ORIGINAL RESEARCH ARTICLE

Improved Statistical Signal Detection in Pharmacovigilance by Combining Multiple Strength-of-Evidence Aspects in vigiRank

Retrospective Evaluation against Emerging Safety Signals

**Ola Caster · Kristina Juhlin ·
Sarah Watson · G. Niklas Norén**

Abstract

Background Detection of unknown risks with marketed medicines is key to securing the optimal care of individual patients and to reducing the societal burden from adverse drug reactions. Large collections of individual case reports remain the primary source of information and require effective analytics to guide clinical assessors towards likely drug safety signals. Disproportionality analysis is based solely on aggregate numbers of reports and naively disregards report quality and content. However, these latter features are the very fundament of the ensuing clinical assessment.

Objective Our objective was to develop and evaluate a data-driven screening algorithm for emerging drug safety signals that accounts for report quality and content.

Methods vigiRank is a predictive model for emerging safety signals, here implemented with shrinkage logistic regression to identify predictive variables and estimate their respective contributions. The variables considered for inclusion capture different aspects of strength of evidence, including quality and clinical content of individual reports, as well as trends in time and geographic spread. A reference set of 264 positive controls (historical safety signals

from 2003 to 2007) and 5,280 negative controls (pairs of drugs and adverse events not listed in the Summary of Product Characteristics of that drug in 2012) was used for model fitting and evaluation; the latter used fivefold cross-validation to protect against over-fitting. All analyses were performed on a reconstructed version of VigiBase® as of 31 December 2004, at around which time most safety signals in our reference set were emerging.

Results The following aspects of strength of evidence were selected for inclusion into vigiRank: the numbers of informative and recent reports, respectively; dispropor-

O. Caster (✉) · K. Juhlin · S. Watson · G. N. Norén
Uppsala Monitoring Centre, Box 1051, SE-75140 Uppsala,
Sweden
e-mail: ola.caster@who-umc.org

O. Caster
Department of Computer and Systems Sciences, Stockholm
University, Forum 100, SE-164 40 Kista, Sweden

G. N. Norén
Department of Mathematics, Stockholm University,
SE-106 91 Stockholm, Sweden

Key Points

Today, automated screening of large collections of individual case reports to identify possible drug safety issues often relies on disproportionality analysis, which is based solely on aggregate numbers of reports, disregarding report quality and content

This study identifies the following variables as strong predictors of emerging drug safety issues: the number of informative reports, recent reports, and reports with free-text descriptions; disproportional reporting; and geographic spread. Simultaneously accounting for these aspects of strength of evidence significantly improves the accuracy of automated screening of individual case reports compared with disproportionality analysis alone

Utilizing the identified predictive model can be expected to reduce the number of false alerts and uncover drug safety issues that would otherwise go undetected

tional reporting; the number of reports with free-text descriptions of the case; and the geographic spread of reporting. vigiRank offered a statistically significant improvement in area under the receiver operating characteristics curve (AUC) over screening based on the Information Component (IC) and raw numbers of reports, respectively (0.775 vs. 0.736 and 0.707, cross-validated). *Conclusions* Accounting for multiple aspects of strength of evidence has clear conceptual and empirical advantages over disproportionality analysis. vigiRank is a first-of-its-kind predictive model to factor in report quality and content in first-pass screening to better meet tomorrow's post-marketing drug safety surveillance needs.

1 Introduction

Medicines enable patients to lead better lives and are fundamental to modern healthcare. However, any effective medicine comes with a risk of adverse effects, many of which are serious enough to end or dramatically change the lives of individual patients. Taken together, adverse drug reactions (ADRs) impose an enormous burden on society, causing hundreds of thousands of deaths annually [1, 2] at a cost of several billion $US [3].

Pre-marketing clinical trials include too few patients from groups that are too homogeneous to capture a drug's full spectrum of possible adverse effects, which makes extensive post-marketing surveillance a necessity [4]. The detection of a previously unknown safety issue can have great impact on the overall benefit–risk balance of a drug, as evidenced for example by the market withdrawal of rofecoxib upon the detection of its increased myocardial infarction risk.

Individual case reports of suspected harm from medicines remain the primary basis to detect unexpected risks of medicines [5, 6]. Their most important strength is that they draw on the astute observations of patients and health professionals to capture information specifically for the purpose of causality assessment [7]. Further, they cover all types of medicines and all patient groups in a single system. Their major limitations include that far from all suspected ADRs are identified as such, with fewer still being reported to the authorities, and that many reports lack the necessary detail to support solid causality assessment. A number of broad research initiatives currently explore the use of longitudinal health data for safety signal detection, and the results suggest that these may complement but not replace individual case reports for this purpose [8].

A major practical issue in the analysis of individual case reports to uncover possible new drug risks is their sheer numerosity. For many concerned organizations, both the number of submitted reports in itself and the number of associated drug–ADR combinations are too large for exhaustive manual review. This necessitates the use of triages to guide clinical assessment [9–11]. Generally, such triages can consider three main criteria: strength of evidence, novelty, and medical impact [12]. Since the 1990s, computerized evaluation of strength of evidence has relied largely on so-called disproportionality analysis [13–15].

Reflecting its practical value, disproportionality analysis has evolved from a specialist to a mainstream activity over the past 15 years [5, 16]. Nevertheless, it provides a very basic analysis that only accounts for statistical associations in the co-reporting of drugs and suspected ADRs. It is entirely based on aggregate numbers of reports and naively disregards the strength of individual reports. More recently, multivariate methods have been proposed that account for co-reported medicines and the indication for treatment [17–20]. These methods should yield strength-of-association measures less liable to confounding, although their practical value for pharmacovigilance is yet to be definitively demonstrated [21]. In any case, even the multivariate methods do not consider other information on a given report beyond its listed drug(s) and ADR(s), and occasionally the indication for treatment. Automated screening based on deviating patterns of reported time to onset [22] is conceptually different, but still focuses on a single aspect of strength of evidence.

This is contrasted by clinical assessment of potential safety signals, which attempts to manually account for all relevant aspects [11, 12]. In particular, the quality and content of individual reports is of fundamental importance [23]. Consequently, we aim in this paper to combine the benefits of automation with the breadth of aspects used in clinical assessment. Specifically, the aim of this paper is to derive a predictive model for emerging signals of suspected ADRs from large collections of individual case reports, accounting for a broad variety of aspects of strength of evidence, ranging from disproportionality analysis and case series diversity to individual report quality and content. This algorithm, denoted vigiRank, is inspired by similar work for adverse drug interaction surveillance [24].

2 Methods

vigiRank is a predictive model for emerging safety signals, here implemented with shrinkage logistic regression to identify predictive variables and estimate their respective contributions.

2.1 Variables Considered for Inclusion into vigiRank

The variables considered for inclusion into vigiRank capture inherently different aspects of a reporting pattern that

might strengthen our suspicion of a true causal association between the drug and the suspected ADR. This includes variables that measure the quality and clinical content of individual reports, as well as more quantitative aspects of a reporting pattern such as trends in time and geographic spread. Each considered variable is defined at the level of a drug–ADR pair and is based on the reporting pattern of that pair in a specific collection of individual case reports. Table 1 provides an overview of all proposed variables, with descriptions and motivations. The final selection of variables to be empirically evaluated was made by the co-authors through consensus. Candidate variables were initially identified through consultation with experienced pharmacovigilance professionals with clinical (four), pharmaceutical (five), and statistical (one) expertise. Below, two non-trivial variables are described in greater detail.

2.1.1 Informative Reports

The 'informative reports' variable measures the number of reports with a vigiGrade completeness score of 0.90 or more [25]. This score is 1 for a report with full information on the type of report, type of notifier, time to onset, country of origin, patient age and sex, indication for treatment, dosage, outcome, and free text description. It then decreases with a pre-specified multiplicative factor for each of these fields that is undocumented. A report with a vigiGrade completeness score of 0.90 or more may lack information on only one of the following information items: type of report, type of notifier, outcome, or free text description; all other items described above must be documented.

2.1.2 Disproportional Reporting

The 'disproportional reporting' variable is a binary indicator of whether the combination is disproportionally reported, either on the full dataset or a subset thereof. In this study, disproportionality is measured using the Information Component (IC), a disproportionality measure based on shrinkage observed-to-expected ratios [13, 26]. The IC is computed as:

$$IC = \log_2 \frac{O + 0.5}{E + 0.5} \tag{1}$$

where O and E are the observed and expected numbers, respectively, of reports on the drug–ADR pair. E is given by $(N_A \times N_D)/N$, where N_A is the total number of reports on the ADR; N_D is the total number of reports on the drug; and N is the total number of reports. Credibility intervals for the IC are obtained via the Gamma distribution [26], and IC_α denotes the α percentile of the posterior distribution for the IC.

Here, a drug–ADR pair is considered disproportionally reported when one or more of the following criteria are fulfilled: $IC_{025} > 0$ on the full data; $IC_{005} > 0$ on the full data when adjusting for age group, notifier, calendar year, or country; $IC_{0005} > 0$ in a stratum of the data based on age group or notifier; $IC_{005} > 0$ in two strata for the same covariate based on age group, notifier, or country; or $IC_{0005} > 0$ for the full data when excluding one stratum based on age group, notifier, country, or calendar year. Here, IC_{025}, IC_{005}, and IC_{0005} denote the lower limits of 95, 99, and 99.9 % credibility intervals for the IC, respectively. Stratifications and interval widths have been selected to protect against spurious associations from multiple comparisons according to Hopstadius and Norén [27].

2.2 Reference Set

To be able to identify reporting characteristics of emerging safety signals, a reference set of positive and negative controls was created. A total of 532 historical European Medicine Agency (EMA) safety signals from September 2003 to March 2007 for 267 European centrally authorized products (CAPs) were considered as potential positive controls of our reference set [28]. Each had an index date reflecting when the EMA first became aware that it required investigation.

Data on the proposed variables were extracted from the World Health Organization (WHO) global individual case safety report database, VigiBase® [29], for all potential controls. As of March 2014, VigiBase included more than 8.5 million reports from 118 countries. Here, vaccine reports and suspected duplicate reports were excluded [30]. Since we focused on emerging safety issues, a VigiBase reconstruction as of 31 December 2004 was used; this date is the end of the quarter in which the median of the index dates fell. No attempt was made to recreate the actual composition of VigiBase at that time. Instead, reports were time-stamped based on the date of the suspected ADR, or an approximation thereof.

After updating to the Medical Dictionary for Regulatory Activities (MedDRA®), version 14.1, matching substance names to the preferred base level of WHO Drug Dictionary Enhanced™, excluding vaccines, and requiring reports from at least two countries, 264 positive controls from 65 CAPs remained. As intended, by the rollback date of 31 December 2004, a vast majority of these were emerging rather than established safety signals: by this date only 30 of 264 (11 %) had led to regulatory action by the EMA, and for 118 (45 %) this date preceded the point in time when EMA first became aware of the potential safety signal. The reference set also included randomly selected negative controls, at a 20:1 ratio with the positive controls. Negative controls were pairs of the initial 267 CAPs and

Table 1 Variables considered for inclusion into vigiRank

Variable	Concept	Motivation	VigiBase[a] implementation
Informative reports (INF)	Reports on the drug and the ADR with sufficient information to allow a causality assessment of the individual case	Reporters may be more likely to provide detailed information when they have a strong suspicion that the adverse event was drug related	Reports with vigiGrade completeness score ≥ 0.9 (for details, see Sect. 2.1.1)
Narrative (NAR)	Number of reports with free text information available	Free text information may strengthen the causality assessment of a case. In addition to this, reporters may be more likely to provide free text information when they have a strong suspicion that the adverse event was drug related.	Reports with narrative information, excluding purely numerical narratives and standard phrases, e.g. 'none provided'
Dechallenge (DCH)	Reports indicating that the adverse event subsided upon withdrawal of the drug	Resolution of the adverse event upon withdrawal of the drug strengthens the causality assessment for the individual case	Reports with positive dechallenge (by definition including positive rechallenge, see below)
Rechallenge (RCH)	Reports indicating that the adverse event recurred upon re-exposure to the same drug	Repeated occurrence of the adverse event upon exposure to the drug strengthens the causality assessment for the individual case	Reports with positive rechallenge
Causality assessment (CAU and CAU+)	Reports indicating a positive result of causality assessment of the individual case	Strengthens the causality assessment of the individual case	Implemented as two separate variables: number of reports with causality probable/certain (CAU) and number of reports with causality certain (CAU+)
Time-to-onset (TTO)	Reports with a plausible time between the intake of the drug and the adverse event	Time-to-onset information may strengthen the causality assessment of the individual case	Reports with reported time-to-onset less than 90 days[b]
Solely reported (SOL)	Reports with no concomitant or co-suspected drugs	To capture reports with low likelihood of other drugs having contributed to the reaction	Reports with no concomitant or co-suspected drugs
Multiple reporting elements (MUL)	Reports fulfilling multiple defined criteria, strengthening the causality of the case	Several criteria that speak in favour of a causal relationship naturally strengthens the overall causality assessment of the individual case	Reports fulfilling at least two of the following: solely reported, dechallenge, narrative, causality probable/certain
Recent reporting (REC)	Reports entered during the last 3 years	To capture emerging safety issues; lack of recent reports may speak against a causal relationship	Reports entered during the last 3 years
Disproportional reporting (DIS)	Information on whether the drug–ADR pair is reported more often than expected	An unexpectedly large number of reports on the drug–ADR may strengthen the likelihood of a causal relationship	Disproportionate reporting as measured by a lower limit of the credibility interval for the IC either on the full dataset or on a subset of the data (for details, see Sect. 2.1.2)
Geographic spread (GEO)	Number of geographic regions contributing reports on the drug–ADR pair of interest	True ADRs might be expected to occur not just in a single geographic region	Countries with IC > 0 for the drug–ADR pair of interest[c]
Time trend (TRE)	Increase in the reporting frequency of the drug–ADR pair	Emerging safety issues may be expected to exhibit an increase in reporting with time	Growing IC values over the three 6-month periods up to the dataset end date

ADR adverse drug reaction, *IC* information component, *WHO* World Health Organization

[a] VigiBase®, the WHO global individual case safety report database, is the particular collection of individual case reports considered in this study. As of March 2014 it contained more than 8.5 million reports from 118 countries

[b] Admittedly a crude attempt to capture plausible time-to-onset, which may rather filter out implausible temporal relations, with the exception of long-latency reactions

[c] For rare drugs and ADRs this is likely to equal the number of countries with at least one report of the drug–ADR pair

MedDRA® preferred terms for which no other preferred term from the same MedDRA® high-level term was included in Section 4.8 of the 2012 European summary of product characteristics (SPC) of that drug. MedDRA®-encoded European SPCs were obtained from the PROTECT database, available at http://www.imi-protect.eu/methodsRep.shtml. All 5,544 drug–ADR pairs of the reference set were required to have reports from at least two countries in VigiBase by 31 December 2004.

2.3 Fitting the vigiRank Model

vigiRank was derived by fitting a shrinkage logistic regression model to the reference set, using all variables described in Table 1 as candidate explanatory variables. Shrinkage regression was employed on account of its ability to avoid over-fitting to sampling variability of the training data [31]. The specific type of shrinkage called 'lasso' was used, as it simultaneously selects variables and estimates their relative contributions to the predictive model; that is, it sets the weights of most explanatory variables to zero [32].

2.3.1 Variable Transformation

Prior to regression modelling, all numerical variables, i.e. all except 'disproportional reporting' and 'time trend' (see Table 1), were subjected to mathematical transformation (see Fig. 1). The main aim of this was to impose a law of diminishing returns, gradually decreasing the impact from additional reports of a given kind, e.g. with positive dechallenge. This is important in order not to exclusively rank very large case series at the top.

As seen in Fig. 1, the transformations all have an identical shape but differ in the first number of reports (or countries) to be rewarded. The shape of the transformation was selected prior to modelling and not altered in light of its results. The threshold for each covariate was determined empirically as the lowest number that applied to fewer than 20 % of a random sample of negative controls, separate to those used to subsequently fit and evaluate the model. For example, 30 % of negative controls in the random sample have two or more dechallenge reports, whereas only 19 % have three or more. Hence, the threshold for 'dechallenge' was set to three reports. Whereas the transform for 'recent reporting' follows the same shape, it yields values between −1 and 0 instead of between 0 and 1. It therefore penalizes drug–ADR pairs with few recent reports rather than rewards pairs with many recent reports.

2.3.2 Regression Modelling

All regression analysis relied on the specific implementation of lasso logistic regression provided by Friedman
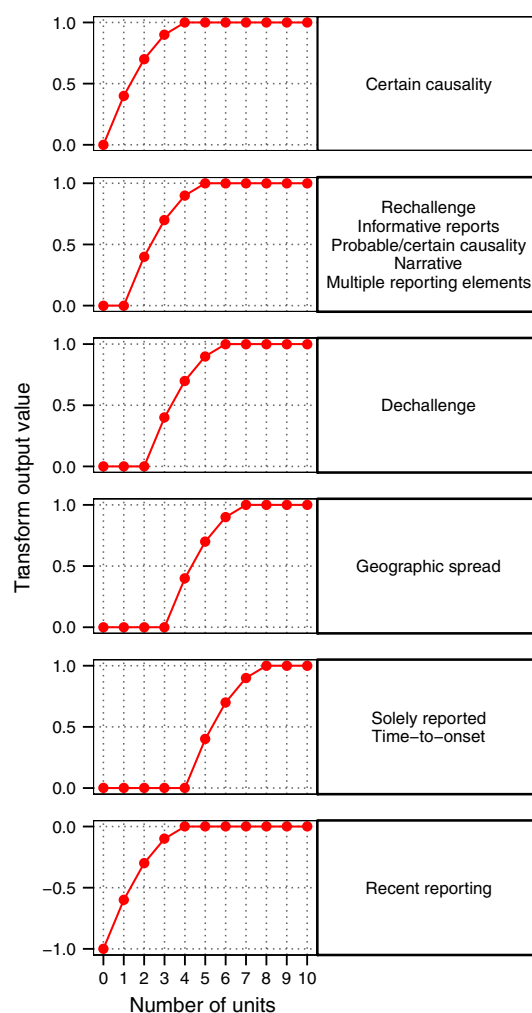

Fig. 1 Mathematical transforms used for numerical variables to gradually decrease the reward from additional units (i.e. reports or reporting countries). The respective points at which the transforms begin to increase were determined empirically based on the variables' frequency among the negative controls

et al. [33]. Prior to model fitting, all transformed variables were standardized to unit variance. However, all presented coefficients apply to the transformed variables on their original scale, as shown in Fig. 1. For details on the lasso logistic regression model, please see the Appendix.

Shrinkage regression requires selecting a degree of shrinkage that determines the model size. In this study, all considered variables were designed to be positive predictors, i.e. to independently add to the likelihood that a drug–ADR pair represents a safety signal, if at all included. Therefore, any estimated negative coefficient was taken to suggest over-fitting, and consequently the selected shrinkage was set to yield the largest possible model without negative coefficients.

2.4 Performance Evaluation

vigiRank utilizes a logistic regression model, and produces a score for each drug–ADR pair reflecting the probability that it represents an emerging safety signal (for details, see the Appendix). This score has no natural threshold, for which reason performance was evaluated in terms of receiver operating characteristic (ROC) curves, which display sensitivity and specificity at varying thresholds. The area under the ROC curve (AUC) was used as aggregate measure of predictive performance. AUC is a standard summary metric for comparisons over all possible thresholds in ROC analyses.

To study its performance relative to standard disproportionality analysis, vigiRank was compared with the IC (see Sect. 2.1.2). Following routine use, the IC was employed without adjustments or stratifications, and IC_{025} was used to decide which drug–ADR pairs to highlight at a given threshold. Comparison was also made against a crude screening strategy based on a drug–ADR pair's raw number of reports, representing the first paradigm of quantitative analysis in pharmacovigilance. The performance of all three methods was evaluated with the reference set of historic EMA safety signals as benchmark.

2.5 Cross-Validation

Because the same data were used to develop and evaluate vigiRank, there might be a risk of over-estimating its predictive performance. To manage this risk, regression modelling and performance evaluation was repeated within fivefold cross-validation. First, the entire reference set was randomly divided into five equally sized folds. Then, withholding one fold at a time, a lasso logistic model was fitted as described for the entire reference set above; notably, the degree of shrinkage was determined anew in each of the five iterations. For each of the five models, the withheld fold was used as benchmark in the evaluation.

Accurate ascertainment of predictive performance was not the only purpose of this cross-validation scheme. Investigation of vigiRank's robustness was enabled by comparing the model fitted from the entire reference set with the five models obtained in cross-validation.

3 Results

3.1 vigiRank: A Predictive Model for Emerging Safety Signals

The variables selected by lasso logistic regression for inclusion into vigiRank are shown in Fig. 2a: 'informative reports', 'recent reporting', 'disproportional reporting', 'narrative', and 'geographic spread'. Further, Fig. 2b displays the models obtained in cross-validation. Robustness does not appear to be an issue: the above five variables were selected in each of the five cross-validation iterations, with little variability in the magnitude of the estimated coefficients. The only other variable selected was 'dechallenge', which was included with small coefficients in two iterations.

Figure 3 provides a fictional example to schematically describe how vigiRank works in practice.

3.2 Overall Predictive Performance

The results from the performance evaluation based on the entire reference set are shown in Fig. 4a. The ROC curve corresponding to vigiRank is consistently above the other two, indicating that, at any given specificity, vigiRank will yield higher sensitivity than both IC_{025} and raw numbers of reports; and conversely, at any given sensitivity, vigiRank will yield higher specificity. This obviously includes the region around the natural threshold for IC_{025}, zero. Furthermore, vigiRank offered a statistically significant improvement in AUC over IC_{025} (0.775 vs. 0.736, cross-validated), which is greater in magnitude than the difference between IC_{025} and raw numbers of reports (0.736 vs. 0.707). For reference, when considered in isolation, the predictive performance of disproportionality analysis extended for subgroup discovery as described in Sect. 2.1.2, was 0.743.

For vigiRank, the mean AUC value across the five cross-validation iterations differed only marginally compared with that obtained with the entire reference set, and there was no difference at all for the other two methods (see Fig. 4b). In all five iterations, vigiRank performed best and IC_{025} second best.

3.3 Examples of Emerging Signals Highlighted or Overlooked by vigiRank

In line with vigiRank's overall better performance, some emerging signals missed by IC_{025} were highly ranked by vigiRank. Others yet obtained lower ranks, reflecting vigiRank's less-than-perfect performance as a first-pass filter. Table 2 shows the three positive controls with negative IC_{025} that obtained highest ranks and the three positive controls that obtained overall lowest ranks, according to vigiRank.

The example olanzapine–bradycardia (see Table 2) illustrates vigiRank's ability to consider local stratum-specific or adjusted effects. This allows it to highlight emerging safety signals not detected by standard disproportionality analysis: while the global IC_{025} is negative,
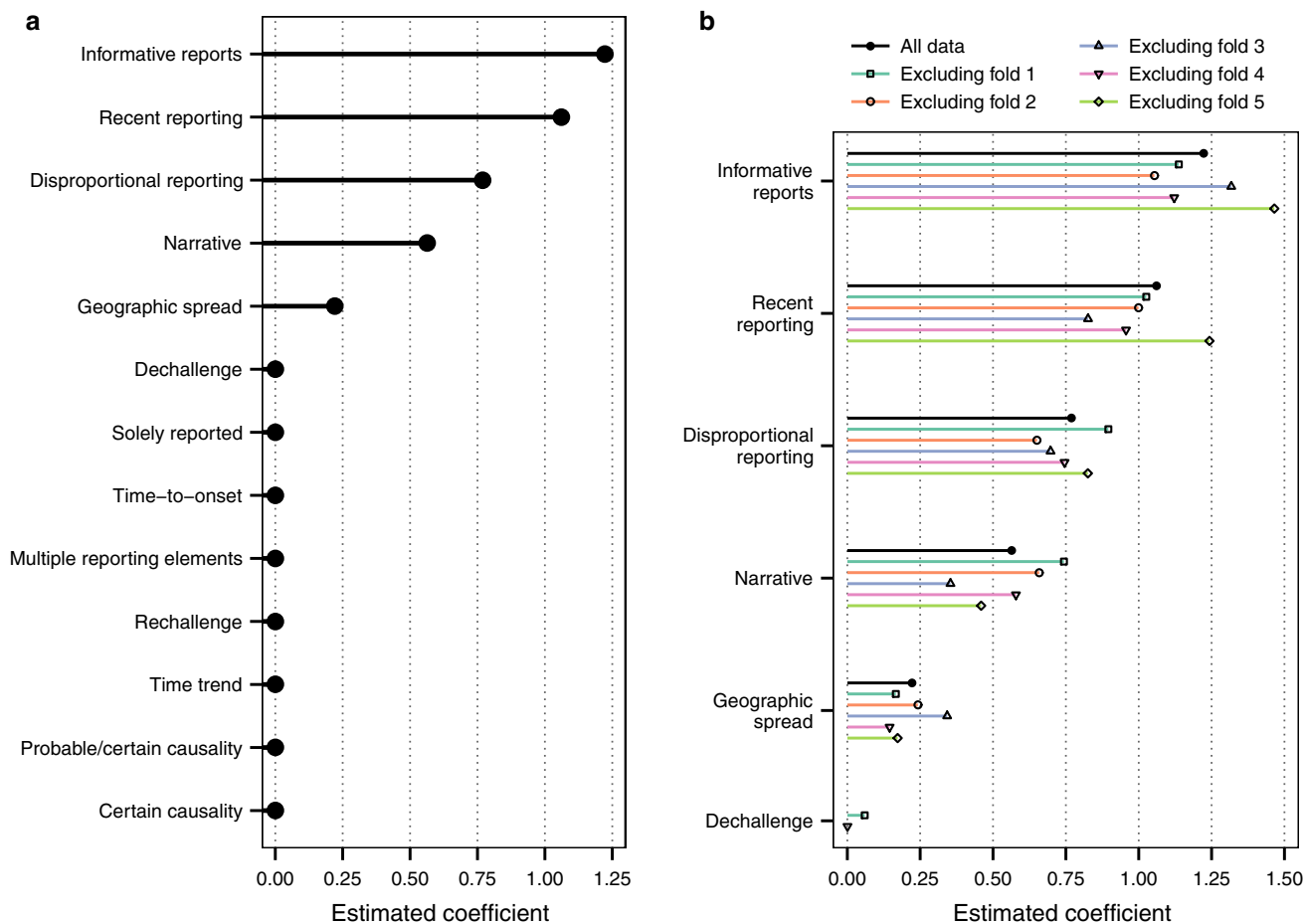
Fig. 2 a Estimated coefficients by lasso logistic regression for all considered variables. The top five variables with non-zero coefficients define our new screening algorithm, vigiRank. The model intercept was −3.45. **b** All non-zero coefficients estimated by lasso logistic regression for any variable, either based on all data (as in **a**) or when excluding a fold during fivefold cross-validation. Positive and negative controls were randomly assigned to one of five folds, so that each fold contains 20 % of the entire reference dataset

there is disproportional reporting ($IC_{005} > 0$) for the age groups 12–17 years and 18–44 years.

Even in the complete absence of disproportional reporting, vigiRank can utilize other reporting patterns to yield high ranks, as for the two clopidogrel signals in Table 2, which are driven by the other four included predictors. Each of these three signals have additional support from variables not included in this vigiRank implementation, such as 'dechallenge', 'rechallenge', 'probable/certain causality', and 'solely reported'.

The three low-ranked positive controls naturally have minimal reporting on the variables included in vigiRank. Two of them, raloxifene–arterial thrombosis and pramipexole–hyperkinesia (see Table 2), have very little empirical support overall. However, for nelfinavir–erythema multiforme, the underlying data—in particular the five positive dechallenge reactions—might have been able to support a positive statement on causality in a subsequent clinical assessment.

4 Discussion

Adverse effects from medicines are a major impediment to beneficial pharmacological therapy: they cause mortality and morbidity, and trigger patients to stop otherwise effective treatment. With many risks unknown at the time of marketing, detecting those risks as timely and accurately as possible is of utmost importance. One of the greatest opportunities towards this end is to adapt first-pass screening to account for the varying strength of individual case reports. As in clinical assessment, computational methods can then look beyond aggregate numbers and prioritize case series more likely to contain convincing reports and reporting patterns. Our results indicate that this may improve performance as much as did the previous paradigm shift from screening based on raw numbers of reports to disproportionality analysis.

Our new screening algorithm, vigiRank, considers the amount and type of information on individual reports, the
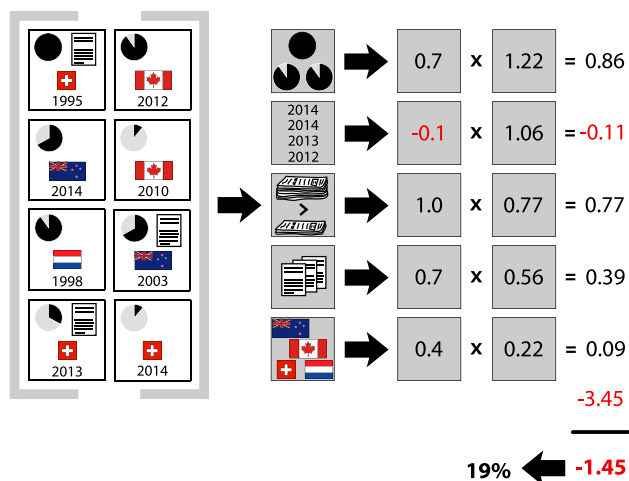
Fig. 3 Outline of how vigiRank applies to a set of eight reports on a fictional drug–adverse drug reaction pair. The *first part* shows a conceptual summary of each report. As an example, the *top left* report is a report from Switzerland that includes a case narrative, attains a vigiGrade completeness score of 1.0, and was received in 1995. The *second part* shows the raw data for each of the predictors (three informative reports, four recent reports, disproportionality = TRUE, three reports with case narratives, and four countries of origin with positive Information Component). The *third part* displays the corresponding transformed values that are multiplied with their corresponding estimated coefficients. The *fourth part* sums the independent contributions from all variables with the intercept (−3.45) to produce the overall score of −1.45 on logit scale, which corresponds to a 19 % probability. Either the score or the probability could be used for ranking purposes. Note: here the year 2014 is used as a reference point to determine whether or not a report is recent

timing and geographic spread of the case series, and the unexpectedness of the reporting pattern. Our version of disproportionality analysis goes further than what is customary today: it also identifies local associations unique to, for example, geographic regions or age groups [27]. Markers of presumed clinical relevance of individual reports were considered as potential predictors but not selected for the final model. These include explicit assessment of causality by the reporter or national center, records of dechallenge interventions where the adverse event had resolved upon end of treatment, and records of rechallenge interventions where the adverse event had reappeared upon re-exposure to the drug. The latter was identified by experts as an important aspect of strength of evidence for pharmacovigilance triages [12]. However, an important difference is that their ascertainment of positive rechallenges is based on manual review, whereas we rely on already encoded information in the database, which may not be perfectly trustworthy. As regards strong causality assessment, this may be unlikely for adverse events not yet known to be associated with the drug, as considered here.

Conversely, the predictive ability of informative reports and case narratives was not expected. These predictors measure the amount of information, but not whether it

supports a causal association. As a case in point, the vigiGrade completeness measure looks at whether there is information on time to onset, but not whether the reported time to onset is suggestive of a causal association. Of course, the amount of information on a report may correlate with the strength of suspicion and perceived importance of the case. Viewed differently, case series with no informative reports or case narratives may be less likely to reflect true adverse effects.

An advantage of vigiRank over rule-based alternatives is that it imposes no hard requirement on any individual predictor. Lack of informative reports can be compensated for, for example, by disproportional reporting, narratives, and recent reporting; and fewer than expected reports overall can be compensated for, for example, by informative reports and geographic spread. The previously described triage for adverse drug interaction surveillance in VigiBase [24] conforms to the vigiRank framework. One important new development in the implementation presented here compared with that for interactions is the principle of diminishing returns for additional reports fulfilling the same criteria (recent, informative, or with narratives), and for additional countries with disproportional reporting rates. This is important since emerging safety signals must be detected early, and it would be unwise to place undue focus on series of very large numbers of reports. By design, vigiRank is less likely than basic disproportionality analysis to highlight report series with as few as three reports. Three reports has been taken as a rule of thumb of minimal support for a safety signal, but the original publication from which this derives actually calls for three strongly supportive cases, which is quite different [23].

We developed and evaluated a first-pass screening algorithm for VigiBase, which is the world's largest database of its kind. Its implementation in a different context would require the predictor variables to be adapted. For example, an informative report may be defined as one that carries information in a minimum number of record fields important for the database of interest, and geographic spread in a national dataset could be measured in terms of states or regions. As for unexpectedness, a different measure of disproportionality and threshold can replace our proposed combination of global and local IC analysis. We believe that vigiRank would also bring value over basic disproportionality analysis in other databases. Ideally, the weights of the model should be optimized to the new setting. If so, the transformations for predictor variables based on numbers of reports and geographical units could be adapted to peak earlier for smaller databases. With less resource, one may consider recycling the weights presented here. In either case, empirical evaluation of performance for the dataset of interest is crucial.
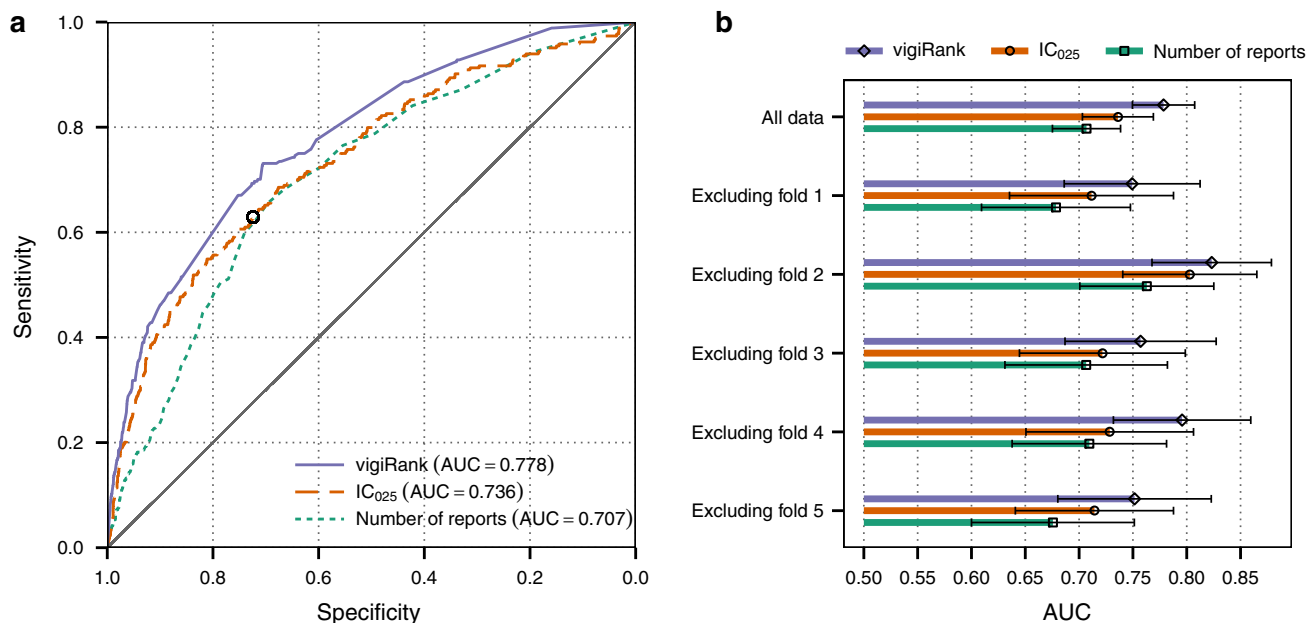
Fig. 4 a Receiver operating characteristic curves for our new screening algorithm, vigiRank, standard disproportionality analysis (IC_{025}), and raw numbers of reports, relative to the benchmark based on historic European Medicines Agency safety signals. The difference between vigiRank and IC_{025} is statistically significant ($p < 0.001$ using DeLong's test [36]). The *circle* corresponds to the standard threshold for IC_{025}, 0, and the 45° *line* corresponds to random guessing. **b** Area under the curve (AUC) values for the three methods from the evaluation on all data as well as from the individual iterations of the cross-validation. The *error bars* indicate 95 % confidence intervals. The mean values over the cross-validation folds are 0.775, 0.736, and 0.707 for vigiRank, IC_{025}, and raw numbers of reports, respectively. *IC* information component, IC_{025} is the lower limit of the two-sided 95 % credibility interval for the IC disproportionality measure

The main limitation of our study is the lack of an external benchmark to evaluate predictive performance. Cross-validation indicates that vigiRank is not over-fitted to sampling variability of the reference dataset: the mean cross-validated AUC is marginally below that for the reference dataset as a whole, the selected predictors and their coefficients are remarkably consistent over the five cross-validation iterations, and vigiRank consistently outperforms IC_{025} over all iterations. However, vigiRank could be over-fitted to the nature of the reference set in question. The strengths of the reference set from Alvarez et al. [28] include its focus on risks identified post-marketing, and the possibility to backdate the analysis to before their discovery. However, it is restricted to European safety signals in the early 2000s, and these could differ systematically from safety signals in different regions and time periods. Our positive controls do have lower proportions of North American reports than does VigiBase overall, but beyond that European reports are not over-represented compared with other continents. This is important since the proportion of informative reports in VigiBase is higher from Europe than from other regions [25].

This implementation of vigiRank is not a proposed final solution, but a first step towards a new paradigm that will allow more effective automated screening of individual case reports. It has clear advantages over basic disproportionality analysis, conceptually as well as empirically, and should be adopted for routine use. However, further enhancements must follow. Other predictive models should be considered, including ensemble methods, which have proven powerful in other settings [34]. Other benchmarks should be utilized to ensure generalizability. Some of the predictors developed here, but not selected, should be considered further, particularly 'dechallenge', which was included in two of the cross-validation models. Additional 'negative' predictors aimed at eliminating false positives, similar to (lack of) recent reports in the current version of vigiRank is an interesting possibility. As an example, lack—or a very low proportion—of reports without co-reported drugs could potentially eliminate false positives. We should also consider novel predictors based on recent innovations such as the measure of unexpected reported time to onset proposed by Van Holle et al. [22]. Combination with multivariate methods to account for co-reported drugs and indications for treatment at the report level through regression [17, 19, 20] or propensity score models [18] is particularly interesting, since these aspects are not captured by any of the predictors in the current model, and have yielded similar increases in performance [18, 20]. As they are conceptually distinct, their combination may well be synergistic.

Table 2 Underlying data on the three positive controls not detected by standard disproportionality (IC$_{025}$ > 0) that were highest ranked by vigiRank, and the three positive controls that were lowest ranked by vigiRank

Drug	Adverse reaction	Nr. of reports	Data on considered variables[a]													vigiRank rank[b]	IC$_{025}$	IC$_{025}$ rank[b]
			INF	NAR	DCH	RCH	CAU	CAU+	TTO	SOL	MUL	REC	DIS	GEO	TRE			
Olanzapine	Bradycardia	105	4 (1.10)	7 (0.56)	25	4	4	2	46	32	12	67 (0.00)	Yes[c] (0.77)	10 (0.22)	No	54	−0.30	1882
Clopidogrel	Stevens-Johnson syndrome	25	5 (1.22)	5 (0.56)	5	0	3	0	13	4	3	22 (0.00)	No (0.00)	5 (0.15)	No	157	−0.73	2448
Clopidogrel	Myalgia	69	4 (1.10)	15 (0.56)	12	3	5	0	21	11	9	48 (0.00)	No (0.00)	6 (0.20)	No	173	−0.82	2582
Raloxifene	Arterial thrombosis	3	0 (0.00)	1 (0.00)	0	0	0	0	0	2	1	1 (−0.42)	No (0.00)	2 (0.00)	No	4720	−1.04	2862
Pramipexole	Hyperkinesia	3	0 (0.00)	0 (0.00)	0	0	0	0	2	0	0	1 (−0.42)	No (0.00)	2 (0.00)	No	4720	−1.71	3712
Nelfinavir	Erythema multiforme	10	0 (0.00)	0 (0.00)	5	1	1	1	6	1	1	0 (−1.06)	No (0.00)	3 (0.00)	No	5291	−1.15	3006

ADR adverse drug reaction, *IC* information component, *INF* informative reporting

[a] Numbers in parentheses indicate how much the variables contribute to the algorithm's score, for a particular drug–ADR pair. For example, the four reports on INF for olanzapine–bradycardia are transformed to 0.9 (see Fig. 1) and multiplied with 1.22 (the INF coefficient, see Fig. 2a) to yield 1.10. Full names of all variables are provided in Table 1

[b] Out of 5,544 drug–ADR pairs in total. vigiRank's rank for a given drug–ADR pair is based on the predicted probability (see Fig. 3)

[c] IC$_{005}$ > 0 in the two age groups 12–17 years and 18–44 years (see Sect. 2.1.2)

5 Conclusions

Accounting for multiple aspects of strength of evidence has clear conceptual and empirical advantages over disproportionality analysis. vigiRank is a first-of-its-kind predictive model to factor in report quality and content in first-pass screening to better meet tomorrow's post-marketing drug safety surveillance needs.

Appendix: vigiRank as a lasso logistic regression model

To appreciate the lasso logistic regression model chosen for the purpose of deriving the new screening algorithm vigiRank in this study, consider first the ordinary logistic regression model:

$$\log \frac{P(y|x)}{1 - P(y|x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2}$$

where p denotes the number of candidate variables, x_j denotes the jth predictor variable, and y the outcome (in our case emerging safety signal or not). The β_j's can be viewed as log odds ratios adjusted for the impact of the other covariates [35].

As shown in Fig. 1, all variables except those that are binary undergo transformation. This can be described in the following way:

$$\log \frac{P(y|x)}{1 - P(y|x)} = \beta_0 + \beta_1 f_1(x_1) + \cdots + \beta_p f_p(x_p) \tag{3}$$

where f_j is the transformation applied to the jth variable. (Trivially, for the binary variables $f(x) = x$.)

Now, lasso logistic regression extends standard logistic regression by imposing an L$_1$ constraint on the coefficients:

$$\sum_{j=1}^{p} \left| \beta_j \right| \le t \qquad (4)$$

There are multiple ways to estimate the β_j coefficients under this constraint. In this study, the algorithm provided by Friedman et al. [33] was used, which allows the coefficients to be estimated for the whole path of possible values for the tuning parameter t in Eq. 4. This parameter determines the amount of shrinkage and therefore the size of the model; see Sect. 2.3.2 for a description of how the model size of vigiRank was selected.

Since vigiRank is a logistic regression model, once its coefficients have been estimated as described above, it can be used to make predictions for arbitrary drug–ADR pairs regarding their likelihood of representing emerging safety signals. Solving Eq. 3 for $P(y|x)$ yields:

$$P(y|x) = \frac{\exp(\beta_0 + \beta_1 f_1(x_1) + \cdots + \beta_p f_p(x_p))}{1 + \exp(\beta_0 + \beta_1 f_1(x_1) + \cdots + \beta_p f_p(x_p))} \qquad (5)$$

Figure 3 provides a detailed description of how Eq. 5 is applied to a set of eight reports for a fictional drug–ADR pair, using the actual transformations f_j and coefficients β_j of the final vigiRank model. In practice, there are a large number of drug–ADR pairs, which are ranked according to their respective predicted $P(y|x)$.

References

1. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta- analysis of prospective studies. J Am Med Assoc. 1998;279(15):1200–5.

2. Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18,820 patients. Br Med J. 2004;329(7456):15–9.

3. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. J Am Med Assoc. 1997;277(4):301–6.

4. Stricker BH, Psaty BM. Detection, verification, and quantification of adverse drug reactions. Br Med J. 2004;329(7456):44–7.

5. CIOMS Working Group XIII. Practical aspects of signal detection in pharmacovigilance. Geneva: CIOMS; 2010.

6. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the united states, 1969–2002: the importance of reporting suspected reactions. Arch Intern Med. 2005;165(12):1363–9.

7. Vandenbroucke JP. In defense of case reports and case series. Ann Intern Med. 2001;134(4):330–4.

8. Coloma PM, Trifirò G, Schuemie MJ, Gini R, Herings R, Hippisley-Cox J, et al. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? Pharmacoepidemiol Drug Saf. 2012;21(6):611–21.

9. van Puijenbroek EP, van Grootheest K, Diemont WL, Leufkens HG, Egberts AC. Determinants of signal selection in a spontaneous reporting system for adverse drug reactions. Br J Clin Pharmacol. 2001;52(5):579–86.

10. Ståhl M, Lindquist M, Edwards IR, Brown EG. Introducing triage logic as a new strategy for the detection of signals in the WHO drug monitoring database. Pharmacoepidemiol Drug Saf. 2004;13(6):355–63.

11. Waller P, Heeley E, Moseley J. Impact analysis of signals detected from spontaneous adverse drug reaction reporting data. Drug Saf. 2005;28(10):843–50.

12. Levitan B, Yee CL, Russo L, Bayney R, Thomas AP, Klincewicz SL. A model for decision support in signal triage. Drug Saf. 2008;31(9):727–35.

13. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54(4):315–21.

14. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10(6):483–6.

15. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat. 1999;53(3):177–90.

16. Hauben M, Madigan D, Gerrits CM, Walsh L, van Puijenbroek EP. The role of data mining in pharmacovigilance. Expert Opin Drug Saf. 2005;4(5):929–48.

17. Caster O, Norén GN, Madigan D, Bate A. Large-scale regression-based pattern discovery: the example of screening the WHO global drug safety database. Stat Anal Data Min. 2010;3(4):197–208.

18. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med. 2012;4(125):125ra31.

19. Berlin C, Blanch C, Lewis DJ, Maladorno DD, Michel C, Petrin M, et al. Are all quantitative postmarketing signal detection methods equal? Performance characteristics of logistic regression and multi-item Gamma Poisson Shrinker. Pharmacoepidemiol Drug Saf. 2012;21(6):622–30.

20. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clin Pharmacol Ther. 2013;93(6):539–46.

21. Caster O, Norén GN, Madigan D, Bate A. Logistic regression in signal detection: another piece added to the puzzle. Clin Pharmacol Ther. 2013;94(3):312.

22. Van Holle L, Zeinoun Z, Bauchau V, Verstraeten T. Using time-to-onset for detecting safety signals in spontaneous reports of adverse events following immunization: a proof of concept study. Pharmacoepidemiol Drug Saf. 2012;21(6):603–10.

23. Edwards IR, Lindquist M, Wiholm BE, Napke E. Quality criteria for early signals of possible adverse drug reactions. Lancet. 1990;336(8708):156–8.

24. Strandell J, Caster O, Hopstadius J, Edwards IR, Norén GN. The development and evaluation of triage algorithms for early discovery of adverse drug interactions. Drug Saf. 2013;36(5):371–88.

25. Bergvall T, Norén GN, Lindquist M. vigiGrade: a tool to identify well-documented individual case reports and highlight systematic data quality issues. Drug Saf. 2014;37(1):65–77.

26. Norén GN, Hopstadius J, Bate A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. Stat Methods Med Res. 2013;22(1):57–69.

27. Hopstadius J, Norén GN. Robust discovery of local patterns: Subsets and stratification in adverse drug reaction surveillance. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. ACM; Miami (FL); 2012:265–74.

28. Alvarez Y, Hidalgo A, Maignen F, Slattery J. Validation of statistical signal detection procedures in EudraVigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling. Drug Saf. 2010;33(6):475–87.

29. Lindquist M. VigiBase, the WHO global ICSR database system: basic facts. Drug Inf J. 2008;42(5):409–19.

30. Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Discov. 2007;14(3):305–28.

31. Copas JB. Regression, prediction and shrinkage. J R Stat Soc Ser B Methodol. 1983;45(3):311–54.

32. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.

33. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.

34. Jahrer M, Töscher A, Legenstein R. Combining predictions for accurate recommender systems. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. ACM; Washington, DC; 2010:693–702.

35. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley; 1989.

36. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.