

REVIEW

Methods for drug safety signal detection using routinely collected observational electronic health care data: A systematic review

Astrid Coste¹  | Angel Wong¹ | Marleen Bokern¹ | Andrew Bate^{1,2}  | Ian J. Douglas¹

¹Department of Non-Communicable Disease Epidemiology, LSHTM, London, UK

²Global Safety, GSK, Brentford, UK

Correspondence

Astrid Coste, Department of Non-Communicable Disease Epidemiology, LSHTM, London, UK.
Email: astrid.coste@lshtm.ac.uk

Funding information

GlaxoSmithKline

Abstract

Purpose: Signal detection is a crucial step in the discovery of post-marketing adverse drug reactions. There is a growing interest in using routinely collected data to complement established spontaneous report analyses. This work aims to systematically review the methods for drug safety signal detection using routinely collected health-care data and their performance, both in general and for specific types of drugs and outcomes.

Methods: We conducted a systematic review following the PRISMA guidelines, and registered a protocol in PROSPERO. MEDLINE, EMBASE, PubMed, Web of Science, Scopus, and the Cochrane Library were searched until July 13, 2021.

Results: The review included 101 articles, among which there were 39 methodological works, 25 performance assessment papers, and 24 observational studies. Methods included adaptations from those used with spontaneous reports, traditional epidemiological designs, methods specific to signal detection with real-world data. More recently, implementations of machine learning have been studied in the literature. Twenty-five studies evaluated method performances, 16 of them using the area under the curve (AUC) for a range of positive and negative controls as their main measure. Despite the likelihood that performance measurement could vary by drug-event pair, only 10 studies reported performance stratified by drugs and outcomes, in a heterogeneous manner. The replicability of the performance assessment results was limited due to lack of transparency in reporting and the lack of a gold standard reference set.

Conclusions: A variety of methods have been described in the literature for signal detection with routinely collected data. No method showed superior performance in all papers and across all drugs and outcomes, performance assessment and reporting were heterogeneous. However, there is limited evidence that self-controlled designs, high dimensional propensity scores, and machine learning can achieve higher performances than other methods.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Pharmacoepidemiology and Drug Safety* published by John Wiley & Sons Ltd.

KEYWORDS

drug safety surveillance, pharmacoepidemiology, pharmacovigilance, real world data, signal detection, systematic review

Key Points

- There has been a growing interest in the last 15 years to use routinely collected data to complement spontaneous reports for drug safety signal detection.
- This is the first systematic review including 101 studies, which quantified the use of a wide variety of methods for drug safety signal detection with routinely collected data and assessed their comparative performance. While self-controlled methods performed overall well there were no direct comparisons of all approaches in the 25 performance assessment studies.
- Transparency, replicability and, due in part to the lack of a gold standard reference set, comparability between studies was limited.
- Although the suitability of epidemiological methods varies by nature of exposure and outcome, stratified performance was only available in 9.9% of studies, adding difficulty to the identification of useful methods for signal detection.

1 | INTRODUCTION

Signal detection is the process of identifying emerging true associations as early as possible, ideally leading to further action while effectively avoiding false positives. For decades, spontaneous reports (SRs) have been the primary approach for detecting adverse drug reactions (ADRs) not picked up in clinical trials,¹ and remain so despite their well-recognized limitations.^{2,3} There is a growing interest in using real-world data (RWD), including claims data and electronic health records (EHRs). Their potential for signal detection has been recognized as a hope for potentially faster and more efficient post marketing surveillance.⁴ Several initiatives have provided methodological input for drug safety signal detection using RWD^{5,6} and have evaluated the performance of various methods against a set of positive and negative controls.

Methods for signal detection with RWD were reviewed by Arnaud et al.⁷ until 2016, focusing on both their overall performance regardless of types of drugs and outcomes and secondly understandability by stakeholders. However, epidemiological methods are differentially valid depending on the nature of the drug and outcome studied, and a single method applied to a wide range of drugs and outcomes without consideration of its optimal application could lead to poor detection⁸ It is therefore useful to explore whether this issue has been considered in signal detection, or whether a one fits all approach has been largely used for simplicity. Further, novel methods have also been developed since this review.^{9,10}

Therefore, this systematic review aimed to: (1) update the list of methods for drug safety signal detection using routinely collected data and quantify the extent of their published; (2) summarize and compare methods performance regarding ability to detect signals in routinely collected observational data; and (3) assess the performance of each method for specific types of exposures and outcomes.

2 | METHODS

2.1 | Search strategy

The systematic review was conducted following the protocol registered at PROSPERO (registration number CRD42021267610). We searched MEDLINE and EMBASE via OVID, Web of science, Scopus, PubMed, and the Cochrane Library with no restriction on the period on July 13, 2021.

Keywords and Medical Subject Headings (MeSH) based on (1) routinely collected data, (2) pharmacoepidemiology or drug safety. and (3) signal detection were used (Appendix S1). The reference lists from identified literature reviews were screened to identify additional works.

Included studies were (1) describing an epidemiological study design or statistical method for signal detection using routinely collected observational data; (2) evaluating their performance; or (3) applying these methods to screen drug-outcome pairs. We excluded studies relying on free-text data because methods mainly rely on natural language processing which are different to that used for structured data¹¹ as well as conference abstracts. The original protocol was modified by not including vaccine related studies as methods for vaccine signal detection have their specific limitations and different considerations from other medications.¹²

Articles were firstly screened by title and abstracts, followed by a full-text evaluation for eligible papers. A second reviewer assessed all the included publications and a sample of the excluded ones. Any disagreement was resolved by discussion.

2.2 | Data extraction

We extracted data based on the RECORD Pharmacoepidemiology Checklist,¹³ focusing on the details of the methods: design, statistical

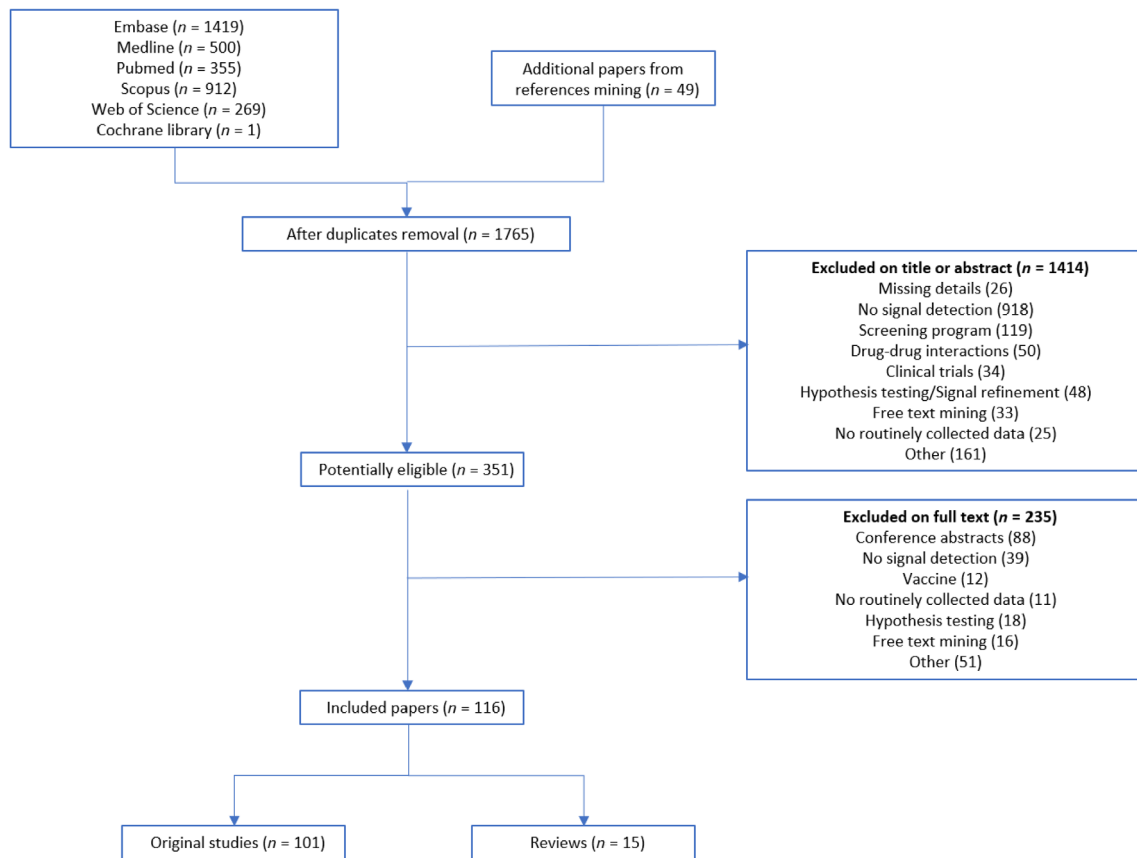


FIGURE 1 Flowchart of inclusion

outputs; exposure(s), outcome(s), results, and performances of the methods.

The risks of bias and confounding, the appropriateness of the ADR testing and the degree to which the database captures outcomes were also assessed.

2.3 | Data analysis

The characteristics of the included studies and the methods for drug safety signal detection were reported. Methods for drug safety signal detection using RWD were described and the number of times they were used was quantified. The performance of these methods was assessed using measures presented in the literature, both in general for all drug/outcome pairs and by drug and outcome when this was available.

3 | RESULTS

3.1 | Studies identified

We screened 1765 titles and abstracts. After applying inclusion and exclusion criteria, 351 papers were classified as potentially eligible

(Figure 1). Of those, 116 relevant studies were included in the review, with 101 original studies and 15 reviews.

Of the included studies, 38.6% purely described methods (Table 1), 24.8% were about performance assessment and 23.8% were observational studies without performance assessment. Among the studies, 5.9% of them compared the use of EHRs and SRs for signal detection.^{1,14-18} The remaining 6.9% included a recent PhD thesis,¹⁹ two commentaries,^{20,21} a study aiming to establish a reference standard for signal detection²² and 3 studies looking at the significance of signal detection results.²³⁻²⁵ Most studies (88.1%) used traditional EHRs or claims data, while 6.9% used abnormal laboratory results^{9,26-31} and a prescription only dataset (5%) where prescriptions are used as proxies for diagnoses.³²⁻³⁶ The aim of our systematic review was to identify original research and so any review articles we identified within scope were only used to provide potential further original research publications for inclusion and their contents were not extracted.^{4,7,23,37-47} A third of the studies were published after 2016, year of the latest review on the topic, as shown on Figure 2.

3.2 | Quality assessment

There are no standard criteria to assess the quality of signal detection studies beyond general quality assessment tools and guidance for

RWD studies. Often, the definitions for the chosen drugs and outcomes were not specified, and specific implementation in the databases was rarely specified. The codes and code lists were rarely made available. Notably, the Observational Medical Outcome Partnership

TABLE 1 Summary characteristics of the publications included in the review

| Characteristic | Number of publications |
|---|------------------------|
| Primary objective of the paper | |
| Method description | 39 (38.6%) |
| Performance assessment | 25 (24.8%) |
| Data comparison between EHRs and SRs | 6 (5.9%) |
| Application of method without performance assessment | 24 (23.8%) |
| Other ^a | 7 (6.9%) |
| Location of data | |
| United States | 43 (42.6%) |
| Europe | 37 (36.7%) |
| Asia/Australia | 18 (17.8%) |
| International | 3 (3.0%) |
| Approach | |
| Outcome based | 40 (39.6%) |
| Exposure based | 26 (25.7%) |
| Both drugs and outcomes specified | 6 (5.9%) |
| All drugs and outcomes in the database(s) | 4 (4.0%) |
| None (purely methodological) | 25 (24.8%) |
| Type of data used by the method | |
| Method based on prescription and diagnoses codes | 89 (88.1%) |
| Method based on prescription data only | 5 (5.0%) |
| Method based on the comparison of laboratory test results | 7 (6.9%) |

^aOther = PhD thesis, commentaries, reference standard.

(OMOP) initiative has now switched to the Observational Health Data Sciences and Informatics (OHDSI), so that previous OMOP reports are not publicly available on the website as of 1st June 1, 2022. This limits the reproducibility of some included studies. Other more recent studies published Supporting Information (Supplement S1 such as details on outcome definition or on performance results.^{48,49}

3.3 | Methods for drug safety signal detection

A wide range of methods were described in the included studies, and are summarized in Tables 2 and 3 following a classification used by Arnaud et al.⁷ Overall, the literature focussed on adapting disproportionality analysis methods to signal detection and implementing traditional epidemiological designs. Other methods, using Bayesian network models, the Weibull shape parameter or likelihood ratio tests were proposed in methodological papers but used in a single or no observational study so are not included in the following tables.^{102,103}

3.4 | Performance of the methods

Performance was defined implicitly across papers as the ability of a method to correctly detect signals among a set of positive (well-established drug-outcome associations) and negative controls (drugs known not to cause certain outcomes).¹⁰⁴

Among the 25 performance assessment papers, 19 reported quantitative measures of performance, 6 only reported qualitative results. Such measures included the area under receiver operating the curve (AUC) in 16 of the 19 studies (84.2%), an estimate of predictive accuracy. It ranges between 0 and 1, the latter corresponding to a perfect prediction of positive controls. A value of 0.5 is identical to random guessing.¹⁰⁵ The specificity, sensitivity^{8,36,48,49,70,81,106,107} and coverage probability (proportion of the 95% confidence interval estimates that included the true parameter value, being 1 for negative controls)^{48,49,52,55,60,108,65,107} were each used in eight papers (42.1%).

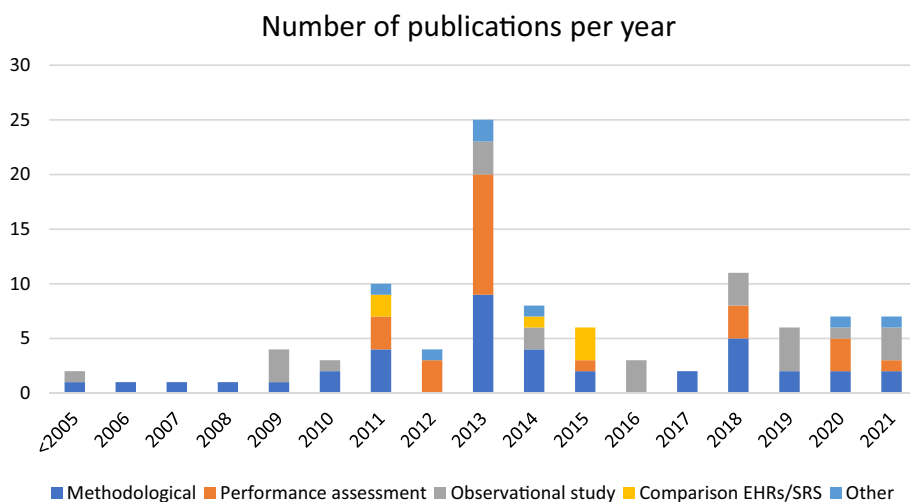


FIGURE 2 Number of studies by year. “Observational study” in the graph refers to the category “application of method without performance assessment” in Table 1.

TABLE 2 Number of times each method applied twice or more was used across the publications of the review

| Method | Number of papers using the design ^a |
|---|--|
| Disproportionality analysis | |
| PRR | 9 (17.3%) |
| ROR | 8 (15.4%) |
| BCPNN | 9 (17.3%) |
| GPS/MGPS | 6 (11.5%) |
| LGPS/LEOPARD | 12 (23.1%) |
| Other | 8 (15.4%) |
| Subtotal | 52 (100.0%) |
| Traditional epidemiological designs | |
| Self-controlled case series | 15 (34.1%) |
| Self-controlled cohort | 5 (11.4%) |
| New-user cohort | 5 (11.4%) |
| Case-control | 13 (29.5%) |
| Case-crossover | 3 (6.8%) |
| Case-population | 3 (6.8%) |
| Subtotal | 44 (100.0%) |
| Temporal association | |
| Temporal pattern discovery | 10 (50.0%) |
| MUTARA/HUNT | 6 (30.0%) |
| Fuzzy-based logic | 4 (20.0%) |
| Subtotal | 20 (100.0%) |
| Sequence symmetry analysis | 6 (100.0%) |
| Sequential testing | |
| MaxSPRT | 4 (66.7%) |
| CSSP | 2 (33.3%) |
| Subtotal | 6 (100.0%) |
| Tree-based scan statistic | 9 (100.0%) |
| Other designs including machine learning | 13 (100.0%) |
| Lab results | 9 (100.0%) |
| Prescription only methods | 5 (100.0%) |

Abbreviations: BCPNN, Bayesian Confidence Propagation Neural Network; CSSP, Conditional Sequential Sampling Procedure; GPS, Gamma Poisson Shrinker; HUNT, Highlighting Unexpected TARs Neglecting TARs; LEOPARD, Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs; LGPS, Longitudinal Gamma Poisson Shrinker; MaxSPRT, Maximized Sequential Probability Ratio Test; MGPS, Multi-Item Gamma Poisson Shrinker; MUTARA, Mining Unexpected Temporal Association Rules (TARs) Given the Antecedent; PRR, Proportional Reporting Ratio; ROR, Reporting Odds Ratio.

^aStudies exploring more than one method were counted for each of the methods they considered, so that the total number of papers in this table does not correspond to the number of included studies.

The average precision^{70,72,106} was used in 4 studies (21.1%). The mean squared error^{48,49,105} and negative and positive predictive values^{36,106,107} were used in 3 of the 19 papers (15.8%) each, whilst the bias,¹⁰⁵ partial area under the curve at 30% false-positive rate (PAU30)¹⁰⁶ and recall at 5% false-positive rate¹⁰⁶ have each been reported in two studies or less.

Fifteen of the 25 studies used datasets from three main projects, which aimed to assess the performance of methods for drug safety signal detection (Figure 3 and Table 4). Each study tested up to 126 unique parameter combinations.

Not all methods had their performance assessed. Disproportionality-based methods, traditional epidemiological designs and Temporal Pattern Discovery (TPD) were evaluated in three or more papers. The tree-based scan statistic, sequential analysis, the Mining Unexpected Temporal Association Rules (TARs) Given the Antecedent (MUTARA) and Highlighting Unexpected TARs Neglecting TARs (HUNT) algorithms were assessed in two papers or less, with few to no head-to-head method comparison. Some studies describing machine learning frameworks also computed measures of performance as a secondary objective, using test sets as reference standards.^{9,34,94,98,101}

Seven studies presented AUC values for >1 method and a large range of drug-outcome pairs, typically >50 pairs (Table 5). The average AUCs across all pairs and databases were as low as 0.47–0.50, below random guessing, for the New-User Cohort and Bayesian Confidence Propagation Neural Network (BCPNN), a disproportionality-based method. The maximum AUC was 0.81 for the Self-Controlled Cohort. Overall, self-controlled methods achieved higher AUCs than other methods.⁴ The High Dimensional Propensity Score (HDPS) method, used in conjunction with a new user, active comparator design achieved the highest AUCs in two papers.^{66,106} TPD had higher AUCs than other methods in all studies except one,⁷⁰ whilst MUTARA and HUNT had lower than average AUCs (0.57–0.60).⁷⁰ The Maximized Sequential Probability Ratio Test (MaxSPRT) and the Conditional Sequential Sampling Procedure (CSSP) had low reported AUCs in the 2011 OMOP report, in the range of 0.23–0.38.⁶⁶

Among Machine Learning (ML) techniques, which were not evaluated within the seven studies above, the supervised Bradford Hill had a reported AUC of 0.86,⁹⁴ which is the highest reported average AUC among all performance assessment papers. The Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs (LEOPARD) algorithm was found to improve the average AUCs of all methods in one study⁵³ when applied to OMOP methods. A lack of differential performance between methods was observed in several papers.^{53,70}

One paper evaluated the performances of different algorithms for laboratory-based signals. ML models achieved the highest AUCs (0.80–0.82), the Comparison of Extreme Laboratory Test results (CERT) and Prescription pattern Around Clinical Event (PACE) algorithms had AUCs in the range of 0.52–0.56, and disproportionality methods were the lowest performing, with AUCs of 0.52–0.56.⁹

3.5 | Performance stratified by drug or outcome

From Table 3, no method can theoretically perform equally well for all drugs and outcomes (e.g., some methods are more suited to acute or rare outcomes). The average AUC discussed above is not representative of the full potential of a method as it represents average

TABLE 3 Overview of methods for drug safety signal detection using RWD mentioned or applied in more than one study as reported in the included papers

| Method | Stated general concept | Reported advantages | Reported weaknesses | Additional comments |
|--|---|---|---|--|
| Disproportionality analysis (DP) | | | | |
| PRR and ROR | Originally applied to SRs to determine the degree of disproportionality between the reporting of a condition for a given drug. ⁵⁰ Focus on a 2×2 contingency table to compare the observed number of records to an expected number of records. ^{51,52} | Easy to calculate and to implement. | Unstable with small number of events and large confidence intervals leads to high false-positive rates for rare events. ⁴⁵ Designed for cross sectional data, so the total number of cases available in EHR databases is not used. | |
| Information Component (BCPNN) | A fully Bayesian (predefined prior) disproportionality method using shrinkage of observed-to-expected (O/E) scores to calculate an Information Component (IC). ⁵³ Sometimes implemented in a neural network (BCPNN) where the IC is the weight. | Addresses the high false-positive rate issue of the SR-like methods by greater shrinkage when little data support reducing spurious chance findings. ⁵⁰ | As above, in addition harder interpretability as impact of Bayesian shrinkage adds additional complexity to result interpretation. | |
| GPS/MGPS | GPS is another Bayesian method for shrinkage of the observed/expected ratio. ⁵⁴ MGPS is an extension for analysis of drug-drug-event interactions. ⁴⁵ | Same as BCPNN, empirical shrinkage so shrinkage strength adapts to specific data set. | Primarily same as BCPNN, and impact of shrinkage variable. Computationally intensive. | |
| LGPS | Adaptation of the GPS method to longitudinal data. ⁵³ LGPS compares the incidence rate of outcome during exposure risk period to the background rate for all people. | Has been used in conjunction with the LEOPARD method, to theoretically handle protopathic and indication biases. LEOPARD compares the rate of prescription prior to the events to that after. ⁵⁴ | Impact of Bayesian prior as above. | |
| Traditional epidemiological designs | | | | |
| Self-controlled case series (SCCS) | Comparison of the event rate during exposed and unexposed time within the same individual. ⁵⁵ | Eliminate time invariant confounders, advantageous when baseline covariates not measured with sufficient precision. Cases only: computational savings. ⁵⁵ | Accurate dating of outcomes is crucial, best applies to intermittent exposures and transient or acute events. ⁵⁶ | Several further modifications have been proposed, but not yet implemented in signal detection studies. ^{10,57,58} |
| Self-controlled cohort (SCC) | Comprising both a cohort and self-controlled adjustment. | Utilizes an external control group to adjust for remaining time-varying confounding after the self-controlled component. ⁸ | Sensitive to differences between risk and control period specific to the exposed group, such as protopathic bias. ⁷ | |
| New user cohort (NUC) | Compares the rate of events in a cohort initiating the drug of interest versus a cohort not initiating this drug. ^{59,60} | Broadly applicable method, ⁵⁹ active comparator approach to address confounding by indication. ⁶⁴ Computation of the absolute risk of events. ⁷ | Higher computational requirements than self-controlled methods. Between person confounding. Need for a predefined comparator(s) as its appropriateness is difficult to assess in real-world settings. | |

TABLE 3 (Continued)

| Method | Stated general concept | Reported advantages | Reported weaknesses | Additional comments |
|--|---|--|---|---------------------|
| Case-control (CC) method | Compares the frequency of exposure of "cases" who experienced the outcome with that of matched "controls" who did not experience the outcome. ⁶¹ | Effective for rare outcomes. ⁶⁵ | Possibly challenging control selection, susceptible to between person confounding. High computational requirements, slowest method in OMOP. ⁶⁶ | |
| Case-crossover | Uses within-person comparison of exposure in the case period compared to that in the control period. ⁶² | Similar to other self-controlled designs. | Subject to bias when exposure time trend is present. ⁶² | |
| Case-population | Similar to case-control design, but using the entire population as control group. ⁶³ | Increased statistical precision compared to CC. ⁶³ | Higher computational requirements. | |
| Temporal association | | | | |
| Temporal pattern discovery (TPD) | Based on the observed-to-expected ratio from DP, adding a comparison of the exposed time with a control period prior to first drug exposure to identify a temporal association. ⁶⁷ Similar to LGPS. | Bayesian shrinkage to protect against spurious associations, especially for rare events. ⁶⁸ Use of chronographs for graphical interpretation, possibility to detect a greater variety of patterns. ⁶⁷ | More difficult to identify patterns related to very common events ⁶⁷ | |
| MUTARA and HUNT | MUTARA is a data-mining algorithm looking for Unexpected Temporal Association Rules by finding any events occurring unexpectedly after the drugs of interest within a predefined risk period and excluding common events unlikely to be ADRs using a reference period to shortlist important ADRs. Treatment failure can lead to recording of events after the drug of interest, causing spurious signals. HUNT, a modified version of MUTARA, re-ranks signals by taking account of treatment failure ^{69,70} | Since MUTARA is event-orientated, it should have the theoretical ability to signal infrequent ADRs ⁷¹ | MUTARA had difficulty distinguishing between adverse events and treatment failures. ⁷² HUNT was developed to consider these treatment failures. | |
| Fuzzy-based logic with causal-leverage measure | This is a computation model called fuzzy recognition-primed decision model. It uses cues such as temporal association between drugs and outcomes, rechallenge, dechallenge to capture potential causal associations within each drug-exposed person. The potential causalities are then incorporated into new measures (causal-leverage and reverse causal value), which were used to rank drug-outcome pair and remove spurious signals. ^{73,74} | Attempt to rank the important signals by quantifying the degree of association of a drug-outcome pair and remove background noise (spurious signals) using new measures. Developed specifically for infrequent associations. ⁷⁵ | The parameter, e.g. hazard period to capture outcomes, and fuzzy variables (e.g. temporal association, rechallenge and dechallenge) could be difficult to optimise. | |

(Continues)

TABLE 3 (Continued)

| Method | Stated general concept | Reported advantages | Reported weaknesses | Additional comments |
|--|--|---|--|--|
| Sequence Symmetry Analysis (SSA) | Investigate the sequence of events related to the initiation of a drug. Testing for asymmetry based on the null hypothesis that if there is no association, one would expect a symmetrical distribution of the outcome before and after the initiation of a drug. ⁴¹ | Computationally efficient, robust to confounders stable over time. ⁴¹ Possible to evaluate all drugs and all outcomes in a database. ⁷⁶ | A non-symmetrical pattern does not necessarily indicate a signal. ⁴¹ Triage needed to find potentially interesting associations, high number of false negatives. Assumes an appropriate single point of symmetry. | Mainly applied in large-scale settings in the literature. ^{32,77–80} |
| Sequential testing: MaxSPRT, CSSP, Log Linear Model for Poisson Data (LLMP) | Suite of methods for near-real time or prospective surveillance, building on repeated testing to detect associations. Patients are added gradually. ⁸¹ The maxSPRT is a statistical test adjusting for multiple testing. ^{82,83} The CSSP method uses a conditional probability of having an outcome more extreme than the observed event rate and stratifies the population. ⁸¹ The LLMP is a log-linear Poisson based method. | Ability to handle rare events. ⁸¹ Near real-time surveillance ⁸³ makes early detection of new ADRs possible, improving the timeliness of signal detection. ⁸⁴ CSSP has been developed to handle chronic exposures. ⁸⁵ | The maxSPRT requires large amount of historical data to have precise estimates of the expected number of events. This problem is handled in the CSSP method. ⁸⁶ CSSP may struggle maintaining the type I error with frequent testing or numerous strata. ⁸⁵ Limited for handling continuous confounders. ⁸¹ | |
| Tree-based scan statistic | Health events are classified in a tree hierarchical system, based on diagnoses codes at different levels. Data mining technique looks for excess risk in individual cells of the tree as well as in closely related cells. ⁸⁷ Hypothesis testing is performed using the likelihood ratio test. ⁸⁸ | Simultaneous scan for signals at different levels of granularity. Minimum prior assumption about the type of health outcome of interest. ⁸⁸ Adjust for multiple hypothesis testing. ⁸⁸ | Potential concern of false-negative signals due to the control of the type I error. ⁸⁹ | Has been flexibly applied with a new user, active comparator design, propensity scores techniques ^{64,90,91} and self-controlled designs ^{92,93} |
| Other machine learning (ML) based approaches | Different approaches have been proposed. Some aim to train a classifier with a range of positive and negative controls. ⁹⁴ Others use data driven strategies to look for apparent outliers. ^{67,95} | Ability to cope with large, high-dimensional and sparse data, ⁹⁶ although computational performance can be decreased. | Need for a large training dataset. Less easily implemented by non-specialists of ML. Relies on the ability of the learning stage to deal with noisy healthcare data. | Used in different ways for signal detection ^{9,34,36,94,96–101} including propensity score calculation ⁹⁹ or as a whole approach using Bradford Hill's causality considerations. ⁹⁴ |
| Other | Variation of the SSA where drugs are taken as proxies for the Health Outcomes of Interest (HOIs). ^{32,35} Methods compared abnormal or extreme laboratory results before and after exposure to a drug. ³¹ | | | ML has also been used for prescription data only. ^{34,36} ML was used to combine the Comparison of Extreme Laboratory Test results (CERT), the Comparison of Extreme Abnormality Ratio (CLEAR) and the Prescription pattern Around Clinical Event (PACE) algorithms in a single value. ⁹ |

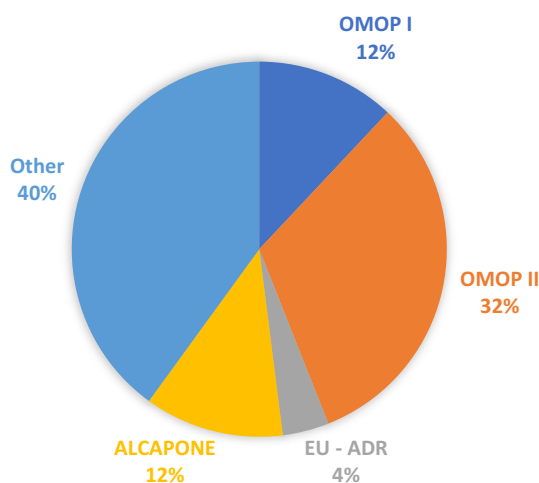


FIGURE 3 Proportion of the 25 performance assessment papers which used one of the main reference sets described in Table 4.

performance across all drug-outcome pairs. In this section, we aim to investigate the performance of the methods for specific types of drugs and outcomes. Only 8 of the 19 quantitative performance assessment papers reported performance measures stratified by type of outcome, and one proposed an analysis per drug. One additional paper discussed stratified results qualitatively. Overall, they were not consistent in their approach.

We accessed one OMOP report which classified drugs and outcomes in subgroups,⁶⁶ including: (1) high and low prevalence drugs and events, (2) acute and non-acute time to event, (3) long and short exposure. The results were presented for a single database. Altogether, methods had higher AUCs with high prevalence Drugs of Interest (DOIs), except for case-control and case-crossover. DP performed better with high prevalence DOIs than with low prevalence ones and had a high false-positive rate with common outcomes. CSSP and MaxSPRT achieved AUCs below 0.5 for all subgroups.

Other studies provided AUC values for each of the 4 OMOP outcomes (Table 6). The Alert generation using the case-population approach (ALCAPONE) project studied the performance of case-based designs for upper gastrointestinal (GI) bleeding and acute liver injury. They achieved higher AUCs for acute liver injury than in OMOP. DP methods were either close or even below random guessing for different outcomes. Self-controlled designs were consistently the best analytic choice for all databases and all outcomes in OMOP, except in one database, where TPD lead to the highest AUC for acute MI and upper GI bleed.¹⁰⁵

In Zhou et al.,⁵⁶ the Self-Controlled Case Series (SCCS) was able to highlight all acute events of interest in the primary analysis, such as fractures or GI perforation, including some outcomes that were not explored in other projects. Regarding slower onset outcomes, two were not highlighted but the association between adalimumab and lymphoma was signaled.

Several studies explored slower onset outcomes including cancer using a case-control design^{111–114} and one paper used a case-crossover design¹¹⁵ but none reported performance. Kulldorff et al.⁸⁸

mentioned the possibility to use the tree-based scan statistic with chronic events but this has not been tested so far.

According to several studies,^{70,72} many methods achieved low performances with rare ADRs. A study⁷² found that all MUTARA, HUNT, and reporting odds ratio (ROR) did not achieve a higher mean average precision (MAP) than 0.03 when restricted to rare ADRs, compared to MAPs ranging from 0.04 to 0.09 for all outcomes.

Only one performance assessment paper took a drug-based approach, investigating 6 drug families with various lengths of treatment (short vs. long term). They computed TPD, HUNT, MUTARA and ROR. However, no differential pattern of performance was observed.⁷⁰

4 | DISCUSSION

4.1 | Principal findings

There is an increasing interest in implementing RWD in signal detection (Figure 2) and several major initiatives have contributed to advances in methods development and performance assessment. However, performance assessment was heterogeneous, with a lack of agreement on the definition of a gold standard and what good performance looks like, making comparison difficult across methods, studies and data sources.

4.2 | Overall performance

Overall, the self-controlled methods tended to achieve higher AUCs than other methods, including case-control and disproportionality ones. The results were consistent across several OMOP papers and their replication in Europe. The HDPS and TPD methods also achieved higher AUCs, both on average and in certain subgroups. However, they were not evaluated in many studies and their running time was longer than for other methods.⁶⁶ Disproportionality methods, widely used in SRs, seem not to be able to distinguish between positive and negative controls as they had reported AUCs close to random guessing.⁵² This result was anticipated as SRs have different properties to that of RWD. Although the tree-based scan statistic did not undergo a formal performance assessment, it was able to capture known signals⁸⁸ and could be useful for assessing outcomes at different levels of granularity, particularly in a drug-based approach. Similarly, performance of ML has been evaluated heterogeneously, but preliminary results highlight its potential for signal detection.

Performance measures were generally reported on average across all drugs and outcomes in the reference set, even though every epidemiological study design performs better with some exposure and outcome types than others. Therefore, reported overall performance could hide particularly strong or weak performance for sets of similar exposure-outcomes combinations.

Performance was mainly assessed and presented with the AUC, which is a single measure and does not incorporate aspects such as bias.¹¹⁶ It assumes that every threshold of sensitivity and specificity is

TABLE 4 Summary of the most commonly used reference sets investigating performance of drug safety signal detection methods

| Name | OMOP | EU-ADR (Exploring and Understanding ADRs by integrative mining of clinical records and biomedical knowledge) | ALCAPONE (Alert generation using the case-population approach) |
|-------------------------|--|--|---|
| Experiment | OMOP I | OMOP II | |
| Dates | 2008–2013 | 2008 - 2012 | 2016 - 2018 |
| Country of data origin | US | Europe (4 countries) | France |
| Database(s) | 6 administrative claims (4 commercial) and 4 EHRs | 4 administrative claims and 1 EHR | Système National des Données de Santé (SNDS) (administrative claims) |
| Size of the database(s) | 130 million | 75 million | 65 million |
| Aims | To test a range of methods for drug safety signal detection and determine the best strategy to implement an active drug surveillance program. | To design and assess a system to exploit EHR data for the early detection of ADRs ¹⁰⁹ | Comparing and calibrating case-based methods for signal detection |
| Outcomes | Angioedema, Aplastic anaemia, Acute Liver injury, Bleeding, Myocardial infarction (MI), Hip fracture, Mortality after MI, Renal failure, Gastrointestinal (GI) Ulcer Hospitalization | Acute liver failure, Acute MI, Acute renal failure, Upper GI bleeding | Acute Liver Failure, Acute MI, Acute Renal Failure, Upper GI bleeding |
| Reference set | 53 drug-outcome pairs, 9 positive and 44 negative controls | 399 drug-outcome pairs, 165 positive controls and 234 negative controls | Combination of the OMOP II and the EU-ADR reference set, 273 drug-outcome pairs |
| Metrics | AUC Positive predictive value Sensitivity/Specificity Partial AUC at 30% false-positive rate Recall at 5% false-positive rate Average Precision | Average AUC Mean squared error “Bias” as the average difference between the log relative risk and zero for negative controls | AUC Mean squared error Coverage probability Calibrated p-values |

TABLE 5 Average AUC for each method and different publications

| Reference | Countries | Databases | SCCS | SCC | CC | CCO | DP | HDPS | TPD | LGPS ^a | NUC | PRR | ROR | BCPNN | GPS/MGPs |
|--------------------------------|-----------------------------|----------------------------|------|------|------|------|------|------|------|-------------------|------|------|------|-------|----------|
| Ryan et al. ¹⁰⁶ | United States | 7 from OMOP | 0.74 | 0.68 | 0.62 | 0.66 | 0.68 | 0.77 | 0.73 | 0.59 | 0.59 | | | | |
| Schuemie et al. ¹¹⁰ | Denmark, Italy, Netherlands | 6 from EU-ADR | 0.67 | 0.75 | 0.61 | 0.6 | 0.6 | 0.67 | 0.67 | 0.59 | 0.59 | | | | |
| Ryan et al. ¹⁰⁵ | United States | 5 from OMOP | 0.74 | 0.81 | 0.54 | 0.53 | 0.53 | 0.75 | 0.75 | 0.58 | 0.69 | | | | |
| DuMouchel et al. ⁵² | United States | 5 from OMOP | | | | | | | | | 0.57 | 0.50 | 0.54 | | |
| Murphy et al. ⁶⁶ | United States | PHS ² from OMOP | 0.57 | 0.61 | 0.61 | 0.63 | 0.68 | 0.65 | 0.65 | 0.47 | 0.47 | | | | |
| Schuemie et al. ⁵³ | Denmark, Italy, Netherlands | 7 from EU-ADR | 0.74 | 0.75 | 0.75 | | | | | 0.78 | 0.72 | 0.71 | 0.72 | 0.69 | |
| Reps et al. ⁷⁰ | United Kingdom | THIN | | | | | | | 0.56 | | | | 0.55 | | |

Note: Where possible, the AUC is measured as the average AUC of the best performing combination across all drug-outcome pairs and databases. Only the methods evaluated in at least two papers are displayed.

Abbreviation: PHS, partners' healthcare system.
^awith LEOPARD filtering.²

equally important, which in practice is not the case for signal detection, and while objective in practice may provide a misleading view of signal detection value. Other measures were sparsely reported and could not be compared.

4.3 | Performance stratified by type of drug or outcome

Only 10 papers proposed an analysis by subgroup of drugs and outcomes, in a heterogeneous manner. It is encouraging to see increased performances in subgroup analyses compared to the average AUCs reported earlier, meaning that some methods are able to perform well when restricted to certain subgroups of DOIs and HOIs. Further work is needed to assess the reliability and reproducibility of these results.

Self-controlled methods were optimal for all acute outcomes in OMOP¹⁰⁵ except in one of the databases where TPD led to the highest AUCs. Zhou et al.⁵⁶ supported these results and suggested that self-controlled methods may identify slow onset outcomes if the signal is strong. However, they did not investigate negative controls so the specificity of their findings is unknown.

Most of the papers were non-specific in their selection of outcome and its characterization or focused on rapid onset AEs. The best method for detecting long-term ADRs, if any to date, remains understudied and therefore unclear. Further work is needed in this area as routinely collected data can have a great advantage of recording long-term outcomes over SRs. Since they can happen years after exposure, it is clearly an even more difficult signal detection problem to associate the outcome with a drug exposure with SRs.

4.4 | Comparability and generalizability of the findings

There was a lack of agreement on a possible gold standard for performance assessment. The findings were strongly influenced by the three main projects described earlier since most of the studies used one of the specific references sets that were proposed therein, which while large still represent a small proportion of all safety knowledge and have well published limitations.^{23,117,118} These reference sets used different outcome definitions. Some were limited to strong signals, and slower onset outcomes were mostly excluded.

There is an inherent variation of the AUCs between the databases, which was shown to be 20-30% for each method between U.S. databases in the OMOP experiment with the same reference set.¹⁰⁶ Comparison across studies using different databases is therefore not possible. However, study replication in several databases can increase precision and power to detect certain signals.¹¹⁹

Signal detection capabilities also depend greatly on the chosen analytic configuration.^{4,24,105} In Ryan et al.,¹⁰⁵ at least one configuration led to an AUC close or equal to 0.5 for each method-drug-outcome combination. In this review, the optimal configuration across all outcomes and databases was chosen as the reference

TABLE 6 AUC of different methods for (a) acute liver injury, (b) acute renal failure, (c) upper gastrointestinal bleeding, and (d) acute myocardial infarction

| (a) Acute liver injury | | | | | | |
|--|--------------------------------|------------------------------|------------------------------|--------------------------------|--------------------------------|-----------------------------|
| Acute liver injury | Schuemie et al. ¹⁰⁸ | Madigan et al. ⁶¹ | Suchard et al. ⁵⁵ | Schuemie et al. ¹¹⁰ | DuMouchel et al. ⁵² | Thurin et al. ⁴⁹ |
| Reference set | OMOP II | OMOP II | OMOP I | OMOP II | OMOP II | ALCAPONE |
| Country | United States | United States | United States | Europe | United States | France |
| LGPS + LEOPARD | 0.57 | | | | | |
| Case-control | 0.59 | | | | | 0.90 |
| Case-population | | | | | | 0.85 |
| SCCS | | | 0.61 | 0.73 | | |
| PRR | | | | | | 0.62 |
| BCPNN | | | | | | 0.57 |
| MGPS | | | | | | 0.50 |
| (b) Acute renal failure | | | | | | |
| Acute renal failure | Schuemie et al. ¹⁰⁸ | Madigan et al. ⁶¹ | Suchard et al. ⁵⁵ | Schuemie et al. ¹¹⁰ | DuMouchel et al. ⁵² | |
| Reference set | OMOP II | OMOP II | OMOP I | OMOP II | OMOP II | |
| Country | United States | United States | United States | Europe | United States | |
| LGPS + LEOPARD | 0.58 | | | | | |
| Case-control | 0.62 | | | | | |
| SCCS | | | 0.85 | 0.94 | | |
| PRR | | | | | | 0.61 |
| BCPNN | | | | | | 0.42 |
| MGPS | | | | | | 0.59 |
| (c) Upper gastrointestinal bleeding | | | | | | |
| Upper GI bleeding | Schuemie et al. ¹⁰⁸ | Madigan et al. ⁶¹ | Suchard et al. ⁵⁵ | Schuemie et al. ¹¹⁰ | DuMouchel et al. ⁵² | Thurin et al. ⁴⁸ |
| Reference set | OMOP II | OMOP II | OMOP I | OMOP II | OMOP II | ALCAPONE |
| Country | United States | United States | United States | Europe | United States | France |
| LGPS + LEOPARD | 0.67 | | | | | |
| Case-control | 0.64 | | | | | 0.62 |
| Case-population | | | | | | 0.67 |
| SCCS | | | 0.82 | 0.84 | | |
| PRR | | | | | | 0.47 |
| MGPS | | | | | | 0.53 |
| (d) Acute myocardial infarction | | | | | | |
| Acute MI | Schuemie et al. ¹⁰⁸ | Madigan et al. ⁶¹ | Suchard et al. ⁵⁵ | Schuemie et al. ¹¹⁰ | DuMouchel et al. ⁵² | |
| Reference set | OMOP II | OMOP II | OMOP I | OMOP II | OMOP II | |
| Country | United States | United States | United States | Europe | United States | |
| LGPS + LEOPARD | 0.662 | | | | | |
| Case-control | 0.65 | | | | | |
| SCCS | | | 0.73 | 0.79 | | |
| PRR | | | | | | 0.60 |
| MGPS | | | | | | 0.60 |

Abbreviation: Gastrointestinal (GI); LEOPARD, Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs; LGPS, Longitudinal Gamma Poisson Shrinker; MGPS, Multi-Item Gamma Poisson Shrinker; PRR, Proportional Reporting Ratio; Self-controlled case series (SCCS).

measure, but higher AUCs could be achieved when applying the optimal configuration to a single outcome and database. Gruber et al.²³ suggested that design choices need to be specific to the characteristics of the drug outcome pairs to avoid highlighting spurious associations.

4.5 | Strengths and limitations of the review

To our knowledge, this is the first systematic review to explore the performance of methods for signal detection stratified by drugs and outcomes. Moreover, we updated the literature by including methods

that are recently developed. We comprehensively described methods used for signal detection, evaluated the quality of the included studies narratively as well as compared the main measures of performance reported from the literature.

We also recognize some limitations. First, relevant studies might have been missed if they did not mention specific keywords in their abstract or full text as signal detection terminology is not standardly used in current literature. We added manual searching and screening bibliography of reviews to improve sensitivity. Quantitative comparison of performance was limited by the heterogeneity of the publications and the lack of gold standard, replicability of the studies was insufficient to perform re-analyses.

4.6 | Recommendations

Further research on the methods' performances for specific types of drugs and outcomes, focusing on inherent strengths and limitations of each method is needed. We also encourage more comprehensive reporting of the performance for individual or subgroups of drug-outcome pairs. We would like to see more head to head comparisons of methods for a larger range of drug-outcome pairs, including slower-onset outcomes. As all reference sets have inherent limitations, we would encourage the development of multiple and diverse reference sets publicly available for reuse. Ideally, generic and accessible codes that can be implemented in any database could be developed, with the use of common data models. We would also like to see results on the timeliness of signal detection with RWD, which was investigated only a single paper included in this review.¹⁵

5 | CONCLUSIONS

No method using routinely collected data showed superior performance across all drugs and outcomes, with heterogeneous performance assessment and reporting. However, some evidence showed that self-controlled designs, HDPS and ML achieved higher AUCs compared to other methods. Performance assessment for methods with slower onset outcomes is lacking.

An ideal approach is likely to involve more than one method to detect multiple drug-outcome pairs since none appears to have universal application to all outcomes and drugs. The aim of a signal detection programme, the type of drugs and outcomes under consideration and the drug- or outcome-based approach taken should be guiding the choice of the method. Future studies should investigate the performance of methods stratified by type of drug and outcome.

CONFLICT OF INTEREST

Astrid Coste is funded by a GSK PhD studentship to undertake this review. Andrew Bate is an employee of GSK and holds stocks and stock options. Ian Douglas holds grants and shares from GSK.

ORCID

Astrid Coste  <https://orcid.org/0000-0003-4207-7447>

Andrew Bate  <https://orcid.org/0000-0003-3151-3653>

REFERENCES

1. Patadia VK, Coloma P, Schuemie MJ, et al. Using real-world healthcare data for pharmacovigilance signal detection—the experience of the EU-ADR project. *Expert Rev Clin Pharmacol*. 2015;8:95-102. doi: [10.1586/17512433.2015.992878](https://doi.org/10.1586/17512433.2015.992878)
2. CIOMS. Working Group VIII. *Practical Aspects of Signal Detection in Pharmacovigilance*. Council for International Organizations of Medical Sciences (CIOMS); 2010.
3. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf*. 2009;18:427-436. doi: [10.1002/pds.1742](https://doi.org/10.1002/pds.1742)
4. Moore TJ, Furberg CD. Electronic health data for Postmarket surveillance: a vision not realized. *Drug Saf*. 2015;38:601-610. doi: [10.1007/s40264-015-0305-9](https://doi.org/10.1007/s40264-015-0305-9)
5. Honig PK. Advancing the science of pharmacovigilance. *Clin Pharmacol Ther*. 2013;93:474-475. doi: [10.1038/clpt.2013.60](https://doi.org/10.1038/clpt.2013.60)
6. Trifirò G, Patadia V, Schuemie MJ, et al. EU-ADR healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection. *Stud Health Technol Inform*. 2011;166:25-30. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-79960993889&partnerID=40&md5=aad7976293e7849d6c44f9553b65b6e0>
7. Arnaud M, Bégaud B, Thurin N, Moore N, Pariente A, Salvo F. Methods for safety signal detection in healthcare databases: a literature review. *Expert Opin Drug Saf*. 2017;16:721-732. doi: [10.1080/14740338.2017.1325463](https://doi.org/10.1080/14740338.2017.1325463)
8. Norén GN, Bergvall T, Ryan PB, et al. Empirical performance of the calibrated self-controlled cohort analysis within temporal pattern discovery: lessons for developing a risk identification and analysis system. *Drug Saf*. 2013;36:S107-S121. doi: [10.1007/s40264-013-0095-x](https://doi.org/10.1007/s40264-013-0095-x)
9. Jeong E, Park N, Choi Y, Park RW, Yoon D. Machine learning model combining features from algorithms with different analytical methodologies to detect laboratory-event-related adverse drug reaction signals. *PLoS One*. 2018;13:e0207749.
10. Morel M, Bacry E, Gaïffas S, et al. ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection. *Bio-statistics*. 2020;21:758-774. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emexa&NEWS=N&AN=633182054>
11. Luo Y, Thompson WK, Herr TM, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. *Drug Saf*. 2017;40:1075-1089. doi: [10.1007/s40264-017-0558-6](https://doi.org/10.1007/s40264-017-0558-6)
12. Mesfin YM, Cheng A, Lawrie J, BATTERY J. Use of routinely collected electronic healthcare data for postlicensure vaccine safety signal detection: a systematic review. *BMJ Glob Health*. 2019;4:e001065. doi: [10.1136/bmjgh-2018-001065](https://doi.org/10.1136/bmjgh-2018-001065)
13. Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ*. 2018;363:3532. doi: [10.1136/bmj.k3532](https://doi.org/10.1136/bmj.k3532)
14. Kim J, Kim MMJ, Ha J-H, et al. Signal detection of methylphenidate by comparing a spontaneous reporting database with a claims database. *Regul Toxicol Pharmacol*. 2011;61:154-160. doi: [10.1016/j.yrtph.2011.03.015](https://doi.org/10.1016/j.yrtph.2011.03.015)
15. Wahab IA, Pratt NL, Kalisch LM, et al. Comparing time to adverse drug reaction signals in a spontaneous reporting database and a claims database: a case study of rofecoxib-induced myocardial infarction and rosiglitazone-induced heart failure signals in Australia. *Drug Saf*. 2014;37:53-64.

16. Reps J, Feyereisil J, Garibaldi JM, Aickelin U, Gibson JE, Hubbard RB. Investigating the detection of adverse drug events in a UK general practice electronic health-care database. Paper presented at: UKCI 2011 - Proceedings of the 11th UK Workshop on Computational Intelligence; 2011; Manchester UK, 167–172. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84908489832&partnerID=40&md5=95653c8a28085895d4585dc0bc200c7f>.
17. Pacurariu AC, Straus SM, Trifirò G, et al. Useful interplay between spontaneous ADR reports and electronic healthcare Records in Signal Detection. *Drug Saf.* 2015;38:1201-1210.
18. Patadia VK, Schuemie MJ, Coloma P, et al. Evaluating performance of electronic healthcare records and spontaneous reporting data in drug safety signal detection. 37:94-104.
19. Demailly R. *Détection automatisée de signaux en pharmacovigilance chez la femme enceinte à partir de bases médico-administratives*. HAL; 2021.
20. Schuemie MJ. Safety surveillance of longitudinal databases: further methodological considerations. *Pharmacoepidemiol Drug Saf.* 2012; 21:670-672. doi:10.1002/pds.3259
21. Norén GN, Hopstadius J, Bate A, Edwards IR, Noren GN, Bate A. Safety surveillance of longitudinal databases: methodological considerations. *Pharmacoepidemiol Drug Saf.* 2011;20:714-717. doi:10.1002/pds.2151
22. Coloma PM, Avillach P, Salvo F, et al. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. *Drug Saf.* 2013;36:13-23.
23. Gruber S, Chakravarty A, Heckbert SR, et al. Design and analysis choices for safety surveillance evaluations need to be tuned to the specifics of the hypothesized drug–outcome association. *Pharmacoepidemiol Drug Saf.* 2016;25:973-981. doi:10.1002/pds.4065
24. Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. *Ther Adv Drug Saf.* 2013;4:53-62. doi:10.1177/2042098613477445
25. Madigan D, Stang PE, Berlin JA, et al. A systematic statistical approach to evaluating evidence from observational studies. *Annu Rev Stat Appl.* 2014;1:11-39. doi:10.1146/annurev-statistics-022513-115645
26. Yu Y-C, Wei R, Jia L-L, et al. Exploring the drug-induced anemia signals in children using electronic medical records. *Expert Opin Drug Saf.* 2019;18:993-999. doi:10.1080/14740338.2019.1645832
27. Lee S, Choi J, Kim H-SS, et al. Standard-based comprehensive detection of adverse drug reaction signals from nursing statements and laboratory results in electronic health records. *J Am Med Inform Assoc.* 2017;24:697-708. doi:10.1093/JAMIA/OCW168
28. Wei R, Jia LL, Yu YC, et al. Pediatric drug safety signal detection of non-chemotherapy drug-induced neutropenia and agranulocytosis using electronic healthcare records. *Expert Opin Drug Saf.* 2019;18: 435-441. doi:10.1080/14740338.2019.1604682
29. Mansour A, Ying H, Dews P, Ji Y, Michael Massanari R. Fuzzy rule-based approach for detecting adverse drug reaction signal pairs. Paper presented at: 8th Conference of the European Society for Fuzzy Logic and Technology, EUSFLAT 2013—Advances in Intelligent Systems Research, volume 32; 2013; Milano, Italy, 384–391. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84891771416&partnerID=40&md5=78ef71b30f239c3d773639c5fd4cfcab>.
30. Park MY, Yoon D, Lee K, et al. A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. *Pharmacoepidemiol Drug Saf.* 2011;20:598-607.
31. Tham MY, Ye Q, Ang PS, et al. Application and optimisation of the comparison on extreme laboratory tests (CERT) algorithm for detection of adverse drug reactions: transferability across national boundaries. *Pharmacoepidemiol Drug Saf.* 2018;27:87-94.
32. Lai ECC, Hsieh CY, Yang YHK, Lin SJ. Detecting potential adverse reactions of sulpiride in schizophrenic patients by prescription sequence symmetry analysis. *PLoS One.* 2014;9. doi:10.1371/journal.pone.0089795
33. Zhan C, Liu L, Li J, Roughead E, Pratt N. A data-driven method to detect adverse drug events from prescription data. *J Biomed Inform.* 2018;85:10-20.
34. Zhan C, Liu L, Li J, Roughead E, Pratt N. Detecting potential signals of adverse drug events from prescription data. *Artif Intell Med.* 2020; 104:101839. doi:10.1016/j.artmed.2020.101839
35. Pratt N, Chan EW, Choi N-KK, et al. Prescription sequence symmetry analysis: assessing risk, temporality, and consistency for adverse drug reactions across datasets in five countries. *Pharmacoepidemiol Drug Saf.* 2015;24:858-864. doi:10.1002/(ISSN)1099-1557
36. Hoang T, Liu J, Roughead E, Pratt N, Li J. Supervised signal detection for adverse drug reactions in medication dispensing data. *Comput Methods Prog Biomed.* 2018;161:25-38.
37. Karimi S, Wang C, Metke-Jimenez A, Gaire R, Paris C. Text and data mining techniques in adverse drug reaction detection. *ACM Comput Surv.* 2015;47:1-39. doi:10.1145/2719920
38. Kaguelidou F, Durrieu G, Clavenna A. Pharmacoepidemiological research for the development and evaluation of drugs in pediatrics. *Therapie.* 2019;74:315-324. doi:10.1016/j.therap.2018.09.077
39. Jones JK. The role of data mining technology in the identification of signals of possible adverse drug reactions: value and limitations. *Curr Ther Res Clin Exp.* 2001;62:664-672. doi:10.1016/S0011-393X%2801%2980072-2
40. Nelson JC, Ulloa-Pérez E, Bobb JF, et al. Leveraging the entire cohort in drug safety monitoring: part 1 methods for sequential surveillance that use regression adjustment or weighting to control confounding in a multisite, rare event, distributed data setting. *J Clin Epidemiol.* 2019;112:77-86.
41. Lai EC-C, Pratt N, Hsieh C-Y, et al. Sequence symmetry analysis in pharmacovigilance and pharmacoepidemiologic studies. *Eur J Epidemiol.* 2017;32:567-582. doi:10.1007/s10654-017-0281-8
42. Coloma PM, Trifirò G, Patadia V, Sturkenboom M. Postmarketing safety surveillance: where does signal detection using electronic healthcare records fit into the big picture? *Drug Saf.* 2013; 36:183-197.
43. Wisniewski AFZZ, Bate A, Bousquet C, et al. Good signal detection practices: evidence from IMI PROTECT. *Drug Saf.* 2016;39:469-490.
44. Prieto-Merino D, Quartey G, Wang J, Kim J. Why a Bayesian approach to safety analysis in pharmacovigilance is important. *Pharm Stat.* 2011;10:554-559.
45. Suling M, Pigeot I. Signal detection and monitoring based on longitudinal healthcare data. *Pharmaceutics.* 2012;4:607-640.
46. Coloma PM, Trifirò G, Schuemie MJ, et al. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? 2012;21:611-621.
47. Gault N, Castañeda-Sanabria J, de Rycke Y, Guillo S, Foulon S, Tubach F. Self-controlled designs in pharmacoepidemiology involving electronic healthcare databases: a systematic review. *Med Res Methodol.* 2017;17:25. doi:10.1186/s12874-016-0278-0
48. Thurin NH, Lassalle R, Schuemie M, et al. Empirical assessment of case-based methods for identification of drugs associated with upper gastrointestinal bleeding in the French National Healthcare System database (SNDS). *Pharmacoepidemiol Drug Saf.* 2020;29:890-903. doi:10.1002/pds.5038
49. Thurin NH, Lassalle R, Schuemie M, et al. Empirical assessment of case-based methods for identification of drugs associated with acute liver injury in the French National Healthcare System database (SNDS). *Pharmacoepidemiol Drug Saf.* 2021;30:320-333. doi:10.1002/pds.4983
50. Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Stat Methods Med Res.* 2013;22:39-56. doi:10.1177/0962280211403602
51. Choi N-K, Chang Y, Choi YK, Hahn S, Park B-J. Signal detection of rosuvastatin compared to other statins: data-mining study using national health insurance claims database. 2010;19:238-246.
52. DuMouchel W, Ryan PB, Schuemie MJ, Madigan D. Evaluation of disproportionality safety signaling applied to healthcare databases. *Drug Saf.* 2013;36:S123-S132.

53. Schuemie MJ, Coloma PM, Straatman H, et al. Using electronic health Care Records for Drug Safety Signal Detection: a comparative evaluation of statistical methods. *Med Care*. 2012;50:890-897. doi:[10.1097/MLR.0B013E31825F63BF](https://doi.org/10.1097/MLR.0B013E31825F63BF)
54. Schuemie MJ. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiol Drug Saf*. 2011;20:292-299.
55. Suchard MA, Zorych I, Simpson SE, Madigan D, Schuemie MJ, Ryan PB. Empirical performance of the self-controlled case series design: lessons for developing a risk identification and analysis system. *Drug Saf*. 2013;36:S83-S93.
56. Zhou X, Douglas IJ, Shen R, Andrew B, Douglas IJ, Andrew B. Signal detection for recently approved products: adapting and evaluating self-controlled case series method using a US claims and UK electronic medical records database. *Drug Saf*. 2018;41:523-536.
57. Schuemie MJ, Trifirò G, Coloma PM, et al. Detecting adverse drug reactions following long-term exposure in longitudinal observational data: the exposure-adjusted self-controlled case series. *Stat Methods Med Res*. 2016;25:2577-2592.
58. Simpson SE. A positive event dependence model for self-controlled case series with applications in Postmarketing surveillance. *Biometrics*. 2013;69:128-136. doi:[10.1111/j.1541-0420.2012.01795.x](https://doi.org/10.1111/j.1541-0420.2012.01795.x)
59. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf*. 2010;19:858-868. doi:[10.1002/pds.1926](https://doi.org/10.1002/pds.1926)
60. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug Saf*. 2013;36:S59-S72.
61. Madigan D, Schuemie MJ, Ryan PB. Empirical performance of the case-control method: lessons for developing a risk identification and analysis system. *Drug Saf*. 2013;36:S73-S82.
62. Takeuchi Y, Shinozaki T, Matsuyama Y. A comparison of estimators from self-controlled case series, case-crossover design, and sequence symmetry analysis for pharmacoepidemiological studies. *BMC Med Res Methodol*. 2018;18:4. doi:[10.1186/s12874-017-0457-7](https://doi.org/10.1186/s12874-017-0457-7)
63. Thurin NH, Lassalle R, Schuemie M, et al. Empirical assessment of case-based methods for drug safety alert identification in the French National Healthcare System database (SNDS): methodology of the ALCAPONE project. *Pharmacoepidemiol Drug Saf*. 2020;29:993-1000. doi:[10.1002/pds.4983](https://doi.org/10.1002/pds.4983)
64. Wang SV, Maro JC, Gagne JJ, et al. A general propensity score for signal identification using tree-based scan statistics. *Am J Epidemiol*. 2021;190:1424-1433.
65. Grosso A, Douglas I, MacAllister R, Petersen I, Smeeth L, Hingorani AD. Use of the self-controlled case series method in drug safety assessment. *Expert Opin Drug Saf*. 2011;10:337-340. doi:[10.1517/14740338.2011.562187](https://doi.org/10.1517/14740338.2011.562187)
66. Murphy SN, Castro V, Colecchi J, et al. Partners HealthCare OMOP Study Report; 2011.
67. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR. Temporal pattern discovery in longitudinal electronic patient records. *Data Min Knowl Disc*. 2010;20:361-387. doi:[10.1007/s10618-009-0152-3](https://doi.org/10.1007/s10618-009-0152-3)
68. Hopstadius J, Noren GN, Bate A, Norén GN, Hopstadius J, Bate A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. *Stat Methods Med Res*. 2013;22:57-69. doi:[10.1177/0962280211403604](https://doi.org/10.1177/0962280211403604)
69. Jin H, Chen J, He H, Kelman C, Mcaullay D, O'keefe CM. Signaling potential adverse drug reactions from administrative health databases. *IEEE Trans Knowl Data Eng*. 2010;22:839-853. doi:[10.1109/TKDE.2009.212](https://doi.org/10.1109/TKDE.2009.212)
70. Reps JM, Garibaldi JM, Aickelin U, Soria D, Gibson J, Hubbard R. Comparison of algorithms that detect drug side effects using electronic healthcare databases. *Soft Comput*. 2013;17:2381-2397. doi:[10.1007/s00500-013-1097-4](https://doi.org/10.1007/s00500-013-1097-4)
71. Jin H, Chen J, Kelman C, He H, Mcaullay D, O'Keefe CM. Mining unexpected associations for signalling potential adverse drug reactions from administrative health databases. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 3918. LNAI; 2006:867-876. doi:[10.1007/11731139_101](https://doi.org/10.1007/11731139_101)
72. Reps J, Garibaldi JM, Aickelin U, Soria D, Gibson JE, Hubbard RB. Comparing data-mining algorithms developed for longitudinal observational databases. Paper presented at: 2012 12th UK Workshop on Computational Intelligence, UKCI 2012; 2012, Edinburgh, UK. doi:[10.1109/UKCI.2012.6335771](https://doi.org/10.1109/UKCI.2012.6335771)
73. Ji Y, Ying H, Dews P, et al. An exclusive causal-leverage measure for detecting adverse drug reactions from electronic medical records. Paper presented at: Annual Conference of the North American Fuzzy Information Processing Society—NAFIPS; 2011, El Paso. doi:[10.1109/NAFIPS.2011.5751957](https://doi.org/10.1109/NAFIPS.2011.5751957)
74. Ji Y, Ying H, Dews P, et al. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Trans Inf Technol Biomed*. 2011;15:428-437. doi:[10.1109/TITB.2011.2131669](https://doi.org/10.1109/TITB.2011.2131669)
75. Ji Y, Ying H, Tran J, Dews P, Mansour A, Michael MR. A method for mining infrequent causal associations and its application in finding adverse drug reaction signal pairs. *IEEE Trans Knowl Data Eng*. 2013; 25:721-733. doi:[10.1109/TKDE.2012.28](https://doi.org/10.1109/TKDE.2012.28)
76. Wang SV, Gagne JJ, Schneeweiss S, et al. Hypothesis-free screening of large administrative databases for unsuspected drug-outcome associations. *Eur J Epidemiol*. 2018;33:545-555.
77. Arnaud M, Bégaud B, Thiessard F, et al. An automated system combining safety signal detection and prioritization from healthcare databases: a pilot study. *Drug Saf*. 2018;41:377-387.
78. Hellfritzsche M, Rasmussen L, Hallas J, Pottegard A. Using the symmetry analysis design to screen for adverse effects of non-vitamin K antagonist Oral anticoagulants. *Drug Saf*. 2018;41:685-695. doi:[10.1007/s40264-018-0650-6](https://doi.org/10.1007/s40264-018-0650-6)
79. Wahab IA, Pratt NL, Ellett LK, et al. Sequence symmetry analysis as a signal detection tool for potential heart failure adverse events in an administrative claims database. *Drug Saf*. 2016;39:347-354. doi:[10.1007/s40264-015-0391-8](https://doi.org/10.1007/s40264-015-0391-8)
80. Tsiropoulos I, Andersen M, Hallas J. Adverse events with use of anti-epileptic drugs: a prescription and event symmetry analysis. *Pharmacoepidemiol Drug Saf*. 2009;18:483-491. doi:[10.1002/pds.1736](https://doi.org/10.1002/pds.1736)
81. Zhou X, Bao W, Gaffney M, Shen R, Young S, Bate A. Assessing performance of sequential analysis methods for active drug safety surveillance using observational data. *J Biopharm Stat*. 2018;28:668-681.
82. Kulldorff M, Davis RL, Kolczak M, Lewis E, Lieu T, Platt R. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Seq Anal*. 2011;30:58-78. doi:[10.1080/07474946.2011.539924](https://doi.org/10.1080/07474946.2011.539924)
83. Brown JS, Kulldorff M, Chan KA, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiol Drug Saf*. 2007;16: 1275-1284.
84. Brown JS, Kulldorff M, Petronis KR, et al. Early adverse drug event signal detection within population-based health networks using sequential methods: key methodologic considerations. *Pharmacoepidemiol Drug Saf*. 2009;18:226-234.
85. Cook AJ, Tiwari RC, Wellman RD, et al. Statistical approaches to group sequential monitoring of postmarket safety surveillance data: current state of the art for use in the mini-sentinel pilot. *Pharmacoepidemiol Drug Saf*. 2012;21:72-81. doi:[10.1002/pds.2320](https://doi.org/10.1002/pds.2320)
86. Li L. A conditional sequential sampling procedure for drug safety surveillance. *Stat Med*. 2009;28:3124-3138. doi:[10.1002/sim.3689](https://doi.org/10.1002/sim.3689)
87. Kulldorff M, Fang Z, Walsh SJ. A tree-based scan statistic for database disease surveillance. *Biometrics*. 2003;59:323-331.
88. Kulldorff M, Dashevsky I, Avery T, et al. Drug safety data mining with a tree-based scan statistic. *Pharmacoepidemiol Drug Saf*. 2013; 19:S172-S173. doi:[10.1002/pds.2019](https://doi.org/10.1002/pds.2019)

89. Huybrechts KF, Kulldorff M, Hernández-Díaz S, et al. Active surveillance of the safety of medications used during pregnancy. *Am J Epidemiol*. 2021;190:1159-1168. doi:10.1093/aje/kwaa288
90. Wang S v, Maro JC, Baro E, et al. Data Mining for Adverse Drug Events with a propensity score-matched tree-based scan statistic. *Epidemiology*. 2018;29:895-903. doi:10.1097/EDE.0000000000000907
91. Wintzell V, Svanström H, Melbye M, et al. Data Mining for Adverse Events of tumor necrosis factor-alpha inhibitors in pediatric patients: tree-based scan statistic analyses of Danish Nationwide health data. *Clin Drug Investig*. 2020;40:1147-1154.
92. Schachterle SE, Hurley S, Liu Q, Petronis KR, Bate A, Schachterle SE. An implementation and visualization of the tree-based scan statistic for safety event monitoring in longitudinal electronic health data. *Drug Saf*. 2019;42:727-741. doi:10.1007/s40264-018-00784-0
93. Brown JS, Petronis KR, Bate A, et al. Drug adverse event detection in health plan data using the gamma Poisson Shrinker and comparison to the tree-based scan statistic. *Pharmaceutics*. 2013;5:179-200. doi:10.3390/pharmaceutics5010179
94. Reps JM, Garibaldi JM, Aickelin U, Gibson JE, Hubbard RB. A supervised adverse drug reaction signalling framework imitating Bradford Hill's causality considerations. *J Biomed Inform*. 2015;56:356-368. doi:10.1016/j.jbi.2015.06.011
95. Whalen E, Hauben M, Andrew B. Time series disturbance detection for hypothesis-free signal detection in longitudinal observational databases. *Drug Saf*. 2018;22:890-897. doi:10.1097/MLR.Ob013e31825f63bf
96. Karlsson I, Zhao J. Dimensionality reduction with random indexing: an application on adverse drug event detection using electronic health records. Paper presented at: Proceedings of IEEE Symposium on Computer-Based Medical Systems; 2014; New-York, 304-307. doi:10.1109/CBMS.2014.22
97. Reps JM, Garibaldi JM, Aickelin U, Soria D, Gibson JE, Hubbard RB. Signalling paediatric side effects using an ensemble of simple study designs. *Drug Saf*. 2014;37:163-170. doi:10.1007/s40264-014-0137-z
98. Bagattini F, Karlsson I, Rebane J, Papapetrou P. A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. *BMC Med Inform Decis Mak*. 2019;19:7.
99. Escolano S, Tubert-Bitter P, Ahmed I, Demailly R, Haramburu F. Identifying drugs inducing prematurity by mining claims data with high-dimensional confounder score strategies. *Drug Saf*. 2020;43:549-559. doi:10.1007/s40264-020-00916-5
100. Bampa M, Papapetrou P. Mining adverse drug events using multiple feature hierarchies and patient history windows. Paper presented at: IEEE International Conference on Data Mining Workshops, ICDMW, volume 2019; 2019; Beijing, China, 925-932. doi:10.1109/ICDMW.2019.00135
101. Zhao J, Henriksson A, Kvist M, Asker L, Boström H. Handling temporality of clinical events for drug safety surveillance. *AMIA Ann Symp Proc*. 2015;2015:1371-1380.
102. Duan L, Khoshneshin M, Street WN, Liu M. Adverse drug effect detection. *IEEE J Biomed Health Inform*. 2013;17:305-311.
103. Sauzet O, Carvajal A, Escudero A, Molokhia M, Cornelius VR. Illustration of the weibull shape parameter signal detection tool using electronic healthcare record data. *Drug Saf*. 2013;36:995-1006.
104. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. *Drug Saf*. 2013;36:S33-S47. doi:10.1007/s40264-013-0097-8
105. Ryan PB, Stang PE, Overhage JM, et al. A comparison of the empirical performance of methods for a risk identification system. *Drug Saf*. 2013;36(Suppl 1):S143-S158. doi:10.1007/s40264-013-0108-9
106. Ryan PB, Madigan D, Stang PE, et al. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership. *Stat Med*. 2012;31:4401-4415. doi:10.1002/sim.5620
107. Choi N-K, Chang Y, Kim J-Y, Choi Y-K, Park B-J. Comparison and validation of data-mining indices for signal detection: using the Korean national health insurance claims database. *Pharmacoepidemiol Drug Saf*. 2011;20:1278-1286. doi:10.1002/pds.2237
108. Schuemie MJ, Madigan D, Ryan PB. Empirical performance of LGPS and LEOPARD: lessons for developing a risk identification and analysis system. *Drug Saf*. 2013;36:S133-S142.
109. Sturkenboom MCJM, van der Lei J, Trifiro G, et al. *The EU-ADR project: preliminary results and perspective*. *Studies in Health Technology and Informatics*. IOS Press; 2009:43-49. doi:10.3233/978-1-60750-043-8-43
110. Schuemie MJ, Gini R, Coloma PM, et al. Replication of the OMOP experiment in europe: evaluating methods for risk identification in electronic health record databases. *Drug Saf*. 2013;36:159-169. doi:10.1007/s40264-013-0109-8
111. McDowell RD, Hughes C, Murchie P, Cardwell C. A systematic assessment of the association between frequently prescribed medicines and the risk of common cancers: a series of nested case-control studies. *BMC Med*. 2021;19:22. doi:10.1186/s12916-020-01891-5
112. Støer NC, Botteri E, Thoresen GH, et al. Drug use and cancer risk: a drug-wide association study (DWAS) in Norway. *Cancer Epidemiol Biomark Prev*. 2021;30:682-689. doi:10.1158/1055-9965.EPI-20-1028
113. Friedman GD, Udaltsova N, Chan J, Quesenberry CP, Habel LA. Screening pharmaceuticals for possible carcinogenic effects: initial positive results for drugs not previously screened. *Cancer Causes Control*. 2009;20:1821-1835. doi:10.1007/s10552-009-9375-2
114. Pottgård A, Friis SS, Christensen R, Depont R, et al. Identification of associations between prescribed medications and cancer: a Nationwide screening study. *EBioMedicine*. 2016;7:73-79. doi:10.1016/j.ebiom.2016.03.018
115. Patel CJ, Ji J, Sundquist J, Ioannidis JPA, Sundquist K. Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study. *Sci Rep*. 2016;6:31308. doi:10.1038/srep31308
116. Norén GN, Hopstadius J, Bate A, Edwards IR. Safety surveillance of longitudinal databases: results on real-world data. *Pharmacoepidemiol Drug Saf*. 2012;21:673-675. doi:10.1002/pds.3258
117. Hennessy S, Leonard CE. Comment on: "desideratum for evidence-based epidemiology.". *Drug Saf*. 2015;38:101-103. doi:10.1007/s40264-014-0252-x
118. Gagne JJ, Schneeweiss S. Comment on "empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership.". *Stat Med*. 2013;32:1073-1074. doi:10.1002/sim.5699
119. Rossi F, Capuano A, Ferrajolo C, et al. Idiopathic acute liver injury in paediatric outpatients: incidence and signal detection in two european countries. *Drug Saf*. 2013;36:1007-1016. doi:10.1007/s40264-013-0045-7

SUPPORTING INFORMATION

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

How to cite this article: Coste A, Wong A, Bokern M, Bate A, Douglas IJ. Methods for drug safety signal detection using routinely collected observational electronic health care data: A systematic review. *Pharmacoepidemiol Drug Saf*. 2022;1-16. doi:10.1002/pds.5548