measure used in this paper can be found in “Measure.” “Year” and “Reference(s)” refer to the first appearance of the respective measure in the field of pharmacovigilance

| Category                                      | Measure   | Year | Reference(s) |
|-----------------------------------------------|-----------|------|---------------|
| Number of reports                             | # reports | 2014 | 22            |
| Reporting odds ratios                         | ROR       | 2002 | 23            |
| Proportional relative risk                    | PRR       | 2001 | 10, 11        |
| Chi squared                                   | $\chi^2$ YaT | 2002 | 10            |
| Binomial likelihood ratio test                | $\alpha_{\text{binomial}}$ | 2013 | 26            |
| Test of the Poisson mean                      | $\Pi_{\text{poisson}}$ | 1991 | 24            |
| BCPNN (original)                              | $\Pi_{\text{original},05}$ | 1998 | 13            |
| BCPNN (alternative)                           | $\Pi_{\text{alternative},05}$ | 2006 | 1             |
| Gamma Poisson shrinker (GPS)                  | EBGM      | 1999 | 24            |
| Penalized regression                          | LASSO     | 2008 | 8             |

There are $500 \times 500 = 250,000$ drug-AE pairs. The relationship between a drug $X_i$ and an AE $Y_j$ is expressed using a logistic model:

$$\logit(P(Y_j = 1 \mid X_i = x_i)) = \beta_i + \log(OR_j) x_i,$$

where $\logit(x) = \log[x / (1 - x)]$, $\beta_i$ is the intercept and $OR_j$ is the odds ratio between the $i$-th drug and the $j$-th AE. In the case that the drug causes the event (ie, an ADR), the $OR_j \in [1, \infty)$ is drawn from a truncated normal distribution with mean 1.5, 3, or 5. When the drug does not cause the AE, $OR_j = 1$. The intercept, $\beta_i$, is chosen such that the probability of the AE appearing on the report is small (see Appendix A.2).

Modeling an innocent bystander requires to specify the relationships between two drugs and one ADR. Let $X_i$ be the drug that causes the ADR $Y_j$ and $X_h$ be the innocent bystander, that is, $OR_h > 1$ and $OR_h = 1$.

The dependence between the innocent bystander, $X_h$, and drug $X_i$ is defined according to Equation (3): $P(X_i = 1 \mid X_h = 0) = \pi_{x_i}$ and $P(X_i = 1 \mid X_h = 1) = \gamma$. Only drug $X_h$ causes the ADR $Y_j$. The innocent bystander, $X_h$, is often reported alongside $X_i$ and can, therefore, be mistaken as the cause of $Y_j$. 

---

**Note:** The text above is a concise summary of the content from the provided document. For a detailed understanding, please refer to the original document.
ADR pairs. In our simulation, drug The goal of the statistical discovery methods is to identify these drug-

TABLE 3

| Description | Equation | Notation | Values |
|-------------|----------|----------|--------|
| Number of drugs | (1) | m | 500 |
| Number of AEs | (1) | n | 500 |
| Number of reports | (2) | N | 50 000 |
| Number of repetitions | – | – | 50 |
| Number of drug-AE pairs | – | m × n | 250 000 |
| Number of drug-ADR pairs | – | – | 250 |
| Number of innocent bystanders | – | – | 0, 125, or 250 |
| Conditional probability bystander | (3) | γ | .5, .75, or .9 |
| Mean OR of drug-ADR pair | (4) | OR | 1.5, 3, or 5 |

Of the 250 000 drug-AE pairs, 250 have a causal relationship. The goal of the statistical discovery methods is to identify these drug-

The prescription of the first 250 drugs, X1 to X250, is not influenced by the prescription of any other drug. The last 250 drugs, X251 to X500, can be innocent bystanders, for example, X251 can be the innocent bystander for the drug-ADR pair (X1, Y1). In that case, X251 is prescribed more regularly when X1 is prescribed than when not; the parameter γ in Equation (3) is set to .5, .75, or .9. The parameter γ is the same for all innocent bystanders. We consider three cases:

1. No innocent bystanders. All drugs X1, X2, ..., X500 are independent.
2. One hundred twenty-five innocent bystanders, that is, drug X251 is the innocent bystander for drug-ADR pair (X1, Y1). X252 is the innocent bystander for (X2, Y2), etc., up to X375 and (X125, Y125).
3. Two hundred fifty innocent bystanders, that is, drug X251 is the innocent bystander for drug-ADR pair (X1, Y1). X252 is the innocent bystander for (X2, Y2), etc., up to X500 and (X250, Y250).

Table 3 shows an overview of the parameter settings. There are 21 settings in total: three if there are no innocent bystanders, and 9 each if there are 125 and 250 bystanders. For each setting, 50 SRSs with N = 50 000 reports are generated, resulting in 1050 SRSs. Some descriptive statistics of the simulated data sets can be found in Appendix A.6. An implementation of this algorithm together with an example dataset is publicly available at www.github.com/bips-hb/srsim. See Appendix B for a description how the 2 × 2 tables are constructed on the basis of the simulated SRSs.

All 24 measures are applied to all 1050 simulated SRSs. Due to the sparse nature of SR data, it may happen that a measure is not defined for a particular drug-AE pair. These pairs are considered least likely to be associated. Drug-AE pairs that were reported less than four times are excluded. See Appendix D for the results without this restriction.

For each measure and each data set, we construct the precision-recall curve (PRC)27 that shows the changes in precision and recall when the threshold is varied from most stringent to most relaxed. See Appendix C for various PRC plots. The area under the curve (AUC) reflects the measures' overall capability to distinguish between associated and not-associated pairs. Although the receiver-operating curve (ROC) is more commonly used than the PRC, it is known to perform rather poorly when the data are imbalanced.27

4 | RESULTS

Figure 1 shows the boxplots of the resulting AUC scores for four parameter settings. The left column shows the results when the observed ORs of the drug-ADR pairs are relatively small (OR ≈ 1.5), and the right column when the ORs are high (OR ≈ 5). The top row shows the results in case there are no innocent bystanders, and the bottom row if there are 250 innocent bystanders and the conditional probabilities, see Equation (3), are set to γ = .9. The measures are ordered by their average AUC. In contrast to a ROC for which the baseline performance of a random classifier is .5, the baseline performance for a PRC is .001 in this particular case.27

The results show a clear difference in performance when the effect sizes are large (the AUC gets close to 90%) compared with relatively small effect sizes (just over 40%). The hypothesis tests (pPoisson, the χ² and the Fisher’s exact tests, and pBinomial) outperform the Bayesian measures when there is no confounding and the effect sizes are small. The measures based on the BCPNN (eg, IC05 alternative) catch up, however, when the effect size increases.

When the innocent bystander effect is introduced (see the lower boxplots in Figure 1) the ranking changes drastically. The BCPNN-based measures and the LASSO perform better than their competitors based on hypothesis tests. This difference becomes even more apparent when the effect sizes increase. Measures that do not account for sample size (the ROR, PRR, RRR, and the number of reports) perform poorly overall.

The variance of the observed AUCs for the measures based on the GPS (EBGM, EB025, and EB05) is larger than for the other measures. Perhaps this is because the hyperparameters of the prior are fitted to the data that allow for more variability.24

Figure 2 shows boxplots of the AUC scores of the pPoisson, IC05alternative and the LASSO when the ORs are high (OR ≈ 5) for all parameter settings. One can see similar patterns for lower average ORs (although less pronounced).

The test of the Poisson mean, pPoisson, is strongly influenced by the introduction of innocent bystanders. The performance drops from an average AUC of .94 (95% CI = ±2.6 × 10⁻²) when there is no confounding to .53 (95% CI = ±2.3 × 10⁻²) when there are 250 innocent bystanders with γ = .9. A similar stark decrease in performance is also seen for other measures that fall into the hypothesis test category.

Although the innocent bystander effect affects the performance of the IC05alternative, it does not influence it as strongly as for the test of the Poisson mean. In case there are no innocent bystanders, the average AUC is .94 (95% CI = ±2.5 × 10⁻²) against .81 (95%
CI = ± 4.7 × 10⁻²) when the innocent bystander effect is strongest. The same holds for the other Bayesian measures, which explains the change in ranking in Figure 1.

Even though the LASSO does not excel in any setting, its performance is striking since it is not affected by the appearance of innocent bystanders. By employing the data on all drugs simultaneously, it...
The boxplots of the areas under the precision-recall curves when the ORs of the drug-ADR pairs are strong (OR ≈ 5) for the test of the Poisson mean, $p_{\text{Poisson}}$, the BCPNN-based $IC_{05}$ alternative and the LASSO. The left-most column depicts the AUC distribution when there is no confounding by other drugs. The middle three show the results when there are 125 innocent bystanders with varying conditional probabilities, $\gamma$ (see Equation (3)). The last three show the results with 250 innocent bystanders.
successfully manages to distinguish between the drug-ADR pairs and the innocent bystanders.

The figures for the other parameter settings can be found online at https://srs.bips.eu.

5 | DISCUSSION AND CONCLUSIONS

In this article, we compared 24 statistical measures regarding their ability to discover drug-ADR pairs on the basis of spontaneous reporting data. These measures range from simple disproportionality measures, hypothesis tests, Bayesian shrinkage estimates, to sparse regression. This comparison differs from previous work not only in scope but also in that it is the first to explore to what extent confounding by other drugs, that is, the innocent bystander effect, affects the measures’ performance. In addition, by using the area under the precision-recall curve, the results do not depend on individual thresholds for each measure that could potentially skew the outcome.

Figure 1 makes it clear that we cannot point out an absolute winner. The simulation results suggest that hypothesis tests (especially the test of the Poisson mean) are to be preferred when the effect sizes are small and there is no or little confounding. The moment when the level of confounding increases and/or the effect sizes become larger, the measures based on the BCPNN perform best. Which measure to use for a particular SRS, thus, depends on the effect sizes and the level of confounding present in the data. Since it is less likely to discover drug-ADR pairs with an odds ratio around 3 or 5 that eluded detection before, it seems reasonable to take special note of the results for the lower effect sizes (OR ≈ 1.5). Due to their poor performance in this simulation study, the use of ROR, PRR, and RRR is discouraged. The only method that is robust against confounding by other drugs is the LASSO. Although it cannot boast the best overall performance, it is not confounded by the presence of innocent bystanders. It achieves this by considering all drugs simultaneously. It would be very interesting to see how other sparse regression approaches, for example, the elastic net, would perform.

All Bayesian methods such as the GPS and the BCPNN use the relative report rate as their basis. Figure 1 shows, however, that the RRR is consistently outperformed by the reporting odds ratio. It might be fruitful to explore the possibility of applying Bayesian shrinkage to the ROR rather than the RRR. By using a threshold-free performance measure, we avoided the problem of choosing an appropriate threshold for the various measures. For some, it might be easier to choose an appropriate threshold than for others. It is, for example, unclear how to set the thresholds for ROR, PRR, and RRR. For the measures based on hypothesis tests, one could employ a multiple testing correction procedure. In case of the Bayesian methods, there are similar procedures to control the false-discovery rate, for example, the study by Ahmed et al.

A point of caution is that the simulation setup is a simplification and might differ from the SRSs used in daily practice. Real data sets can contain more noise. The results shown here should, thus, be seen as a best-case scenario. In addition, one should not rely solely on statistical analysis. Clinical and pharmacological knowledge is essential when identifying drug-ADR pairs.

The implementation of the measures, the SR simulator, and the code for this comparison study are publicly available as R packages at www.github.com/bips-hb/pvm, www.github.com/bips-hb/srsim and www.github.com/bips-hb/pvcomparison.

ETHICS STATEMENT

The authors state that no ethical approval was needed.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Vanessa Didelez for her input and fruitful discussions. We gratefully acknowledge the financial support from the innovation fund ("Innovationsfonds") of the Federal Joint Committee in Germany (grant number: 01VSF16020).

CONFLICT OF INTEREST

The authors have no conflict of interest to declare relevant for this study.

ORCID

Louis Dijkstra  https://orcid.org/0000-0003-1476-6202

REFERENCES

1. Norén GN, Bate A, Orre R, Edwards IR. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. Stat Med. 2006; 25(21):3740-3757.
2. Onakpoya UJ, Heneghan CJ, Aronson JK. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. BMC Med. 2016;14:10. https://doi.org/10.1186/s12916-016-0553-2.
3. WHO. The importance of pharmacovigilance: safety monitoring of medicinal products. Technical report. 2002. http://apps.who.int/medicinedocs/pdf/s4893e/s4893e.pdf. Assessed 23 Aug 2018.
4. Bailey C, Peddie D, Wickham ME, et al. Adverse drug event reporting systems: a systematic review. Br J Clin Pharmacol. 2016;82(1):17-29.
5. Huang YL, Moon J, Segal JB. A comparison of active adverse event surveillance systems worldwide. Drug Saf. 2014;37(8):581-596.
6. Routledge P. 150 years of pharmacovigilance. Lancet. 1998;351(9110):1200-1201.
7. Suling M, Pigeot I. Signal detection and monitoring based on longitudinal healthcare data. Pharmaceutics. 2012;4(4):607-640.
8. Caster O, Madigan D, Norén GN, Bate A. Large-scale regression-based pattern discovery in international adverse drug reaction surveillance. Proc KDD Work min Med Data KDD Cup. 2008;24-27.
9. Tibshirani R. Regression shrinkage and selection via the Lasso. Stat Med. 1996;58(1):267-288.
10. van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Eggerts ACG. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11(3):1-10.
11. 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):483-486.
12. Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res. 2013;22(1):39-56.
13. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54(4):315-321.
14. Almenoff JS, DuMouchel W, Kindman LA, Yang X, Fram D. Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. *Pharmacoepidemiol Drug Saf*. 2003;12(6):517-521.

15. Candore G, Juhlin K, Manlik K, et al. Comparison of statistical signal detection methods within and across spontaneous reporting databases. *Drug Saf*. 2015;38:577-587.

16. 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*. 2014;93(6):1-19.

17. Pacurariu AC, Straus SM, Trifîrò G, et al. Useful interplay between spontaneous ADR reports and electronic healthcare records in signal detection. *Drug Saf*. 2015;38(12):1201-1210.

18. Ahmed I, Haramburu F, Fourrier-Réglat A, et al. Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. *Stat Med*. 2009;28(13):1774-1792.

19. Rolka H, Bracy D, Russell C, Fram D, Ball R. Using simulation to assess the sensitivity and specificity of a signal detection tool for multidimensional public health surveillance data. *Stat Med*. 2005;24(4):551-562.

20. Li C, Xia J, Deng J, Jiang J. A comparison of measures of disproportionality for signal detection on adverse drug reaction spontaneous reporting database of Guangdong province in China. *Pharmacoepidemiol Drug Saf*. 2008;17(6):593-600.

21. Roux E, Thiessard F, Fourrier A, Bégaud B, Tubert-Bitter P. Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. *IEEE Trans Inf Technol Biomed*. 2005;9(4):518-527.

22. Norén GN, Caster O, Juhlin K, Lindquist M. Zoo or Savannah? Choice of training ground for evidence-based pharmacovigilance. *Drug Saf*. 2014;37(9):655-659.

23. Stricker BH, Tijssen JG. Serum sickness-like reactions to cefaclor. *J Clin Epidemiol*. 1992;45(10):1177-1184.

24. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat*. 1999;53(3):177-190.

25. Ahmed I, Dalmasso C, Haramburu F, Thiessard F, Broët P, Tubert-Bitter P. False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics*. 2010;66(1):301-309.

26. Lian Duan L, Khoshneshin M, Street WN, Liu M. Adverse drug effect detection. *IEEE J Biomed Health Inform*. 2013;17(2):305-311.

27. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):1-21.

28. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Statistical Methodol*. 2005;67:301-320.

29. Deshpande G, Gogolak V, Smith SW. Data mining in drug safety: review of published threshold criteria for defining signals of disproportionate reporting. *Pharmaceut Med*. 2012;24(1):37-49.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Dijkstra L, Garling M, Foraita R, Pigeot I. Adverse drug reaction or innocent bystander? A systematic comparison of statistical discovery methods for spontaneous reporting systems. *Pharmacoepidemiol Drug Saf*. 2020;29:396–403. [https://doi.org/10.1002/pds.4970](https://doi.org/10.1002/pds.4970)