THERAPEUTIC ADVANCES in
*Drug Safety*

*Original Research*

# Evaluation of four machine learning models for signal detection

*Ther Adv Drug Saf*

2023, Vol. 14: 1–15

DOI: 10.1177/20420986231219472



**Daniel G. Dauner, Eleazar Leal, Terrence J. Adam, Rui Zhang and Joel F. Farley**

***Abstract***

**Background:** Logistic regression-based signal detection algorithms have benefits over disproportionality analysis due to their ability to handle potential confounders and masking factors. Feature exploration and developing alternative machine learning algorithms can further strengthen signal detection.

**Objectives:** Our objective was to compare the signal detection performance of logistic regression, gradient-boosted trees, random forest and support vector machine models utilizing Food and Drug Administration adverse event reporting system data.

**Design:** Cross-sectional study.

**Methods:** The quarterly data extract files from 1 October 2017 through 31 December 2020 were downloaded. Due to an imbalanced outcome, two training sets were used: one stratified on the outcome variable and another using Synthetic Minority Oversampling Technique (SMOTE). A crude model and a model with tuned hyperparameters were developed for each algorithm. Model performance was compared against a reference set using accuracy, precision, F1 score, recall, the receiver operating characteristic area under the curve (ROCAUC), and the precision-recall curve area under the curve (PRCAUC).

**Results:** Models trained on the balanced training set had higher accuracy, F1 score and recall compared to models trained on the SMOTE training set. When using the balanced training set, logistic regression, gradient-boosted trees, random forest and support vector machine models obtained similar performance evaluation metrics. The gradient-boosted trees hyperparameter tuned model had the highest ROCAUC (0.646) and the random forest crude model had the highest PRCAUC (0.839) when using the balanced training set.

**Conclusion:** All models trained on the balanced training set performed similarly. Logistic regression models had higher accuracy, precision and recall. Logistic regression, random forest and gradient-boosted trees hyperparameter tuned models had a PRCAUC $\geqslant$ 0.8. All models had an ROCAUC $\geqslant$ 0.5. Including both disproportionality analysis results and additional case report information in models resulted in higher performance evaluation metrics than disproportionality analysis alone.

## Plain language summary

### Evaluating new methods to detect potential harmful adverse drug events in spontaneous report databases

**Background:**The Food and Drug Administration (FDA) adverse event reporting system (FAERS) is a database that contains adverse event reports, medication error reports, and product quality complaints. The FDA uses statistical methods to identify potentially harmful drug-adverse event combinations, also known as signals, within FAERS. This study compared several different methods to identify harmful drug-related events from

Correspondence to:
**Daniel G. Dauner**
Department of Pharmaceutical Care and Health Systems, College of Pharmacy, University of Minnesota Duluth, 232 Life Science, 1110 Kirby Drive, Duluth, MN 55812, USA
**ddauner@d.umn.edu**

**Eleazar Leal**
Department of Computer Science, Swenson College of Science and Engineering, University of Minnesota Duluth, Duluth, MN, USA

**Terrence J. Adam**
Department of Pharmaceutical Care and Health Systems, College of Pharmacy, Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA

**Rui Zhang**
Division of Computational Health Sciences, Department of Surgery, University of Minnesota, Minneapolis, MN, USA

**Joel F. Farley**
Department of Pharmaceutical Care and Health Systems, College of Pharmacy, University of Minnesota, Minneapolis, MN, USA

adverse event reports in FAERS. The performance of each method was compared to see which method worked best.
**Methods:**Logistic regression-based signal detection methods have demonstrated superior performance due to their ability to handle variables that can distort the effect of other variables or hide potential associations. The development of other machine learning models is of interest. Machine learning models can define complex relationships between risk factors and outcomes. Our objective was to compare the signal detection performance of multiple models.
**Results:**Our study show that two models (logistic regression and random forest) were better at identifying true signals than other models.
**Conclusions:**The four methods have differing abilities on how well they identify adverse drug events in voluntarily reported surveillance data. Including both results of searches for unexpected associations between drugs and adverse events and additional case report information in models resulted in identifying more true signals than unexpected association results alone. The models can be replicated or modified for use by drug safety programs.

***Keywords:*** adverse drug events, machine learning, pharmacovigilance, signal detection

Received: 13 July 2023; revised manuscript accepted: 17 November 2023.

## Introduction

Disproportionality algorithms quantify the unexpectedness of specific drug-event combination pairs (DECs) in a spontaneous adverse drug event (ADE) report database. Unexpectedness suggests the number of reports for a specific DEC is higher than expected and can provide a signal that warrants clinical review and further investigation.[1] However, the algorithms give the same weight to information from all reports in a database, which may result in signals being masked or false positives being flagged as signals.[2] Multiple groups have found that logistic regression (LR)-based signal detection algorithms are superior to disproportionality analysis due to their ability to account for potential confounders and masking factors.[2–6]

Despite the demonstrated advantages of LR, it does have limitations. First, interaction terms need to be programmed into the LR model to assess for interacting independent variables. Second, LR does not work well with large databases and outlier observations. Third, LR does not handle complex, nonlinear relationships; or correlated independent variables.[7–9]

Machine learning and deep learning algorithms are able to define complex relationships between risk factors and outcomes.[10] They have mostly been used to help predict ADEs during drug discovery and preclinical trials. Wang *et al.* used a deep neural network to detect potential ADEs in new drugs. Study results showed the overall performance of the model had a mean average precision of 0.523 and the area under the curve (AUC) was 0.844 for ADE prediction.[11] Ietswaart *et al.* developed random forest (RF) models to predict ADEs from *in vitro* pharmacological profiles using *in vitro* pharmacology assay data from Novartis and ADE data from Food and Drug Administration adverse event reporting system (FAERS). The models had high accuracy and precision ranging between 0.9 and 1, recall of 0.6 and an AUC of 0.8.[12]

Two studies have used machine learning algorithms with FAERS data for pharmacovigilance purposes.[13,14] Chen *et al.* developed LR, support vector machine (SVM), RF, and gradient-boosted tree (GBT) models to predict hospitalizations and deaths based on patient demographics and drugs. The accuracy was between 73% and 75%

for predicting hospitalization and 68% and 76% for predicting deaths. The recall (90–99%) and F1 score (83–84%) were also higher for models predicting hospitalizations, and the precision was similar. Part of the difference in performance could be due to the relatively low number of deaths in the data.[13] Pham *et al.* compared the accuracy of multiple methods to detect DEC associations. The methods included frequentist and Bayesian disproportionality analysis, multivariate methods, and machine learning algorithms. Most AUC values were greater than 0.65, with Bayesian confidence propagation neural network having the highest AUC (0.693) and RF the lowest (0.521).[14]

The objective of this study was to compare the performance of LR, GBT, RF, and SVM for signal detection utilizing data from FAERS. Twelve features were used for model development. Accuracy, precision, F1 score, recall, the receiver operating characteristic AUC (ROCAUC), and the precision-recall curve AUC (PRCAUC) were used to compare the performance of the models against the testing set portion of the reference set.

## Methods

### *Data sources*

A cross-sectional study was conducted. The publicly available FAERS quarterly data extract files from 1 October 2017, through 31 December 2020, were downloaded. The Demographic, Drug, Outcome, Reaction, Therapy and Indication files were used. The Demographic, Drug, Outcome and Reaction files were linked on the primary ID (PRIMARYID). The Drug, Therapy and Indication files were linked on both the primary ID (PRIMARYID) and drug sequence (DRUG_SEQ) variables.[15] Deduplication was performed by selecting the highest PRIMARYID for each report. Only the primary suspect drug from a report (ROLE_COD = PS) was included in the analysis. Secondary suspect, concomitant or interacting drugs were excluded in efforts to reduce noise in the data due to the uncertainty of the association between the drug and the ADE.[12,16,17] All ADEs listed on a report were included, and ADE terms were standardized using the Medical Dictionary for Regulatory Activities (MedDRA) preferred terms listed in the Reaction file. Reports missing a primary suspect drug or an ADE were excluded. Generic names were used to identify drugs, and all ADE and drug names were converted to upper case text for standardization.

### *Variables*

Table 1 includes the features included in this analysis. We developed a report completeness measure based on work by The Uppsala Monitoring Centre and the Pharmacovigilance Programme of India to quantify the amount of information available in an ADE report.[18,19] The features used are displayed in Table 2. Time-to-onset is defined as the time from treatment initiation to the suspected ADE. The completeness of report score starts at 1 and for every missing variable the corresponding penalty factor (Table 2) is applied. The score is calculated using equation (1),

$$C = \prod_{i=1}^{7}(1 - P_i), \quad (1)$$

where $P_i$ is the penalty listed in Table 2 for variable $i$.[18] The completeness of report score ranges from a minimum of $1 \times 0.5 \times 0.7^3 \times 0.9^3 = 0.125$ to a maximum of 1 (zero penalties imposed). A report was considered serious if OUTC_COD contained a valid value. The reporter was considered a healthcare provider if OCCP_COD equaled physician (MD), pharmacist (PH) or other healthcare professional (OT).[15] Disproportionality signals from multi-item gamma Poisson shrinker (MGPS), proportional reporting ratio (PRR), and subgrouped PRR analyses equalled 1 if a signal was identified and 0 if not. All numeric variables, except for the disproportionality measures, were standardized by subtracting the mean and dividing by the standard deviation. All data preparation and wrangling were conducted using R (v4.0.2).

### *Reference data set*

A reference set of positive and negative controls was developed to evaluate and compare multiple SDAs as part of a larger study examining pharmacovigilance for direct-acting antivirals used for the treatment of chronic hepatitis C virus infection.[20] The reference set focused on the following ADEs: dysglycaemia, hepatic decompensation and hepatic failure, and angioedema.[21–23] A reference set was developed to evaluate the ability of models to detect these ADEs. It included nine

**Table 1.** Description of features from FAERS included in models.[15]

| Feature | Data element | Feature definition | Feature coding |
|---|---|---|---|
| Report completeness | N/A | Number of informative or complete reports | Numeric |
| Dechallenge | DECHAL | Was there a positive dechallenge | 0 = No/missing, 1 = Yes |
| ADE report type | REPT_COD | Type of ADE report | Number reports expedited, periodic, or direct |
| Seriousness | OUTC_COD | Seriousness of outcome resulting from an ADE [death (DE), life-threatening (LT), hospitalization (HO), disability (DS), congenital anomaly (CA), intervention required to prevent permanent impairment or damage (RI), other serious or an important medical event (OT)] | 0 = Non-serious, 1 = Serious |
| Reporter | OCCP_COD | Occupation of reporter listed on ADE report | 0 = Non-healthcare worker, 1 = Healthcare worker |
| Recent reporting | FDA_DT | ADE reports from last 18 months | 0 = Not within last 18 months, 1 = Within last 18 months |
| MGPS | N/A | Signal meeting MGPS thresholds | 0 = No signal, 1 = Disproportionality signal |
| PRR | N/A | Signal meeting PRR thresholds | 0 = No signal, 1 = Disproportionality signal |
| Subgrouped PRR | N/A | Signal meeting age or sex subgrouped PRR thresholds | 0 = No signal, 1 = Disproportionality signal |

ADE, adverse drug event; FAERS, Food and Drug Administration adverse event reporting system; MGPS, multi-item gamma Poisson shrinker; PRR, proportional reporting ratio.

MedDRA preferred terms: angioedema, ascites, encephalopathy, hepatic encephalopathy, hyperglycaemia, hypoglycaemia, jaundice, oesophageal varices haemorrhage and varices oesophageal. Positive controls are known associated DECs.[2,24,25] Negative controls included drugs that do not include one of the nine preferred terms and no other MedDRA preferred term from the same MedDRA high-level term listed in their prescribing information.[5,24] A control variable was attached to each DEC to classify it as either a positive control (1) or negative control (0). The reference set included 155 DECs from 60 drugs with 110 DECs for positive controls and 45 DECs for negative controls (Supplemental Material 1).[20]

### *Statistical analysis*

*Disproportionality analysis.* Disproportionality analysis was conducted utilizing PRR, PRR subgrouped by age or sex, and MGPS. For PRR analyses, a signal was defined by the accepted thresholds of PRR $\geq 2$, number of reports $\geq 3$ and a $\chi^2 \geq 4$.[26] A subgrouped PRR analysis was conducted for each age and sex, and a signal for a DEC was counted if it met the signal criteria within any strata. For the MGPS analysis, a signal was defined as a DEC with a lower 95% confidence interval limit $\geq 2$.[27,28] Proportional reporting ratio and MGPS analyses represented frequentist and Bayesian disproportionality analyses, respectively, in this study.[29] All DECs

**Table 2.** Description of features included in the report completeness score.

| Feature | Description | Notes | Penalty (%) |
|---|---|---|---|
| Time-to-onset | Time from start of treatment to reported ADE | Missing or incomplete information to assess if drug was started before the ADE | 50 |
| | | If incomplete dates or time difference is greater than 120 days | 30 |
| | | If time difference is 0 | 10 |
| Indication | Indication for drug | N/A | 30 |
| Sex | Patient sex | Unknown sex is treated as missing | 30 |
| Age | Patient age at time of reported ADE | Unknown age is treated as missing | 30 |
| Dose | Dose of drug | If either dose amount or units are missing | 10 |
| Reporter | Occupation of who reported the ADE | N/A | 10 |
| Report type | Expedited, periodic or direct | N/A | 10 |

ADE, adverse drug event.

entered into FAERS during the study period were included in the disproportionality analyses. Disproportionality analyses were performed using R (R Core Team, v4.0.2).

*Models.* The dataset was split into training (80%) and testing (20%) sets. The outcome is imbalanced, containing 72.2% positive controls and 27.8% negative controls. To maintain the distribution during model building, the training and testing sets were stratified on the control variable using the stratify parameter in the train_test_split() Python function to ensure both sets have the same distribution as the dataset. This training set will be referred to as the balanced training set.

Due to the outcome imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was also assessed. In SMOTE, the k-nearest neighbours belonging to the minority class are determined for each minority observation. Then a synthetic minority observation at some intermediate point along the line joining $x$ to one of its randomly chosen k-nearest neighbour, $x_k$, is generated.[9,30] In this study, the negative controls were the minority class. LR, GBT, RF and SVM models were trained using the balanced and SMOTE training sets.

Models were optimized by fine-tuning the algorithm hyperparameters, including the regularization parameters of LR and SVM; and the number of trees, features to consider, tree depth and samples for GBT and RF. A randomized search on hyperparameters was conducted using RandomizedSearchCV() utilizing a repeated ($n=3$), stratified fivefold cross-validation. Lasso regression was used to identify significant features in the LR model. Feature importance was used to identify the most important features in the GBT and RF models and was computed as the (normalized) total reduction of the Gini impurity caused by that feature. The higher the value the more important the feature.[31] Feature coefficients that are not either equal to or near zero were used to identify the most important features in the SVM model. For each model, both a crude model without hyperparameter tuning or feature selection (crude) and a model with tuned hyperparameters and feature selection (HPT) were run. Machine learning models were built, trained and tested using Python (Python Software Foundation, v3.9.7) *via* the Spyder IDE (v5.1.5).

### *Performance evaluation*

Accuracy, precision, recall, F1 score, ROCAUC and PRCAUC were used to evaluate and compare the performance of the classification models. These metrics provide important context regarding the ability of the models to correctly identify true positives and true negatives. Both ROCAUC and PRCAUC were calculated to get a more complete assessment and comparison of the models. The PRCAUC is better at evaluating a model's ability to identify true positives and provides a better estimation of performance in unbalanced datasets.[32,33] Recall, precision, ROCAUC and PRCAUC were the primary metrics used for comparing model performance against the testing set portion of the reference set. Performance evaluation was performed using Python (Python Software Foundation, v3.9.7) *via* the Spyder IDE (v5.1.5). The reporting of this study conforms to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (Supplemental Material 2).[34]

## Results

After the FAERS quarterly data extract files were cleaned and merged, the data set contained 17,702,846 reported DECs from 5,497,137 ADE reports. Next, in order to identify the positive and negative control DECs in FAERS, the FAERS data was merged with the reference set by DEC to attach a control label for classification. This resulted in 10,282 reported DECs from 10,079 ADE reports being labelled either a positive control (1) or a negative control (0). Four DECs (fluoxetine-jaundice, fluoxetine-oesophageal varices haemorrhage, gabapentin-varices oesophageal and metformin-oesophageal varices haemorrhage) contained in the reference set were not in the downloaded FAERS data. Finally, after aggregating the 10,282 DECs there were 109 positive control DECs and 42 negative control DECs.

### *Feature inclusion and importance*

For the LR model trained on the balanced training set, 9 of 12 features were selected by lasso LR for inclusion: dechallenge, number of expedited, periodic and direct reports; report completeness, PRR signal, MGPS signal, PRR age subgrouped signal and PRR sex subgrouped signal. For the LR model trained on the SMOTE training set only two features were selected by lasso LR for inclusion: dechallenge and reporter. The most important features for the GBT and RF models trained on the balanced and SMOTE training sets are in Figure 1(a) and (b) and Figure 2(a) and (b), respectively. All GBT and RF models had the same top seven most important features: number of periodic reports, dechallenge, recent reporting, reporter, number of expedited reports, seriousness and report completeness. The most important features for the SVM models trained on the balanced and SMOTE training sets are in Figure 3(a) and (b), respectively. There were three features (number of direct reports, sex subgrouped PRR signal and age subgrouped PRR signal) that were not important for the SVM model trained on the balanced training set, and only MGPS signal was not an important feature when training on the SMOTE training set.

### *Performance evaluation*

The performance evaluation metrics for each of the models are in Table 3. Overall, the GBT crude model trained on the SMOTE training set had the highest ROCAUC (0.657), and the GBT HPT model trained on the balanced training set (0.646) was the second highest. The range of ROCAUC values for models trained with the balanced training set was 0.08 (0.566, 0.646) with the GBT crude model achieving 0.566 and the GBT HPT model achieving 0.646. The range of ROCAUC values for models trained with the SMOTE training set was 0.157 (0.5, 0.657) with the SVM crude model achieving 0.5 and the GBT crude model achieving 0.657.

When examining PRCAUC, the SVM crude model trained on the SMOTE training set had the highest PRCAUC (0.855), and the RF HPT model trained on the SMOTE training set (0.848) was the second highest. It should be noted that the SVM crude model trained on the SMOTE training set did not predict any positive control outcomes. The range of PRCAUC values for models trained with the balanced training set was 0.099 (0.740, 0.839) with the GBT crude model achieving 0.740 and the RF crude model achieving 0.839. The range of PRCAUC values for models trained with the SMOTE training set was 0.07 (0.785, 0.855) with the LR HPT model achieving 0.785 and the SVM crude model achieving 0.855.

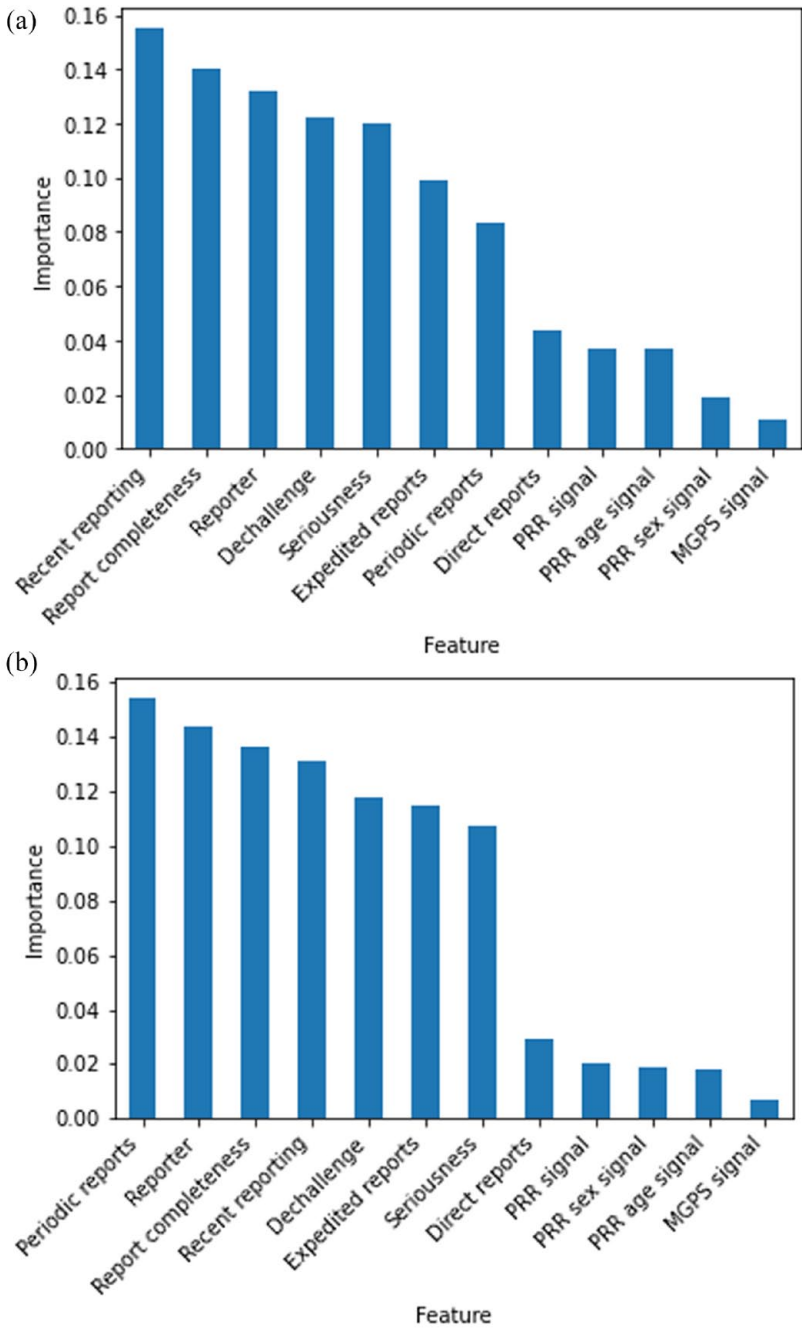


**Figure 1.** Feature importance for gradient-boosted tree model trained on the (a) balanced training set and (b) SMOTE training set.
PRR age signal and PRR sex signal are signal identified when subgrouping PRR by age and sex, respectively.
PRR, proportional reporting ratio; MGPS, multi-item gamma Poisson shrinker; SMOTE, Synthetic Minority Oversampling Technique.

The remaining performance evaluation metrics were higher for models trained on the balanced training set. Their accuracy, F1 score, and recall were higher compared to models trained on the SMOTE training set. The precision was higher for the LR and GBT models in the balanced training set, higher for RF in the SMOTE training set, and mixed between the two training sets for SVM (Table 3).

When focusing on models trained on the balanced training set, the LR, RF, and GBT HPT

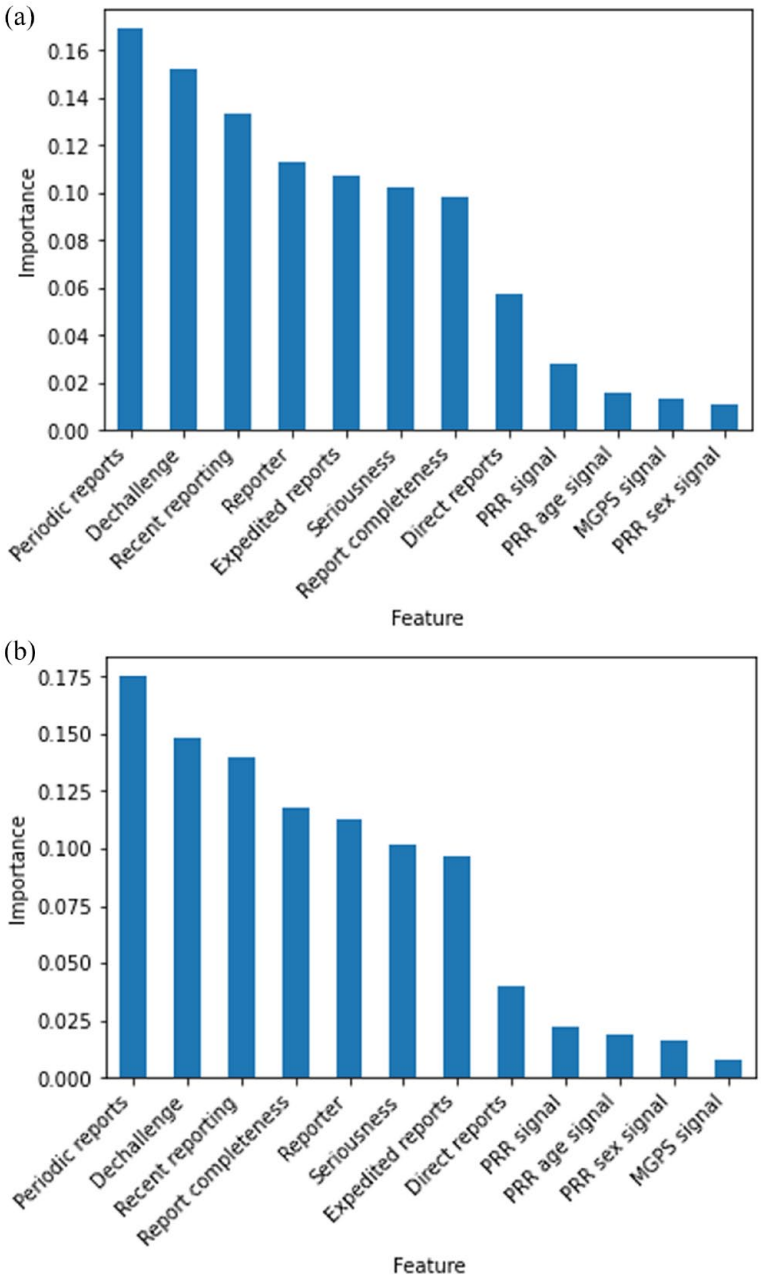


**Figure 2.** Feature importance for random forest model trained on the (a) balanced training set and (b) SMOTE training set.
PRR age signal and PRR sex signal are signal identified when subgrouping PRR by age and sex, respectively.
PRR, proportional reporting ratio; MGPS, multi-item gamma Poisson shrinker; SMOTE, Synthetic Minority Oversampling Technique.

models had a PRCAUC $\geq 0.8$. All models had an ROCAUC $\geq 0.5$. This is represented in the PRC and ROC curves (Figures 4 and 5). The HPT models had higher ROCAUC values in the LR, GBT, and SVM models. The LR model had equivalent or higher performance evaluation metrics compared to the other three models. It had the highest accuracy, precision, and recall metrics.

## Discussion

Our study evaluated the performance of LR, GBT, RF and SVM against a reference set and

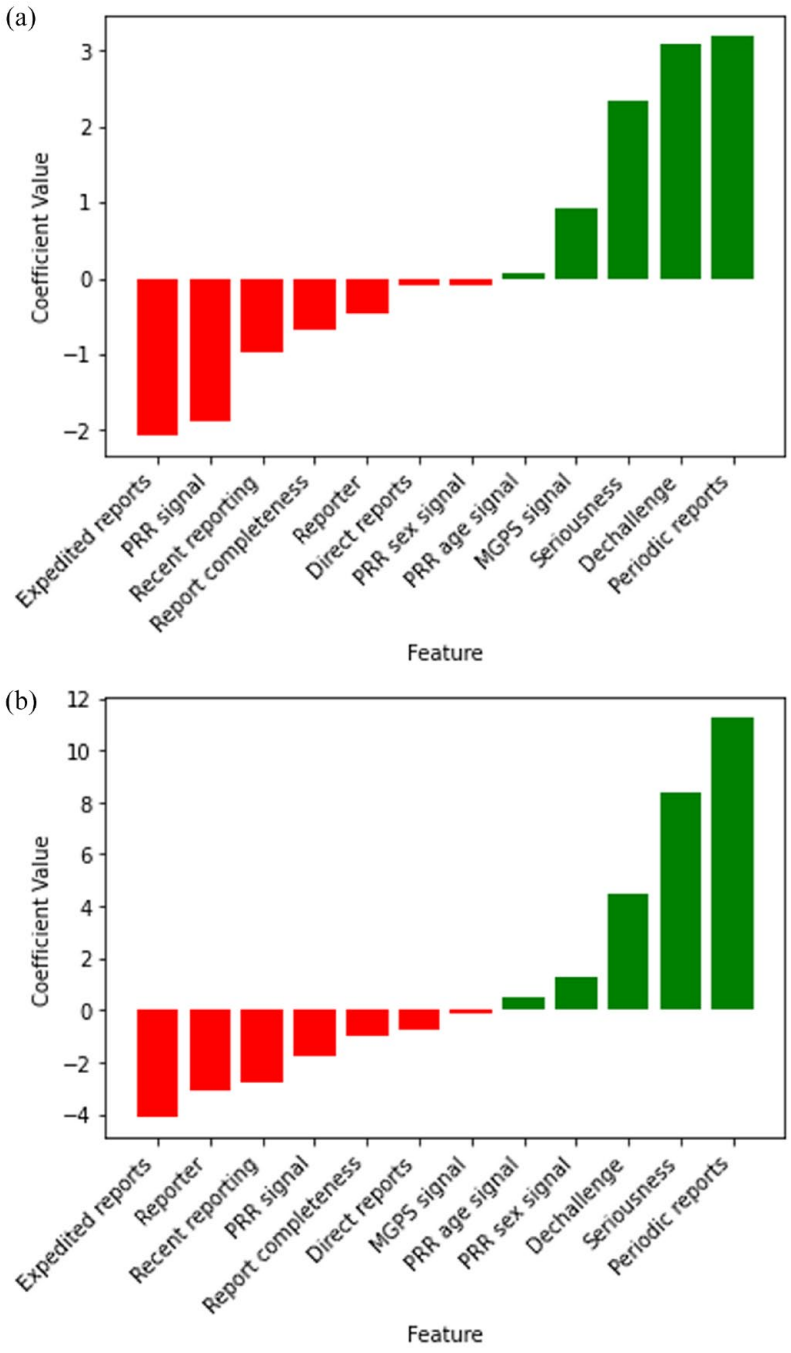


**Figure 3.** Feature importance for support vector machine model trained on the (a) balanced training set and (b) SMOTE training set.
PRR age signal and PRR sex signal are signal identified when subgrouping PRR by age and sex, respectively.
PRR, proportional reporting ratio; MGPS, multi-item gamma Poisson shrinker; SMOTE, Synthetic Minority Oversampling Technique.

included 12 features in the models (Table 1). All models performed similarly (Table 3). In general, models trained on the balanced training set had higher evaluation metric values than models trained on the SMOTE training set. Possible reasons for this include that SMOTE can only generate new synthetic minority class observations within the space of the existing minority class observations, it will not improve the representation of the minority class outside the

**Table 3.** Performance evaluation metrics for models against test data.

| Model | Data set | Metric | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Accuracy** | **F1 score** | **Precision** | **Recall** | **ROCAUC** | **PRCAUC** |
| Logistic regression | Balanced Crude | **0.710** | 0.824 | 0.724 | 0.955 | 0.591 | 0.811 |
| | Balanced HPT | **0.710** | 0.816 | 0.741 | 0.909 | 0.601 | 0.819 |
| | SMOTE Crude | 0.452 | 0.514 | 0.692 | 0.409 | 0.606 | 0.826 |
| | SMOTE HPT | 0.516 | 0.634 | 0.684 | 0.591 | 0.543 | 0.785 |
| Gradient-boosted trees | Balanced Crude | 0.645 | 0.766 | 0.720 | 0.818 | 0.566 | 0.740 |
| | Balanced HPT | 0.645 | 0.776 | 0.704 | 0.864 | 0.646 | 0.825 |
| | SMOTE Crude | 0.645 | 0.766 | 0.720 | 0.818 | **0.657** | 0.803 |
| | SMOTE HPT | 0.613 | 0.750 | 0.692 | 0.818 | 0.641 | 0.847 |
| Random forest | Balanced Crude | 0.645 | 0.776 | 0.704 | 0.864 | 0.641 | 0.839 |
| | Balanced HPT | **0.710** | **0.830** | 0.710 | **1** | 0.596 | 0.808 |
| | SMOTE Crude | 0.645 | 0.766 | 0.720 | 0.818 | 0.624 | 0.832 |
| | SMOTE HPT | 0.645 | 0.766 | 0.720 | 0.818 | 0.636 | 0.848 |
| Support vector machine | Balanced Crude | **0.710** | **0.830** | 0.710 | **1** | 0.586 | 0.789 |
| | Balanced HPT | 0.677 | 0.792 | 0.731 | 0.864 | 0.591 | 0.794 |
| | SMOTE Crude | 0.290 | 0* | 0* | 0* | 0.500 | **0.855** |
| | SMOTE HPT | 0.548 | 0.588 | **0.833** | 0.455 | 0.611 | 0.817 |
| Disproportionality analyses | MGPS | – | – | 0.750 | 0.275 | 0.519 | – |
| | PRR | – | – | 0.673 | 0.321 | 0.458 | – |
| | PRR age subgrouped | – | – | 0.740 | 0.523 | 0.523 | – |
| | PRR sex subgrouped | – | – | 0.721 | 0.450 | 0.499 | – |

Bolded values are the highest value for each performance metric.
Crude, model without hyperparameter tuning or feature selection; HPT, model with hyperparameter tuning and feature selection; MGPS, multi-item gamma Poisson shrinker; PRR, proportional reporting ratio; SMOTE, Synthetic Minority Oversampling Technique.
*This model did not predict any positive control outcomes for the test data.

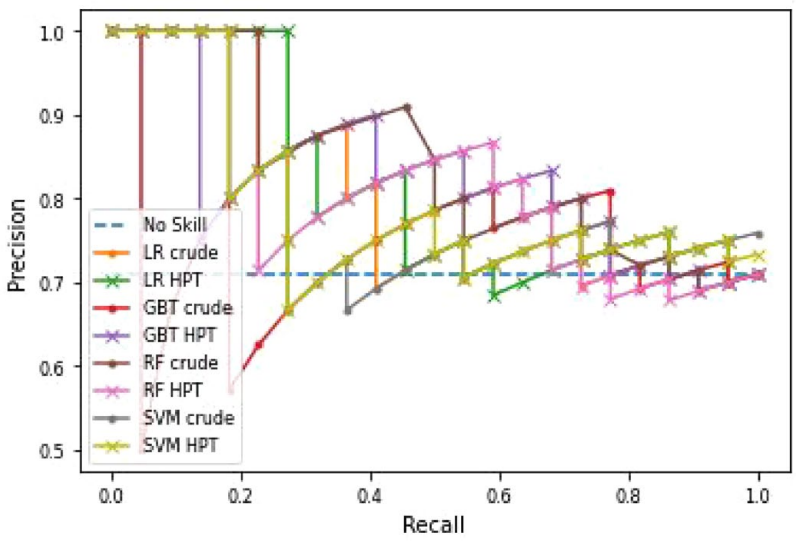


**Figure 4.** Precision-recall curve comparing the performance of the LR, GBTs, RF and SVM algorithms.
Crude, model without hyperparameter tuning or feature selection; HPT, model with hyperparameter tuning and feature selection; LR, logistic regression; GBT, gradient-boosted tree; RF, random forest; SVM, support vector machine.

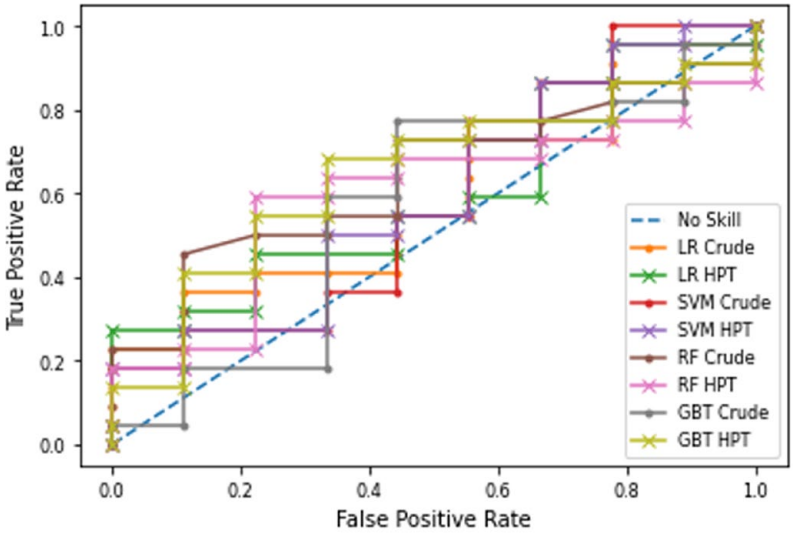


**Figure 5.** Receiver operating characteristic curve comparing the performance of the LR, GBTs, RF and SVM algorithms.
Crude, model without hyperparameter tuning or feature selection; HPT, model with hyperparameter tuning and feature selection; LR, logistic regression; GBT, gradient-boosted tree; RF, random forest; SVM, support vector machine.

boundary of the existing observations, and it could potentially create synthetic observations in locations where majority class observations are located.[9,35,36] In this study, the reference set had 109 positive controls (72.2%) and 42 negative controls (27.8%), and this classifies as low imbalanced data.[35] Depending on the location of the new synthetic negative control observations, the models may have a more difficult time correctly distinguishing between the classes, leading to the decline in the metrics seen in the results.

Our models go beyond standard disproportionality analysis for signal detection by incorporating

the amount and type of information in individual case reports in addition to disproportionality analysis. Table 3 includes the performance evaluation metrics for the disproportionality analyses. The models have higher recall and ROCAUC values; and comparable precision.

A precision-recall curve plot shows the relationship between precision and recall, of which precision is particularly important because it measures the fraction of correct predictions among the positive predictions. They provide a visual summary of the susceptibility of models to imbalanced datasets.[32] The high PRCAUC values indicate the models perform well at detecting true positives. In addition, the ROCAUC values are all greater than 0.5, which indicates the performance is better than chance. It has been shown that there is a connection between ROCAUC and PRCAUC since both include the recall measure and that if a model has a strong precision-recall curve, it will also have a strong ROC curve.[37]

An error analysis using decision trees was conducted to identify how to improve classification (Supplemental Materials 3 and 4).[38,39] In general, all models were affected by two features, recent reporting and ADE report type. The models poorly classified reports submitted greater than 18 months from the end of the study period and all ADE report types. The ADE report type feature quantified the number of expedited, periodic and direct reports for a DEC. An individual DEC could have reports in each category, which could complicate classification. Future research could examine using more recent reports only and removing the ADE report type feature.

This study builds off prior research of using LR and machine learning algorithms for signal detection.[2–6] ADE report type was structured as the number of expedited, periodic or direct reports and subgrouped PRR analyses were conducted for age and sex. In total, 12 features were used in model development. Prior studies investigated the use of age, sex, report year and ADE time-to-onset as predictors.[2–5] Our study included recent reporting as a feature, but age, sex and ADE time-to-onset were accounted for in the report completeness feature. Our LR model did not find recent reporting as a significant feature. However, it was an important feature in GBT, RF and SVM models. Caster *et al.*[5] suggested dechallenge should be investigated further as it was included in two of their cross-validation models. In this study, the dechallenge feature represented if there was a positive dechallenge. Dechallenge had the largest coefficient in the LR model, the second highest importance in the RF and SVM models, and fourth highest importance in the GBT model. Scholl *et al.*[6] found percentage of reports from healthcare professionals to be a strong predictor of the presence of a unique DEC association in the Summary of Product Characteristics in their LR model. Our reporter feature represented the number of reports from a healthcare worker. It was not a significant feature in our LR model, but it was one of the most important features in the GBT, RF and SVM models.

### *Limitations*

Our analysis has limitations. One, ADE reports in FAERS are likely underreported and subject to reporting biases such as dilution bias, indication bias, co-prescription bias and competition bias.[29,40,41] It is important to understand the variety of biases involved in ADE reporting in order to properly interpret the data. Two, the ADE reports in FAERS are subject to issues of data quality and consistency due to a lack of required, standardized fields; missing data, and potential duplicate reports. Three, ADE reports submitted to FDA do not undergo extensive validation or verification, and therefore, a causal relationship cannot be established between a product and the reactions listed in a report. Four, the lack of a gold standard for evaluating signal detection algorithm performance is an issue. We developed a reference set based on ADEs that describe dysglycaemia, hepatic decompensation and hepatic failure; and angioedema. It is imbalanced with more positive controls than negative controls, which can influence the calculations of recall, precision and PRCAUC (Table 3). Lastly, results from this study are not generalizable to other spontaneous report databases, different subgrouping variables or use of an alternative reference set. It is possible that changing one or multiple of these aspects may produce different results. However, our results do provide examples of models and features, in particular dechallenge, that can be considered in future research. Future research examining different machine learning models and data sets would be helpful to expand the generalizability of our findings.

## Conclusion

This study compared the performance of LR, GBT, RF and SVM for signal detection utilizing data from FAERS. LR, RF and the GBT hyperparameter tuned models had a PRCAUC $\geq 0.8$, and all models had ROCAUC values $>0.5$. The LR models had higher accuracy, precision and recall. Incorporating additional information from case reports and the disproportionality analysis results into the models resulted in higher performance evaluation metrics than disproportionality analysis alone. The models can be replicated or modified for use by pharmacovigilance programs.

## Declarations

### *Ethics approval and consent to participate*

Not applicable since this is a secondary data analysis of publicly available de-identified data and individual patients were not involved.

### *Consent for publication*

Not applicable.

### *Author contributions*

**Daniel G. Dauner:** Conceptualization; Data curation; Formal analysis; Methodology; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.

**Eleazar Leal:** Methodology; Resources; Supervision; Writing – review & editing.

**Terrence J. Adam:** Conceptualization; Methodology; Supervision; Writing – review & editing.

**Rui Zhang:** Methodology; Supervision; Writing – review & editing.

**Joel F. Farley:** Conceptualization; Methodology; Project administration; Supervision; Writing – review & editing.

### *Acknowledgements*

We thank Viviene Heitlage, PharmD, BCPS for conducting the second independent review of the prescribing information for the negative controls. This research was presented as a poster at the International Society for Pharmacoepidemiology Mid-Year Meeting in Reykjavik, Iceland from April 23-25, 2023.

### *Funding*

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Financial support was provided by the Hadsall-Uden Award for Pharmacy Advancement from the College of Pharmacy, University of Minnesota.

### *Competing interests*

DGD, EL, TJA and RZ have no conflicts of interest to disclose. JFF reports receiving personal fees from Takeda for expert witness testimony and grant support from Astra Zeneca to the University of Minnesota for an unrelated research project.

### *Availability of data and materials*

The FAERS quarterly data extract files are publicly available (https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html).

## ORCID iD

Daniel G. Dauner https://orcid.org/0000-0002-1198-0468

## Supplemental material

Supplemental material for this article is available online.

## References

1. Hauben M and Bate A. Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov Today* 2009; 14: 343–357.
2. Harpaz R, Dumouchel W, Lependu P, *et al.* Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clin Pharmacol Ther* 2013; 93: 539–546.
3. Dumouchel W, Fram D, Yang X, *et al.* Antipsychotics, glycemic disorders, and life-threatening diabetic events: a Bayesian data-mining analysis of the FDA adverse event reporting system (1968–2004). *Ann Clin Psychiatry* 2008; 20: 21–31.
4. Van Holle L and Bauchau V. Use of logistic regression to combine two causality criteria for signal detection in vaccine spontaneous report data. *Drug Saf* 2014; 37: 1047–1057.
5. Caster O, Juhlin K, Watson S, *et al.* Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank: retrospective evaluation against emerging safety signals. *Drug Saf* 2014; 37: 617–628.

6. Scholl JHG, van Hunsel FPAM, Hak E, *et al*. A prediction model-based algorithm for computer-assisted database screening of adverse drug reactions in the Netherlands. *Pharmacoepidemiol Drug Saf* 2018; 27: 199–205.

7. Ranganathan P, Pramesh CS and Aggarwal R. Common pitfalls in statistical analysis: logistic regression. *Perspect Clin Res* 2017; 8: 148–151.

8. Theobald O. *Machine learning for absolute beginners*. 2nd ed. Monee, IL: Scatterplot Press; 2017.

9. Tan PN, Steinbach M, Karpatne A, *et al*. Introduction to data mining. 2nd ed. New York, NY: Pearson Education, 2019.

10. Liu Y, Chen PHC, Krause J, *et al*. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019; 322: 1806–1816.

11. Wang CS, Lin PJ, Cheng CL, *et al*. Detecting potential adverse drug reactions using a deep neural network model. *J Med Internet Res* 2019; 21: e11016.

12. Ietswaart R, Arat S, Chen AX, *et al*. Machine learning guided association of adverse drug reactions with *in vitro* target-based pharmacology. *EBioMedicine* 2020; 57: 102837–102837.

13. Chen AW. Predicting adverse drug reaction outcomes with machine learning. *Int J Community Med Public Health* 2018; 5: 901–904.

14. Pham M, Cheng F and Ramachandran K. A comparison study of algorithms to detect drug-adverse event associations: frequentist, Bayesian, and machine-learning approaches. *Drug Saf* 2019; 42: 743–750.

15. U.S. Food and Drug Administration. *FDA adverse event reporting system (FAERS) quarterly data extract ASC NTS File*. 2016.

16. Sonawane KB, Cheng N and Hansen RA. Serious adverse drug events reported to the FDA: analysis of the FDA adverse event reporting system 2006–2014 database. *J Manag Care Spec Pharm* 2018; 24: 682–690.

17. Dauner DG and Farley JF. Comparing the use of individual and composite terms to evaluate adverse drug event disproportionality: a focus on glucagon-like peptide-1 receptor agonists and diabetic retinopathy. *Expert Opin Drug Saf* 2021; 20: 475–480.

18. Bergvall T, Niklas Norén G and Lindquist M. VigiGrade: a tool to identify well-documented individual case reports and highlight systematic data quality issues. *Drug Saf* 2014; 37: 65–77.

19. Kumar R, Kumar P, Kalaiselvan V, *et al*. Best practices for improving the quality of individual case safety reports in pharmacovigilance. *Ther Innov Regul Sci* 2016; 50: 464–471.

20. Dauner DG, Zhang R, Adam TJ, *et al*. Performance of subgrouped proportional reporting ratios in the US Food and Drug Administration (FDA) adverse event reporting system. *Expert Opin Drug Saf* 2023; 22: 589–597.

21. U.S. Food and Drug Administration. January – March 2019 | Potential signals of serious risks/new safety information identified from the fda adverse event reporting system (FAERS). FDA, https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/january-march-2019-potential-signals-serious-risksnew-safety-information-identified-fda-adverse (2020, accessed 8 February 2021).

22. U.S. Food and Drug Administration. April – June 2019 | Potential Signals of serious risks/new safety information identified by the FDA adverse event reporting system (FAERS). FDA, https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/april-june-2019-potential-signals-serious-risksnew-safety-information-identified-fda-adverse-event (2019, accessed 8 February 2021).

23. U.S. Food and Drug Administration. FDA warns about rare occurrence of serious liver injury with use of hepatitis C medicines Mavyret, Zepatier, and Vosevi in some patients with advanced liver disease, https://www.fda.gov/drugs/drug-safety-and-availability/fda-warns-about-rare-occurrence-serious-liver-injury-use-hepatitis-c-medicines-mavyret-zepatier-and (2019, accessed 8 February 2021).

24. Ryan PB, Schuemie MJ, Welebob E, *et al*. Defining a reference set to support methodological research in drug safety. *Drug Saf* 2013; 36(Suppl. 1): S33–S47.

25. Harpaz R, DuMouchel W, LePendu P, *et al*. Empirical Bayes model to combine signals of adverse drug reactions. In: *KDD' 13: The 19th ACM SIGKDD international conference on knowledge discovery and data mining*, Chicago, IL, USA, 2013, pp. 1339–1347.

26. Evans SJW, Waller PC and Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001; 10: 483–486.

27. Szarfman A, Machado SG and O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected

combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf* 2002; 25: 381–392.

28. Holle LV and Bauchau V. Signal detection on spontaneous reports of adverse events following immunisation: a comparison of the performance of a disproportionality-based algorithm and a time-to-onset-based algorithm. *Pharmacoepidemiol Drug Saf* 2014; 23: 178–185.
29. Noguchi Y, Tachi T and Teramachi H. Detection algorithms and attentive points of safety signal using spontaneous reporting systems as a clinical data source. *Brief Bioinform* 2021; 22: bbab347.
30. Aggarwal CC. *Data mining*. Switzerland: Springer International Publishing, 2015.
31. scikit-learn. sklearn.ensemble.RandomForest Classifier, https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForest Classifier.html. (accessed 29 June 2022).
32. Saito T and Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; 10: e0118432.
33. Saito T and Rehmsmeier M. Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics* 2017; 33: 145–147.
34. von Elm E, Altman DG, Egger M, *et al*. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med* 2007; 4: e296.
35. Mahani A and Ali ARB. *Classification problem in imbalanced datasets*. IntechOpen, 2019.
36. Brandt J and Lanzen E. A comparative review of SMOTE and ADASYN in imbalanced data classification, https://www.diva-portal.org/smash/get/diva2:1519153/FULLTEXT01.pdf (2020, accessed 7 June 2022).
37. Davis J and Goadrich M. The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd international conference on machine learning – ICML '06*. New York, New York: ACM Press, 2006, pp. 233–240.
38. Dataiku. mealy: Model Error Analysis python package, https://github.com/dataiku/mealy (2022, accessed 2 December 2022).
39. Dataiku. Welcome to mealy's documentation! – mealy 0.2.4-git documentation, https://dataiku-research.github.io/mealy/index.html (2021, accessed 2 December 2022).
40. Alatawi YM and Hansen RA. Empirical estimation of under-reporting in the U.S. Food and Drug Administration adverse event reporting system (FAERS). *Expert Opin Drug Saf* 2017; 16: 761–767.
41. Raschi E, Poluzzi E, Salvo F, *et al*. Pharmacovigilance of sodium-glucose co-transporter-2 inhibitors: what a clinician should know on disproportionality analysis of spontaneous reporting systems. *Nutr Metab Cardiovasc Dis* 2018; 28: 533–542.