Drug Safety
https://doi.org/10.1007/s40264-026-01696-0

ORIGINAL RESEARCH ARTICLE

# Optimising Pharmacovigilance Efficiency with MLIT (Machine Learning for Intelligent Triage): A Tool for Statistical Safety Alerts

**Luciano Ciccarelli[1] · Olivia Mahaux[2] · Christie Roshan[3] · Ami Fofana[1] · Anna Kawka[4] · Emilia Occhipinti[1,7] · Mariapia Possidente[1] · Silvia Cenci[1] · Jeffery L. Painter[5] · Andrew Bate[3,6]**

Received: 13 April 2026 / Accepted: 21 June 2026


## Abstract

**Background and Aim** Pharmacovigilance is essential to ensuring patient safety by enabling timely identification of adverse reactions in increasingly complex and voluminous data. Routine quantitative signal detection methods generate statistical alerts for product-event pairs based on predefined criteria; however, most alerts do not warrant further investigation, creating inefficiencies and significant time demands for pharmacovigilance teams. Manual triage of these alerts is often resource-intensive, prone to variability, and challenging to audit, highlighting the need for more reliable, transparent and efficient triage strategies. This study aimed to design, develop and prospectively evaluate an explainable Machine Learning for Intelligent Triage (MLIT) tool to assist pharmacovigilance teams in reviewing statistical alerts for vaccine and drug portfolios. The objective was to enhance signal detection performance without increasing the risk of missing signals, improving operational efficiency and maintaining decision traceability and regulatory compliance.

**Methods** Alert and individual case safety report data were retrieved from the company's safety and signal management databases. Feature selection was guided by prior experience with a published case completeness tool, called Clinical Utility Score for Prioritisation (CUSP), and expert input. Of several ML methods explored, eXtreme Gradient Boosting (XGBoost) emerged as the optimal algorithm, with models trained and tested using a 75/25 split dataset. Iterative model refinement was conducted using Shapley Additive Explanations analyses to ensure explainability and alignment with safety reviewers' decision-making processes. Refined models underwent prospective validation in two four-month prospective validation studies, covering over 20 products across vaccine and drug portfolios. The prospective validations assessed concordance between model predictions and reviewers' decision under real-world conditions, as well as estimated time savings.

**Results** The vaccine model demonstrated robust predictive performance, achieving a weighted-average *F*1 score of 0.81 and an accuracy of 0.79. In the prospective validation phase, 92% of vaccine alerts were closed in alignment with the model's top-ranked prediction, while 98% were closed within the top 3 predictions. The MLIT tool also identified inconsistencies and human errors in manual triage, highlighting its potential role as a quality-control mechanism. Safety reviewers reported a 24% reduction in time spent on triage activities, and explainability analyses confirmed that the model's decision-making was conceptually aligned with safety reviewers' logic. Comparable results were observed for the drug portfolio.

**Conclusion** This study highlights the potential of ML-based tools to improve pharmacovigilance by enhancing signal detection performance, reducing the likelihood of missed signals, while increasing operational efficiency, and strengthening reproducibility and transparency. While MLIT demonstrated high concordance with expert decisions and provided meaningful time savings, human oversight remains essential, especially for low-confidence predictions. Ongoing refinement and user engagement will be critical for broader implementation and further automation, marking a significant step forward in ensuring safer and more efficient drug safety surveillance.

## 1 Introduction

Pharmacovigilance (PV) is central to patient safety, ensuring timely detection of unexpected adverse reactions across a growing and complex data landscape. As safety databases expand and potential drug/vaccine-adverse event (AE) pairs increase, systems must improve the speed and efficiency of

Luciano Ciccarelli and Olivia Mahaux have contributed equally.

Extended author information available on the last page of the article

Published online: 10 August 2026

**Key Points**

We developed a Machine Learning for Intelligent Triage (MLIT) tool that helped improve efficiency, correctly guiding most safety alert decisions and reducing the time spent by safety reviewers on routine triage work.

The tool supported consistency and quality, identifying some human errors and applying decision logic that aligned with expert judgement.

Prospective validation of the MLIT tool shows its potential for broader integration into pharmacovigilance activities, as long as experts remain involved in reviewing difficult or uncertain cases.

identifying emerging signals. Routine quantitative signal detection methods were introduced in the early 2000s [1], but industry-standard statistical algorithms generate large numbers of alerts that rarely become validated signals, creating a substantial false-positive burden [2–4]. Importantly, statistical alerts indicate trends or potential associations rather than definitive evidence of causality and require thorough clinical review by PV teams [5, 6].

Within our company's signal management platform, quantitative and qualitative alerts support drug safety monitoring. Quantitative statistical alerts, generated from spontaneous individual case safety reports (ICSR) data (excluding clinical trial and solicited reports), flag product-event pairs that meet predefined rule-based criteria. In 2023, the platform generated ~ 27,000 alerts across drug and vaccine portfolios; 99.9% were closed with documented rationale. This high close-out rate reflects system sensitivity rather than missed signals, but highlights inefficiencies in current processes and the need for more effective triage. Despite these inefficiencies, quantitative and qualitative alerting remains essential for the early detection of emerging safety signals [7, 8].

Signal management within the company involves three steps: alert triage, signal validation, and signal assessment. Alert triage, the focus of this paper, is the initial review of statistical outputs to identify potential signals. Safety reviewers apply medical and product expertise to aggregated case-series data to decide whether an alert requires deeper investigation. Manual triage remains time-consuming and varies across safety reviewers. Although triage outcomes and decision rationales are recorded for each alert, the underlying clinical reasoning is often only partially captured in free-text comments and is not documented in a structured nor standardised manner. This limits auditability and traceability and poses challenges for transparency and regulatory compliance. A structured, consistent, and explainable triage framework is needed to support reproducible decision-making and withstand regulatory scrutiny.

Advances in artificial intelligence (AI) and machine learning (ML) suggest significant potential to enhance processes relying on large volumes of complex data, such as PV. In PV, AI/ML applications span case processing, signal detection, predictive modelling, and integration of diverse data streams. Growing evidence shows that they can improve efficiency, accuracy, and scalability in safety signal management [9, 10]. However, adoption in PV faces challenges including data heterogeneity/incompleteness, quality issues, and regulatory fragmentation [11]. The highly regulated nature of PV requires harmonised global frameworks and consistent adherence to AI/ML best practices. Transparency, accountability, and explainability in AI/ML are critical for regulatory inspections and audits. Automation and ML should augment, not replace, expert judgement, ensuring that trust, oversight, and safety performance remain central [12, 13].

Traditional disproportionality methods implementation often favours sensitivity over specificity, resulting in high false-positive rates and substantial triage workloads [14]. This limitation persists even with advanced approaches, as signal detection intentionally favours sensitivity to avoid missing emerging safety concerns [15]. Recent studies examined AI/ML to strengthen signal detection and triage [16–19]. While some aim to reduce upstream false-positive alerts [20], our work focuses on reducing reviewer burden during downstream triage. Imran et al. (2022) developed a supervised ML classifier (eXtreme Gradient Boosting [XGBoost]) predicting validation decisions for disproportionality signals, achieving macro-$F1$ 0.53 and weighted-$F1$ 0.83 scores with Shapley Additive Explanations (SHAP)-based interpretability [21]. Although based on similar principles, their model predicts downstream validation outcomes, whereas our aim is upstream triage support, which relies on different evidence levels, case-series attributes, and decision criteria. Differences in architecture, taxonomies, and proprietary data further limit direct adoption. Unlike validation, triage is a repetitive, pattern-recognition task under extreme class imbalance, with most alerts closed after review. Supervised ML can learn historical reviewer patterns and case-series features to prioritise or recommend closure of low-risk alerts, enabling experts to focus on alerts requiring clinical judgement [10, 22].

In ML, features refer to data elements that contribute to the predictive nature of a model. Our feature set was tailored to our triage workflow, with the Clinical Utility Score for Prioritisation (CUSP)—an automated method that evaluates a predefined set of ICSR attributes considered to contain clinically relevant information and uses them to rank ICSRs

by the completeness of data normally judged to be of clinical relevance [23]—serving as one source of candidate case-level attributes. Attributes judged most likely to influence reviewer priorities and triage decisions were included as features. This study aimed to develop and prospectively evaluate an explainable ML tool that predicts triage decisions for vaccine and drug statistical alerts. Alerts were framed as a multiclass classification problem using company-specific triage labels. Routine triage decisions were combined with aggregated trends and key clinical features from ICSRs as model inputs. Guided by explainability and expert feedback, we trained and refined models to create an interpretable ML for Intelligent Triage (MLIT) tool aligned with our workflows and data. The methodological framework is broadly transferable, since core elements such as feature engineering, explainability workflows, and human-in-the-loop safeguards are generalisable. However, triage taxonomies, asset features, internal metadata, and workflows are organisation-specific and would require adaptations.

## 2 Methods

We developed separate models for vaccines and drugs to account for differences in triage, confounders, risk factors, and decision categories. Here, we focused on the vaccine models, which used largely the same methodological framework as the drug models; drug-specific methods and results are presented in the Supplementary Methods and Results.

### 2.1 Data Sources

#### 2.1.1 Alert Data

Alerts were retrieved from the company's signal management system. Alerts for product-event pairs are routinely generated and classified as: (i) disproportionality analysis (DPA), when the Multi-item Gamma Poisson Shrinker lower bound ($5^{th}$ percentile) of the 90% credibility interval (CI) of the Empirical Bayes Geometric Mean (EBGM) (EB05) is $\geq$ 2 [24]; (ii) time to onset (TTO; vaccine model only), when two-sample Kolmogorov-Smirnov tests [25] yield $p$-values <0.01 for TTO distributions within a 30-day window; (iii) designated medical event (DME), the list of which is defined and maintained in-house, when $\geq$ 4 reports of an unlisted DME (or $\geq$ 1 for newly launched products) are reported; or (iv) aberration, when $\geq$ 4 cases exceed the forecast upper bound from a negative binomial time-series regression fitted to the preceding six months [26].

Statistical alerts are generated periodically using cumulative data through each alert-period end. Triage is a repeated process: each review considers new information since the previous triage, but records decisions are based on the cumulative data. For each alert, the reviewer makes one of two primary triage decisions: either to close the alert or to initiate further investigation. When an alert is closed, the reviewer should select a single one of 13 predefined decision-rationale categories (Table 1). An "Other" category is available and should be used in situations where none of the specific categories are judged applicable, or when multiple categories are equally applicable and no single rationale predominates; in such cases, the reviewer must use the free-text comment field to clarify the main rationale(s) for closure. For brevity, we refer to these decision rationale categories as "decisions" in the remainder of the manuscript and in the ML modelling. The retrieved dataset therefore represented one row per product-event-alert-period-end-alert-type combination and included the reviewer's routine triage outcome, selected decision rationale (Table 1), and free-text comment; counts of spontaneous and total reports; new spontaneous reports since 1, 3, 6, 9 and 12 months; counts of serious and fatal reports (if any); EBGM with its 90% CI (EB05, EB95); and flags for DME, label-listed, serious, and medication error events.

The initial vaccine model used cumulative alerts through December 2022 for 17 marketed vaccines. The refined model was retrained on an expanded dataset covering all marketed vaccines ($n = 36$) and including cumulative alerts through July 2024 to capture recent trends.

#### 2.1.2 Individual Case Safety Reports Data

ICSR data were retrieved from the company safety database and included attributes selected for their clinical relevance; several were identified using CUSP. Based on CUSP and expert judgement, we included ICSR attributes such as event duration and outcome, product dosage and indication, action taken with the product, presence of dechallenge or rechallenge, and TTO. Patient information (age, sex, medical history) was also included. These variables (Table 2) were organised by their relation to the report, patient, event, or product.

The initial vaccine model was trained and tested on spontaneous ICSRs received through 16 December 2022. The refined model's train/test dataset included all report types (not only spontaneous) received through 16 September 2024.

#### 2.1.3 Product-Level Metadata

Product-specific features for the refined models were compiled outside the safety database using PV expert knowledge and internal product records. These differed by product type and included vaccine target population, target disease, and novelty status.

**Table 1** Predefined decisions to document closure or investigation of statistical alerts

| Decision | Outcome |
|---|---|
| Event being monitored, no change | Close |
| Event describing suspected or confirmed vaccination failures (vaccine model) | Close |
| Event is likely explained by other factors | Close |
| Event term is uninformative or non-specific | Close |
| Implausible TTO distribution (vaccine model) | Close |
| Implausible relationship based on known properties/mode of action | Close |
| Information does not change previous assessment | Close |
| Insufficient information available | Close |
| Listed event including synonym or symptom of | Close |
| Medication error with no specific pattern of concern | Close |
| Other | Close |
| Reporting bias related to coding/media coverage/regional reporting practice | Close |
| Event being investigated, assessed | Close |
| Merits further investigation | Investigate |

*TTO* time to onset

## 2.2 Data Pre-processing

### 2.2.1 Alert Data

Because alerts were generated repeatedly, the dataset was restricted to the most recent alert per product-event pair, so each pair contributed a single observation. However, to increase the number of samples for the decision "Merits further investigation", we kept the most recent alert with that decision.

During pre-processing, fields were added: counts of how often each product-event pair triggered each alert type (DPA/TTO/DME/aberration); the immediately preceding previous decision and period-end for the retained alert; and a flag for alerts migrated from legacy signal management systems that may have been triggered or triaged under earlier processes. Historical alerts migrated from legacy systems with missing statistical scores and event characteristics were excluded to ensure data completeness.

For the refined models, PV experts reviewed alerts with multiple closing decisions in the train/test dataset and, where possible, selected the single most relevant decision. Alerts remaining with multiple decisions were excluded to avoid many low-frequency combinations that would impair model training. For product-event pairs with at least one "Event being monitored, no change" decision, we selected the most recent alert carrying that decision instead of the most recent alert overall, to increase the number of observations.

The "Other" decision category is non-specific and varies in usage. This category was therefore reviewed by experts and, where appropriate, alerts were reclassified into more specific decision categories. Alerts that had been reclassified were then included under the corresponding category for training/testing. For example, "Other" was sometimes selected to indicate that the cases in the alert remained consistent with the previous review and that the decision was unchanged, although the same decision as previously should instead have been selected. In other instances, "Other" was

**Table 2** ICSR attributes retrieved from the company safety database

| Report information | Patient information | Event information | Product information |
|---|---|---|---|
| Report type | Age | Reported event(s) | Suspect products |
| Healthcare professional report | Sex | DME | Concomitant products |
| Country | Weight | Event onset date | Product start date |
| Verbatim narrative | Medical history | Event duration | Product indication |
| Narrative length | Comorbidities | Event outcome | Dose/dosing regimen |
| Maximum significant follow-up date | Historical drug use | Time to onset | Action taken |
| Report seriousness | Test and procedure results | Rechallenge | Lot number |
| Fatal outcome (subpopulation) | | Listedness | Dechallenge |

*DME* designated medical event, *ICSR* individual case safety report

chosen even though the free-text comment aligned with existing decisions such as "Insufficient information available" or "Event term is uninformative or non-specific". For this study, only alerts with a confirmed "Other" decision after this review were retained as "Other" for training/testing.

#### 2.2.2 Individual Case Safety Reports Data

Underrepresented values for event outcome and action taken were grouped with related, more frequent categories.

Several attributes were converted to binary flags (e.g., dose_flag = 1 if dose, daily dose, or total regimen dose > 0 with units provided, or if dose number > 0). Because product indications, lot numbers, and reporter countries had many distinct non-null values, we counted distinct values per product-event pair and counted missing values for indication and lot number.

For the refined models, TTO was supplemented with free text timing expressions (e.g., immediately; less than an hour) that could be mapped to the same TTO categories. For vaccines, missing or unknown actions taken were set to "Not applicable". Dechallenge and rechallenge were retained when explicitly reported as positive or negative and set to "Not applicable" otherwise.

Further pre-processing details and variable definitions are provided in Supplementary Table 1.

### 2.3 Feature Engineering

TTO and event duration were categorised as: 0–< 24 h (hours), 24 h–< 7 days (days), 7–< 30 days (weeks), 30–< 365 days (months), ≥ 365 days (years); negative or missing values were coded as "Unknown".

Alerts and ICSRs were linked by product and Medical Dictionary for Regulatory Activities (MedDRA) Preferred Term (PT) and by matching report dates within the alert period. Reports with significant follow-up after the alert-period end were excluded to ensure alignment with the alert timeline. Where multiple ICSRs matched a single alert, they were aggregated at the product-event alert level by counting cases for each category value and expressing these counts as proportions of the total cases underlying the alert [21].

Categorical variables were "one-hot encoded" (i.e., converted into multiple yes/no (1/0) variables, one per category, indicating whether that category applied) to make the data compatible with ML models [27]. A feature correlation matrix was generated, and highly correlated attributes were removed after consultation with PV experts. For the initial model, the DME and event listed in product label flags from the alert data source were removed because they were highly correlated with the corresponding flags from the ICSR data source and were missing for some historical alerts. The number of spontaneous reports (Spont_N), which was strongly correlated with the number of fatal reports (Fatal_N), and the EBGM for spontaneous reports (Spont_EBGM), which was highly correlated with the lower 90% confidence bound of the EBGM for spontaneous reports (Spont_EB05) were also removed. Table 3 summarises the features retained for the initial model after feature selection.

For the refined models, we applied targeted feature engineering guided by explainability reviews and PV expert input. Redundant binary features were consolidated to retain a single representative value. Highly correlated features were removed. Table 4 lists the features used in the refined models, comprising a subset of those retained for the initial model plus several newly engineered features identified during expert review.

### 2.4 Machine Learning Framework

We evaluated several multiclass algorithms (Decision Tree [28], K-Nearest Neighbours [29], Linear Support Vector Machine [30], Logistic Regression, Random Forest [31], and XGBoost [32, 33]), prioritising interpretability alongside performance given the regulatory context. XGBoost is a supervised tree-based ensemble method that implements an optimised gradient-boosting framework and is well suited to fast, scalable learning from structured (tabular) data.

Models were developed in Python (v3.8.10) using Scikit-learn (v0.24.1) [34] and executed on JupyterHub and Databricks. Performance was assessed on a held-out test set using a combination of class-level precision, recall, and $F1$-score, with weighted-average $F1$ as the primary selection metric to account for class imbalance. Both Random Forest and XGBoost achieved acceptable accuracy (> 60%). After hyperparameter tuning and applying class weights, XGBoost performed best on our large, sparse, high-dimensional, imbalanced dataset (data not shown).

Hyperparameters for the XGBoost (v2.1.1.) classifier were optimised via grid search [34] over learning rate, tree depth, number of estimators, and regularisation terms using a 3-fold cross-validation to reduce overfitting, chosen as a compromise between variance reduction and computational cost given the large dataset and model complexity. The weighted $F1$-score was used as the tuning metric to address the imbalanced multiclass setting. We used the "multi:softprob" objective to produce per-class probabilities for each instance, and used these raw, uncalibrated probabilities to assess prediction confidence and output reliability.

Models were trained on 75% of the data and tested on the remaining 25% (test dataset). To maintain representativeness across all alert triage classes and remove any potential influences of data ordering, random stratified sampling was used during data partitioning. This effectively mitigated sampling bias, ensuring an equitable distribution of examples in both

**Table 3** Features included in the initial model

| Feature | Source | Type | Engineering/short description |
|---|---|---|---|
| Spont_EB05 | Alert data | Numeric | MGPS lower 90% CI (EB05) |
| Num_DPA_alerts | Alert data | Count | Cumulative DPA alerts up to period end |
| Num Aberration alerts | Alert data | Count | Cumulative aberration alerts up to period end |
| Num_TTO_alerts | Alert data | Count | Cumulative TTO alerts up to period end (Vx only) |
| Num_Pediat_alerts | Alert data | Count | Cumulative paediatric DPA alerts up to period end (Vx only) |
| Num_Trend_alerts | Alert data | Count | Cumulative trend in EB05 alerts up to period end |
| Prevdecision_flags[a] | Alert data | Binary | One flag per previous decision |
| Historical_flag | Alert data | Binary | Flag for alerts migrated from legacy system |
| Medication_error_flag | Alert data | Binary | Event flagged as Medication error SMQ |
| Serious_flag | Alert data | Binary | Event flagged as serious (safety database auto-seriousness) |
| Outcome_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by event outcome |
| EV_duration_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by event duration bin |
| TTO_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by TTO bin |
| NumDIFindication | ICSR data | Count | Number of distinct product indications in cases |
| NumUNKindication | ICSR data | Count | Number of unknown product indications in cases |
| NumDIFlot | ICSR data | Count | Number of distinct lot number in cases |
| NumUNKlot | ICSR data | Count | Number of unknown lot numbers in cases |
| NumDIFcountry | ICSR data | Count | Number of different countries in cases |
| Dose_percents/counts | ICSR data | Percent/count | %/counts of cases by dose flag |
| Action_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by action taken category |
| Dechall_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by dechallenge category |
| Rechall_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by rechallenge category |
| AgeGroup_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by age group category |
| Sex_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by sex category |
| Fatal_percents/counts | ICSR data | Percent/count | %/counts of cases by fatal flag |
| Seriousness_percents/counts | ICSR data | Percent/count | %/counts of cases by case seriousness flag |
| Narr_len_percents/counts | ICSR data | Percent/count | %/counts of cases by long narrative flag |
| Verbatim_percents/counts | ICSR data | Percent/count | %/counts of cases by verbatim flag |
| HCP_percents/counts | ICSR data | Percent/count | %/counts of cases by HCP flag |
| Lab_res_percents/counts | ICSR data | Percent/count | %/counts of cases by laboratory results flag |
| PatMedHist_percents/counts | ICSR data | Percent/count | %/counts of cases by patient medical history flag |
| PatProdHist_percents/counts | ICSR data | Percent/count | %/counts of cases by patient product history flag |
| Comorbidity_percents/counts | ICSR data | Percent/count | %/counts of cases by comorbidity flag |
| Concomitant_percents/counts | ICSR data | Percent/count | %/counts of cases by concomitant products flag |
| OtherSusProd_percents/counts | ICSR data | Percent/count | %/counts of cases by other suspect products flag |
| DME_flag | ICSR data | Binary | Event flagged as DME (CMQ) |
| Listed_flag | ICSR data | Binary | Event flagged as listed in product label |

*CI* credibility interval, *CMQ* company MedDRA query, *DME* designated medical event, *DPA* disproportionality analysis, *EB05* lower bound ($5^{th}$ percentile) of the 90% CI of the Empirical Bayes Geometric Mean, *HCP* healthcare provider, *ICSR* individual case safety report, *MedDRA* Medical Dictionary for Regulatory Activities, *MGPS* Multi-item Gamma Poisson Shrinker, *SMQ* standard MedDRA query, *TTO* time to onset, *Vx* vaccine model

[a]See Supplementary Table 1 for the full list of PrevDecision classes, feature categories, and exact TTO/duration bin definitions

the training and testing sets. To address class imbalance, class weights were adjusted during training to penalise minority-class misclassification more heavily, encouraging the model to focus on underrepresented classes. Model predictions on the test dataset were compared with routine triage decisions made by safety experts.

## 2.5 Performance Evaluation

Predictive performance was evaluated on the test dataset using classification reports with class-level precision (proportion of predicted positives that are true positives), recall (sensitivity, proportion of true positives correctly identified), and $F1$-score (harmonic mean of precision and recall).

**Table 4** Features included in the refined model

| Feature | Source | Type | Engineering/short description |
|---|---|---|---|
| Spont_EB05 | Alert data | Numeric | MGPS lower 90% CI (EB05) |
| Num_DPA_alerts | Alert data | Count | Cumulative DPA alerts up to period end |
| Num Aberration alerts | Alert data | Count | Cumulative aberration alerts up to period end |
| Num_TTO_alerts | Alert data | Count | Cumulative TTO alerts up to period end (Vx only) |
| Prevdecision_flags[a] | Alert data | Binary | One flag per previous decision |
| *Percent_Prevdecision* | *Alert data* | *Percent* | *% of previous decision by decision category* |
| *New_N_cases* | *Alert data* | *Count* | *Count of new cases since previous decision* |
| Historical_flag | Alert data | Binary | Flag for alerts migrated from legacy system (Vx only) |
| Medication_error_event | Alert data | Binary | Event flagged as Medication error SMQ |
| Serious_flag | Alert data | Binary | Event flagged as serious (safety database auto-seriousness) |
| *DME_flag* | *Alert data* | *Binary* | *Event flagged as DME (CMQ)* |
| *Listed_flag* | *Alert data* | *Binary* | *Event flagged as listed in product label (safety database auto-listedness)* |
| Outcome_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by event outcome |
| EV_duration_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by event duration bin |
| TTO_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by TTO bin |
| NumDIFlot | ICSR data | Count | Number of distinct lot number in cases |
| NumUNKlot | ICSR data | Count | Number of unknown lot numbers in cases (Vx only) |
| NumDIFcountry | ICSR data | Count | Number of different countries in cases |
| Dose_percents/counts | ICSR data | Percent/count | %/counts of cases by dose flag |
| Action_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by action taken category |
| Dechall_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by dechallenge category |
| Rechall_percents/counts[a] | ICSR data | Percent/count | %/counts of cases by rechallenge category |
| Seriousness_percents/counts | ICSR data | Percent/count | %/counts of cases by case seriousness flag (Vx only) |
| Lab_res_percents/counts | ICSR data | Percent/count | %/counts of cases by laboratory results flag |
| PatMedHist_percents/counts | ICSR data | Percent/count | %/counts of cases by patient medical history flag |
| PatProdHist_percents/counts | ICSR data | Percent/count | %/counts of cases by patient product history flag |
| Comorbidity_percents/counts | ICSR data | Percent/count | %/counts of cases by comorbidity flag |
| Concomitant_percents/counts | ICSR data | Percent/count | %/counts of cases by concomitant products flag |
| OtherSusProd_percents/counts | ICSR data | Percent/count | %/counts of cases by other suspect products flag |
| *Pregnancy_subpop_percents/counts* | *ICSR data* | *Percent/count* | *%/counts of cases by pregnancy subpopulation flag* |
| *Pregnancy_event* | *ICSR data* | *Binary* | *Event flagged as pregnancy CMQs (Vx only)* |
| *Reacto_event* | *ICSR data* | *Binary* | *Event flagged as reactogenicity (SOC)* |
| *LOE_event* | *ICSR data* | *Binary* | *Event flagged as lack of efficacy SMQ and CMQs (Vx only)* |
| *YearsInMarket_category*[a] | *Product metadata* | *Binary* | *One flag per years in market category* |
| *Vaccine_category*[a] | *Product metadata* | *Binary* | *One flag per vaccine category* |
| *Vaccine_type*[a] | *Product metadata* | *Binary* | *One flag per vaccine type* |

*CI* credibility interval, *CMQ* company MedDRA query, *DME* designated medical event, *DPA* disproportionality analysis, *EB05* lower bound (5th percentile) of the 90% CI of the Empirical Bayes Geometric Mean, *ICSR* individual case safety report, *MedDRA* Medical Dictionary for Regulatory Activities, *MGPS* Multi-item Gamma Poisson Shrinker, *SMQ* standard MedDRA query, *SOC* MedDRA system organ class, *TTO* time to onset, *Vx* vaccine model

[a]See Supplementary Table 1 for full list of PrevDecision classes, feature categories, and exact TTO/duration bin definitions. Features in *italics* are new features added in the refined models

Aggregated metrics, summarising performance across all classes, included macro-average *F*1-score (equal class weighting) and weighted-average *F*1-score (accounting for class imbalance). This is particularly important for imbalanced datasets, where accuracy alone can be misleading.

Models produce a probability for every class; the highest probability class becomes the model's predicted label (the "top" or 1st-ranked prediction) and is assessed by standard metrics (confusion matrix, precision/recall/*F*1). We also examined the ranked probabilities (2nd/3rd-ranked) to

interpret model uncertainty and alternative plausible classifications and referred to these explicitly when discussing ordered class probabilities.

A confusion matrix [35] visualised prediction performance across decision classes, aiding identification of specific error types and supporting explainability analyses and model refinement. Because triage is a repeated, time-dependent process, we also performed retrospective time-based evaluations: models were trained on alerts up to an earlier cut-off and then applied to alerts from subsequent months that had already been triaged but were unseen by the models. Performance on these time-based test sets was similar to that on the main held-out test set (data not shown), supporting the models' ability to generalise to future alerts without changing the overall conclusions.

### 2.6 Explainability and Refinement

Explainability analyses using SHAP (v0.44.1) [36] were performed to improve the model. SHAP quantifies feature importance and the direction/magnitude of each feature's contribution. We assessed importance at three levels. Model-level global SHAP values identified features that most strongly influence predictions and revealed key patterns and interactions. Class-level global explanations clarified decision logic for each decision category. Sample-level local SHAP values explained individual predictions by showing which features pushed the prediction toward or away from specific classes. Misclassified test samples were examined by comparing SHAP values for the true and predicted classes to identify features driving incorrect predictions and systematic error patterns.

Model refinement proceeded iteratively over a two-year period with PV experts: (1) train model; (2) generate SHAP explanations (global summaries and local plots for misclassifications); (3) expert review; (4) apply refinements (feature additions/removals, recoding, procedural changes); and (5) retrain model. This was repeated until stakeholders were satisfied with performance, interpretability and operational fit.

### 2.7 Prospective Validation

A four-month prospective validation study (November 2024–February 2025) evaluated the refined model's practical utility. During the study, safety reviewers received the model's top 3 predictions and key features supporting each prediction; feature importance details were available on request. Reviewers recorded whether they agreed with one of the top 3 predictions or none and noted if any non-selected prediction was nevertheless an acceptable alternative. Reviewers also estimated time saved during triage when using the model versus the standard manual process.

To assess potential bias from model support, a blinded validation was conducted in the first month for one vaccine. Safety reviewers alternated between using model predictions and performing manual triage: half of the PTs were reviewed by Reviewer A with model support, while Reviewer B triaged them manually; the remaining PTs were reviewed with roles reversed.

## 3 Results

### 3.1 Performance of the Initial Vaccine Model

The initial dataset contained 650,483 ICSRs and 4293 alerts, of which 1074 alerts were used to test the model (Fig. 1a). The initial model demonstrated a weighted-average $F1$-score and an overall accuracy of 0.70, indicating reasonable predictive performance across the dataset. Overall, decisions supported by larger sample sizes tended to have higher $F1$-scores, whereas classes with limited data more often showed lower scores.

Four decision categories accounted for over 80% of the test dataset: "Other", "Listed event including synonym or symptom of", "Event is likely explained by other factors", "Implausible relationship based on known properties/mode of action", with the latter achieving the highest $F1$-score (0.85), supported by a robust sample size of 252. Conversely, categories with limited sample sizes, such as "Implausible TTO distribution", exhibited poor performance, with $F1$-scores as low as 0. "Event being investigated, assessed" was not represented in the test dataset because of its extremely low sample size and therefore does not appear in the performance output. "Other" represented the largest proportion of the dataset (262 samples) but achieved a relatively low $F1$-score of 0.59.

Inspection of the confusion matrix (Fig. 1b) revealed consistent misclassification patterns. When misclassified, a high proportion of decisions was predicted as "Other" category, which on further review was found to be highly heterogeneous and inconsistently used, often overlapping with other decision categories. "Event term is uninformative or non-specific" ($F1 = 0.07$) and "Insufficient information available" ($F1 = 0.47$) were also subject to frequent misclassification, reflecting limited support and overlap with other categories. These findings informed manual review and refinement in subsequent model iterations.

Despite being represented by only three samples in the test dataset, "Merits further investigation" achieved a reasonable $F1$-score of 0.67, reflecting the model's ability to identify this critical class. These decisions are particularly important because they prioritise potential signals among many false-positive alerts and therefore must not be missed. The model correctly predicted two of the three alerts; the

**a**

| | Predicted decision class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| 1 | Event being monitored, no change | 0.50 | 0.50 | 0.50 | 2 |
| 2 | Event describing suspected or confirmed vaccination failures | 0.83 | 0.62 | 0.71 | 8 |
| 3 | Event is likely explained by other factors | 0.63 | 0.57 | 0.59 | 136 |
| 4 | Event term is uninformative or non-specific | 0.11 | 0.05 | 0.07 | 19 |
| 5 | Implausible time to onset distribution | 0.00 | 0.00 | 0.00 | 2 |
| 6 | Implausible relationship based on known properties/mode of action | 0.82 | 0.89 | 0.85 | 252 |
| 7 | Information does not change previous assessment | 0.69 | 0.48 | 0.56 | 23 |
| 8 | Insufficient information available | 0.46 | 0.48 | 0.47 | 77 |
| 9 | Listed event including synonym or symptom of | 0.81 | 0.84 | 0.83 | 249 |
| 10 | Medication error with no specific pattern of concern | 0.77 | 0.87 | 0.82 | 39 |
| 11 | Merits further investigation | 0.67 | 0.67 | 0.67 | 3 |
| 12 | Other | 0.59 | 0.58 | 0.59 | 262 |
| 13 | Reporting bias* | 1.00 | 0.50 | 0.67 | 2 |
| | **Accuracy** | | | **0.70** | **1074** |
| | **Macro average** | **0.61** | **0.54** | **0.56** | **1074** |
| | **Weighted average** | **0.69** | **0.70** | **0.70** | **1074** |

**b**

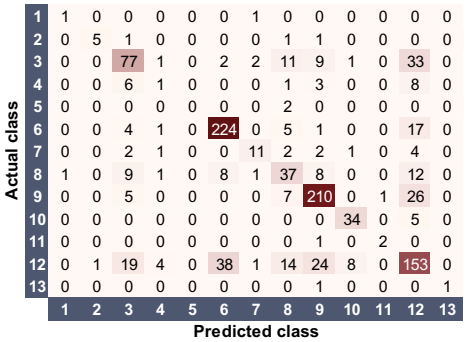


**Fig. 1** EXtreme Gradient Boosting classification report (**a**) and confusion matrix (**b**) for the initial vaccine model. *Related to coding/media coverage/regional reporting practice. Notes: The classification report compares the model's prediction against the actual decision of the safety reviewer on the hold-out test set. For the confusion matrix, the diagonal shows counts of correct predictions for each decision (i.e., the matches between safety reviewer's decisions and the highest probability predictions). Cells off the diagonal show misclassifications (which true decision was confused with which incorrectly predicted decision)

single missed instance had a low top prediction probability, indicating limited model confidence.

## 3.2 Performance of the Refined Vaccine Model

After refinement and fine-tuning, the XGBoost model was retrained and tested on an expanded dataset of 1,395,671 ICSRs and 3850 alert decisions. The refined model achieved a weighted-average $F1$-score of 0.81 and an accuracy of 0.79 (Fig. 2a), up from 0.70 in the initial run. Improvements were concentrated in several high-support classes and triage-critical classes (Supplementary Fig. 1). Performance for "Listed event including synonym or symptom of" improved substantially ($F1 = 0.93$; support = 413), reflecting both increased training data and improvements in feature representation; this class showed very high precision and recall and accounted for a larger share of correctly classified alerts. For "Medication error with no specific pattern of concern", the refined model achieved near perfect performance ($F1 = 0.95$; support = 76), indicating robust identification of medication error patterns after refinement. "Event is likely explained by other factors" ($F1 = 0.77$) and "Insufficient information available" ($F1 = 0.79$) showed improved $F1$-scores with larger support counts, suggesting better calibration of the model for common or non-actionable alert closure decisions. Notably, these four classes contained 82% of total alerts.

"Event describing suspected or confirmed vaccination failure" increased in sample size but showed decreased $F1$-scores. "Merits further investigation" ($F1 = 0.47$; support = 5), exhibited a lower $F1$ despite a few additional samples; however, the model achieved a high recall for this critical category (0.80), correctly identifying four of the five alerts. The remaining alert was incorrectly predicted as "Information does not change previous assessment" with a low probability of 26%, primarily driven by the presence of "Information does not change previous assessment" as previous decision feature, which strongly influenced the model toward repeating this conclusion. This indicates that the PT had already been recognised and previously evaluated by the safety team. Notably, "Merits further investigation" appeared as the model's 3$^{rd}$-ranked prediction (9% probability), and the subsequent review ultimately resulted in closure as a non-validated signal. Additional details on the "Merits further investigation" decisions are provided in Supplementary Table 2.

"Other" remained challenging ($F1 = 0.52$; support = 47), reflecting its inherent heterogeneity despite efforts to improve consistency through manual curation. Notably, "Reporting bias related to coding/media coverage/regional reporting practice" showed a marked reduction in sample size relative to the initial model, indicating less frequent use in recent triage decisions, and no alerts were assigned "Implausible TTO distribution" during the refined model training period, so this category was no longer represented in the dataset. Remaining classes continued to be poorly predicted, largely due to limited training examples.

As observed in the initial model, classes with larger support generally showed higher $F1$-scores after refinement. The expanded training set and iterative feature/label refinements produced notable gains for classes with increased sample counts, while classes with sparse data continued to have limited performance.

a

| | Predicted decision class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| 1 | Event being monitored, no change | 0.56 | 0.60 | 0.58 | 15 |
| 2 | Event describing suspected or confirmed vaccination failures | 0.42 | 0.77 | 0.54 | 13 |
| 3 | Event is likely explained by other factors | 0.86 | 0.70 | 0.77 | 179 |
| 4 | Event term is uninformative or non-specific | 0.55 | 0.62 | 0.58 | 39 |
| 5 | Implausible relationship based on known properties/mode of action | 0.23 | 0.35 | 0.28 | 17 |
| 6 | Information does not change previous assessment | 0.41 | 0.61 | 0.49 | 33 |
| 7 | Insufficient information available | 0.82 | 0.76 | 0.79 | 123 |
| 8 | Listed event including synonym or symptom of | 0.97 | 0.88 | 0.93 | 413 |
| 9 | Medication error with no specific pattern of concern | 0.94 | 0.96 | 0.95 | 76 |
| 10 | Merits further investigation | 0.33 | 0.80 | 0.47 | 5 |
| 11 | Other | 0.43 | 0.68 | 0.52 | 47 |
| 12 | Reporting bias* | 0.67 | 0.67 | 0.67 | 3 |
| | **Accuracy** | | | **0.79** | **963** |
| | **Macro average** | **0.60** | **0.70** | **0.63** | **963** |
| | **Weighted average** | **0.83** | **0.79** | **0.81** | **963** |

b

| Actual class \ Predicted class | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9 | 2 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 10 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 125 | 6 | 10 | 13 | 3 | 2 | 1 | 1 | 18 | 0 |
| 4 | 2 | 0 | 0 | 24 | 1 | 3 | 1 | 1 | 0 | 3 | 4 | 0 |
| 5 | 0 | 0 | 5 | 1 | 6 | 0 | 1 | 1 | 0 | 0 | 3 | 0 |
| 6 | 0 | 0 | 3 | 4 | 1 | 20 | 1 | 2 | 2 | 0 | 0 | 0 |
| 7 | 0 | 4 | 3 | 4 | 5 | 4 | 94 | 3 | 0 | 0 | 6 | 0 |
| 8 | 2 | 7 | 7 | 2 | 3 | 1 | 9 | 365 | 1 | 4 | 11 | 1 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 73 | 0 | 1 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 4 | 0 | 0 |
| 11 | 2 | 1 | 1 | 3 | 0 | 5 | 2 | 0 | 1 | 0 | 32 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |

**Fig. 2** EXtreme Gradient Boosting classification report (**a**) and confusion matrix (**b**) for the refined vaccine model. *Related to coding/media coverage/regional reporting practice. Notes: The classification report compares the model's prediction against the actual decision of the safety reviewer on the hold-out test set. For the confusion matrix, the diagonal shows counts of correct predictions for each decision (i.e., the matches between safety reviewer's decisions and the highest probability predictions). Cells off the diagonal show misclassifications (which true decision was confused with which incorrect predicted decision)

The confusion matrix did not reveal any dominant or systematic misclassification patterns: errors were distributed across multiple classes rather than concentrated in a particular pairwise confusion (Fig. 2b). This suggests the model's mistakes are largely case-specific rather than driven by a single pervasive bias or failure mode.

### 3.3 Explainability Insights

Feature importance and local explanation analyses were performed across multiple model versions for both vaccines and drugs. Here we present representative examples from the vaccine models to illustrate how explainability guided feature selection, debugging, and iteration. The SHAP analysis identified several features that meaningfully contributed to classifying specific decision categories.

At the global model-level (as defined in the Methods, Section 2.6), the top 5 contributors to the initial model (ranked by mean absolute SHAP value) were spontaneous EB05, historical alert flag, number of DPA alerts, serious event flag, and listed event flag (Supplementary Fig. 2). The listed event flag mainly drove predictions for "Listed event including synonym or symptom of", which was expected since reviewers often select this when an event is listed in the label. By contrast, identifying synonyms and symptoms is less rule-based and more subjective, underscoring the potential value of ML to support these triage decisions. The serious event flag predominantly influenced predictions for the "Event being monitored, no change" category, indicating that serious events are more likely to undergo monitoring.

Class-level explanations for the initial vaccine model showed that different features dominated different decision categories. For example, "Medication error with no specific pattern of concern" was primarily associated with alerts flagged for medication error events (Fig. 3a), while "Merits further investigation" was driven by high percentages of unknown action taken with vaccines (Fig. 3b). The second most influential feature for "Merits further investigation" was the percentage of cases with reporter verbatim present, with higher values pushing predictions towards this class.

Percentage with reporter verbatim absent was also influential (ranked 8$^{th}$) but showed the opposite and less consistent effect. This redundancy (two features representing complementary states of a binary attribute) led us to remove the percentage-absent feature (and similar duplicates) in later iterations, to reduce collinearity and simplify interpretation.

For "Other", the previous decision of "Other" strongly influenced the model toward predicting the same category. This decision is used by reviewers only when no other option fits due to the absence of a clear pattern or when more than one category applies and none is largely represented. Other influential features included the historical alert flag and the number of DPA alerts: historical alerts increased the likelihood of "Other", and low numbers of DPA alerts favoured this decision, whereas the effect of high DPA alert counts was less consistent (Fig. 3c).

To further illustrate our refinement approach, we present one pattern among others observed in the initial vaccine model confusion matrix: some alerts closed as "Event is likely explained by other factors" were misclassified as "Insufficient information available". We examined feature importance and local explanations for several such instances. Figure 4 presents one example: four of the five most influential features reflected percentages by age group or sex, while the 2$^{nd}$-ranked feature was the low number of cases with dosage information – the only feature consistent with "Insufficient information available". The series comprised only

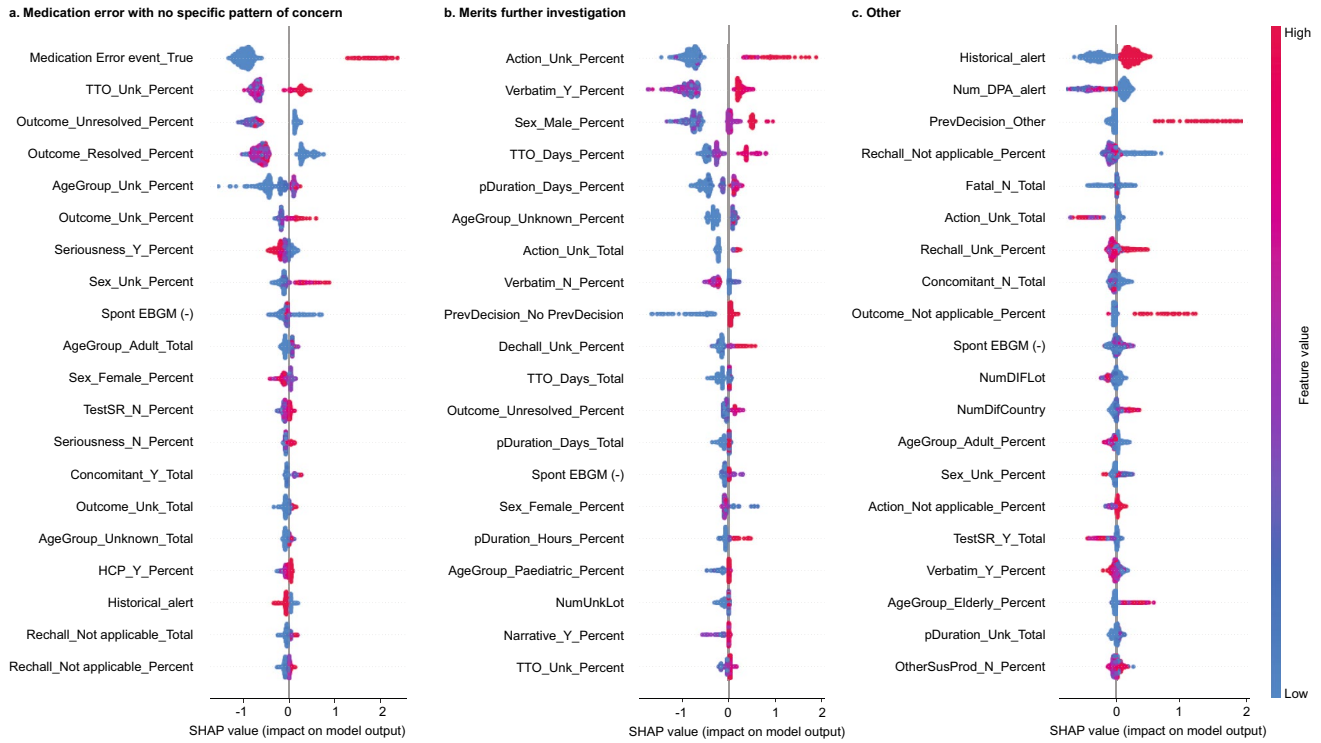


**Fig. 3** Feature importance and contribution for decision classes in the initial vaccine model. *DPA* disproportionality analysis, *EBGM* Empirical Bayes Geometric Mean, *HCP* healthcare provider, *SHAP* Shapley Additive Explanations, *TTO* time to onset. Notes: Class level SHAP summary plots. On the vertical axis, features are sorted by importance (top = most important). The horizontal axis indicates SHAP values (reflecting impact on the model output); points left of the vertical zero line decrease the model's propensity to predict the class, while points to the right increase it. Each dot is one data point (sample); dot colour indicates the feature value (blue = low, red = high). In each panel, one plot is shown per decision class

serious adult female cases. Although these demographic attributes pushed the model toward "Insufficient information available", PV expert review indicated that, in most misclassifications where sex and age dominated, they were unlikely to have driven triage decisions and instead reflected dataset bias from spurious correlations. Consequently, to mitigate bias and to support responsible AI governance, sex and age features were removed in subsequent vaccine model iterations.

Further analysis of misclassified decisions during the vaccine model refinement found examples where the model's top prediction diverged from the recorded decision due to human error in triage documentation, as later confirmed by safety reviewers. For example, the alerting PT *Extra dose administered* was incorrectly closed as "Other" but correctly predicted by the model as "Medication error with no specific pattern of concern". Other discrepancies reflected subjective judgements by less experienced reviewers: the alerting PT *Vaccination site movement impairment* was predicted as "Listed event including synonym or symptom of" (61% probability) but closed as "Event is likely explained by other factors". Given that this PT followed intramuscular vaccination, a "Listed event including synonym or symptom of" decision is reasonable because local reactions (e.g., swelling, pain on movement) that may reduce range of motion at the injection site may be listed for intramuscular vaccines, such as the one in the example.

For many misclassified alerts, examination of the full probability distribution showed the top-ranked prediction was still clinically acceptable. Model confidence correlated with accuracy: for predictions such as "Insufficient information available" (vaccines), "Event term is uninformative or non-specific" (drugs), "Medication error with no specific pattern of concern", "Event is likely explained by other factors", and "Listed event including synonym or symptom of", a predicted probability ≥ 60% was almost always correct. Conversely, when the top probability was low or only marginally above the 2$^{nd}$- and 3$^{rd}$-ranked probabilities, the top prediction was less reliable.

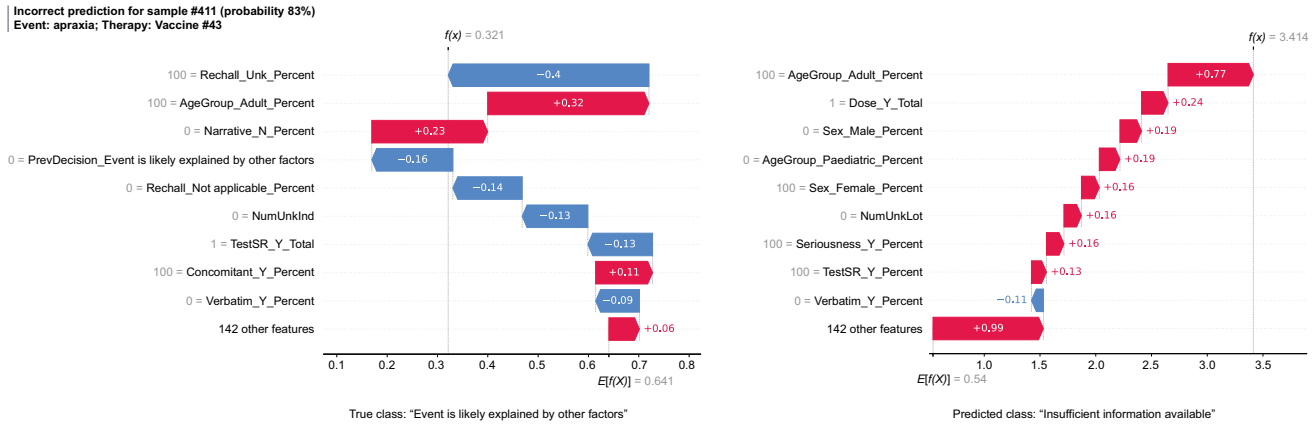


**Fig. 4** Feature importance and contribution for individual predictions in the initial vaccine model. *SHAP* Shapley Additive Explanations. Notes: Sample-level SHAP force plots showing each feature's contribution to a single prediction, shown separately for the true class and the predicted class. The plot starts at the model's expected value (baseline) and illustrates how individual feature SHAP values shift the prediction away from that baseline toward the model output. Red bars indicate features that push the prediction higher; blue bars indicate features that push it lower. Only the top contributing features are shown

## 3.4 Prospective Use

The prospective validation study included six vaccines across maturities: established (A, B), mid-life (C, D), and recently launched (E, F).

### 3.4.1 Top 3 Probability Predictions Performance

During the prospective validation, 881 vaccine alerts were received, including 35 first-time occurrences. Of these, 810 (92%) were closed in agreement with the model's 1st-ranked prediction, 41 (5%) with the 2nd-ranked prediction, and 13 (1%) with the 3rd-ranked prediction. For 652/881 alerts (74%), the top-ranked prediction had a probability ≥ 0.60, and in all of these cases the reviewers' final decision agreed with the model's 1st-ranked prediction. Only 17 PTs (2%) were closed with decisions outside the top 3 predictions. In eight of these, an acceptable decision was nevertheless present within the top 3. For example, for an alert on *Loss of personal independence in daily activities* following vaccination with vaccine E, the reviewer selected "Event is likely explained by other factors", whereas the model's top 3 predictions included "Event term is uninformative or non-specific", which the reviewer judged to be an acceptable alternative. Adjusting for reviewer-judged acceptable options, only 9/881 (1%) vaccine alerts were closed with decisions outside the top 3 (Table 5). Overall, 24/35 (69%) first-time alerts were closed with a decision among the model's top 3 predictions.

### 3.4.2 Inconsistencies and Errors Detected in the Current Process

During the prospective validation, reviewers occasionally selected options outside the top 3 or did not select the 1st-ranked prediction even when it had a high probability and matched the previous decision. For example, the alerting PT *Injection site injury* had a 1st-ranked prediction of "Reporting bias related to coding/media coverage/regional reporting practice" at 78% probability (and matched the previous decision), but the case had been closed as "Listed event including synonym or symptom of", which was the 3rd-ranked prediction (3%). Follow-up review with safety reviewers confirmed these discrepancies were due to human error.

**Table 5** Concordance between top 3 model predictions and safety reviewer decisions during prospective validation of the vaccines refined model

| | Alert PTs, *n* | Alert PTs closed with each prediction, n (%) | | | |
|---|---|---|---|---|---|
| | | 1st-ranked | 2nd-ranked | 3rd-ranked | Non-top 3 |
| All alert types | 881 | 810 (92%) | 41 (5%) | 13 (1%) | 17 (2%) |
| First-time occurrence | 35 | 15 (43%) | 9 (26%) | 0 (0%) | 11 (31%) |

*n* (%) number (percentage) of alert PTs, *PTs* Medical Dictionary for Regulatory Activities preferred term

Notes: The table presents the concordance of top 3 machine learning-generated predictions with the decisions made by safety reviewers for each alert

#### 3.4.3 Investigations Performed during Prospective Validation

One investigation involved vaccine E and PT *Blood glucose increase*. As a first-occurrence alert for a recent vaccine, the reviewer decided to investigate; the model's top prediction was also "Merits further investigation" (33% probability), although the reviewer reported their decision was not influenced by the model. The other involved vaccine B and PT *Pharyngeal erythema,* for which the 1st-ranked prediction was "Insufficient information available" (25%) and the 2nd-ranked was "Merits further investigation" (17%). The reviewer closed the investigation due to insufficient information based on the cumulative review of cases. Both investigations were ultimately concluded as non-validated signals, and notably each had "Merits further investigation" within the top 2 model predictions.

A separate investigation for PT *Bronchitis* with vaccine F was initiated despite the event being outside the model's top 3 predictions and was subsequently closed as non-validated. The model's top prediction for this alert was "Information does not change previous assessment" with a low probability (26%), indicating limited confidence and supporting the reviewer's decision not to follow the prediction (details are available in Supplementary Table 2).

#### 3.4.4 Time Saving

Participants in the prospective validation reported a self-estimated collective time saving of approximately 83 hours attributable to MLIT, representing roughly 24% of the total time spent on triage during the prospective validation study.

## 4 Discussion

We developed, refined, and prospectively evaluated an explainable ML-assisted triage tool—with model variants trained for vaccine and drug safety alerts—through structured retrospective testing and two distinct four-month prospective validation studies. Blinded validation and quality checks showed that, when prediction probabilities were high, the tool did not miss safety signals compared with manual triage. Across all phases, only one alert not predicted by the model was later assessed as a topic of interest. This supports continued use under controlled conditions, indicating that no missed validated signals were observed during the prospective validation.

MLIT improved triage accuracy, reproducibility, and operational efficiency. The refined XGBoost vaccine model achieved a weighted $F1$ of 0.81 (previously 0.70); drug model performance was similar. These gains reflect expert-informed tuning and the model's ability to learn reviewer decision patterns from increasing data volumes.

A major strength is MLIT's explainability, which supports user trust and acceptance. Model rationale, provided through feature importance analyses, aligned with established PV logic. This transparency aided debugging, informed feature selection (e.g., removing redundant complements or excluding inappropriate demographic variables), and reduced "black box" concerns.

Rule-driven decisions such as "Medication error with no specific pattern of concern" performed well due to strong event-tag information. "Listed event including synonym or symptom of" was similarly tag-dependent and showed good overall performance. For straightforward listed events, ML added relatively little beyond the listed event tag itself; however, the models still provided useful support in alerts where the event was a synonym or symptom of a listed concept, which are inherently more challenging. Further improving recognition of such synonym/symptom relationships will be important, particularly as DPA alerts for listed events are no longer triggered. Improving identification of these patterns is needed, particularly as DPA alerts for listed events are no longer triggered.

Four decision classes represented over 80% of alerts in both vaccine and drug datasets, and showed the strongest performance, making them ideal candidates for future automation. Heterogeneous or inconsistently applied classes (e.g., "Other") showed modest performance. Rather than removing these categories entirely, we focused on harmonising their use and excluded only low-support decision classes with insufficient data for reliable modelling (e.g., "Implausible relationship based on known properties/mode of action" and "Reporting bias related to coding/media coverage/regional reporting practice" for drugs). Confusion matrix review revealed interclass overlap, prompting decision harmonisation, misclassification review, and feature refinements that improved later iterations. The rare yet critical "Merits further investigation" class achieved acceptable recall post-refinement but will continue to require human oversight and enriched training data.

During the prospective validation of the vaccine model, 98% of alerts were closed using one of MLIT's top 3 predictions, with 92% matching the top prediction; similar concordance occurred for drugs. These results show strong alignment between model predictions and expert decisions and reinforce the tool's value in routine triage. In particular, the prospective analyses indicated that misclassifications were concentrated among alerts with lower and similar top-ranked probabilities, supporting the use of probability thresholds to distinguish high-confidence predictions from those requiring closer human scrutiny. First-occurrence

alerts showed lower performance due to the absence of prior-decision features. Despite this limitation, MLIT provided valuable insights that helped reduce time spent triaging these complex alerts. Nevertheless, these alerts should remain prioritised for manual review during deployment to ensure accuracy.

A blinded validation demonstrated robust consistency for MLIT: independent reviewers reached 100% agreement when using the tool, indicating that ML support does not bias the reviewer's decision and can standardise decision-making among experienced professionals.

Time-savings were substantial, although self-measured. During the prospective validations, users reported a 24% reduction in triage effort for vaccines, mirrored by results for drugs. These efficiencies are especially important for products generating high alert volumes.

An unexpected benefit was MLIT's ability to identify inconsistencies or human errors in alert closure. High-probability predictions sometimes flagged cases where reviewers had unintentionally altered decisions. Subsequent investigation confirmed user error, highlighting MLIT's potential as a quality-control mechanism.

Nonetheless, some data- and model-related limitations remain. Case data are dynamic and subject to continuous updates, and triage decisions reflect information available at the time alerts are generated. To preserve fidelity to those decisions, cases that were updated substantially after the alert date were excluded from model training; however, this may omit cases that contributed to the original review in an earlier state. Likewise, time lags between case receipt and processing can introduce uncertainty: some information used for model training may have been entered after the initial review and therefore was unavailable for triage. Dataset imbalance also poses challenges, as minority decision classes with fewer examples limit the model's ability to learn reliably, resulting in increased prediction uncertainty. Relatedly, important product and disease-specific knowledge and tacit expert judgement are not always captured in structured ICSR or alert fields, contributing to model uncertainty. Interpretation variability across reviewers also impedes model training. Certain decision categories (for example, "Information does not change previous assessment") were applied inconsistently in practice, sometimes used after signal assessment and other times to reaffirm a prior decision, reducing label consistency and introducing individual reviewer bias.

An additional limitation is the substantial influence of the "Previous decision" feature in SHAP analyses, which can introduce anchoring bias, risking the replication and perpetuation of historical triaging patterns and misclassifications. We partially mitigated this by incorporating features reflecting current evidence and a "Duration since last review" feature to provide temporal context and reduce inertia, but the risk of perpetuating systemic bias from historical data remains. Future work will pursue advanced feature engineering and fairness-aware regularisation to balance historical and current evidence. Concurrently, continuous monitoring for model drift or bias, within a human-in-the-loop system leveraging explainability, is essential to audit, override, and provide feedback on predictions unduly influenced by past outcomes, allowing continuous improvement.

ICSR data are mostly categorical and sparse, producing very high dimensionality that increases computational complexity, memory requirements, and the risk of overfitting. These characteristics also make it harder for the model to learn meaningful patterns from infrequent categories, necessitating extensive hyperparameter tuning and model optimisation, which extended processing time to several hours per run.

Retaining the latest alert per product-event pair prevented a time-based stratified split, thus limiting insights into temporal model performance. Future efforts will prioritise implementing time-aware splitting techniques to enable a more dynamic and robust evaluation.

Operational risks include false positives, when the tool assigns the top probability to "Merits further investigation" despite limited or ambiguous supporting patterns in the data. These are more common where training data are sparse, team-specific patterns are atypical, or model confidence is low. Retrospective testing already indicated that predictions with low probabilities or similar across candidates are not reliable; therefore, human oversight and critical judgement remain essential. In deployment, all three top-ranked predictions and their associated probabilities are presented to reviewers to preserve transparency; low and closely similar probabilities are explicitly highlighted in user training as markers of model uncertainty and triggers for increased scrutiny rather than automation. Periodic retraining and routine capture of user feedback, as part of quality-check processes, are planned to mitigate this risk and improve model performance.

A key concern is missing signals if alerts are incorrectly closed. In addition to human-in-the-loop review, targeted user training, and transparent outputs (top-ranked predictions, associated probabilities, and explanatory features), we implemented prospective validation-phase risk assessments and prospective user testing. Across both prospective validations (>2300 alerts), only two investigations occurred while the decision "Merits further investigation" was not within the top 3 predictions. These were performed for precautionary reasons and did not lead to validated signals; in both cases prediction confidence was low, reinforcing the need for cautious interpretation when certainty is limited. Other risks, such as inconsistent use of decision categories or excessive reliance on MLIT, are mitigated through layered controls

that maintain reviewer accountability while enabling incremental automation.

Before implementation, MLIT will undergo formal computerised system validation. Deployment will require its use in routine triage, with reviewers retaining authority to override MLIT predictions. A minimum six-month parallel phase will support confidence building, performance monitoring, and definition of conditions for expanded automation.

Future enhancements include automated generation of closure rationales, currently provided by safety reviewers as free text, and integration of external triggers (e.g., Health Authority requests). Incorporating such information is challenging due to inconsistent AE coding between internal systems and the fact that events listed in product labels do not always map directly to MedDRA concepts. Additionally, techniques such as semantic similarity metrics [37], leveraging MedDRA hierarchies and other clinical terminologies (e.g., Systematized Nomenclature of Medicine—Clinical Terms), could improve recognition of synonyms and symptom terms associated with listed events. There may be a future role for generative AI (including large language models) to support or interact with the tool to gain further efficiencies, provided that appropriate validation and guardrails are in place [38, 39].

## 5 Conclusion

Overall, our results suggest MLIT represents a meaningful advance in PV technology for signal management. Its integration into routine workflows has demonstrated improved accuracy, reproducibility, and efficiency while maintaining transparency and alignment with expert judgement. Continued refinement, collaborative development, and user engagement remain essential to sustain impact and expand opportunities to enhance drug-safety surveillance.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s40264-026-01696-0.

**Acknowledgements** The authors thank Yaseen Sehgal, Clément Bonal, Caroline Easterbrook, Sreevidya Sivaswamy, Alexandria Lynch, Mike Perrio, Boglárka Cseke, Philip Bryan, and Catherine Tregunno (GSK) for their contribution to the study. The authors would also like to thank the Akkodis Belgium platform for editorial assistance (by Petronela M. Petrar), manuscript coordination, and design support, on behalf of GSK.

**Funding** This work was supported by GSK.

## Declarations

**Conflict of Interest** All authors are/were employed by GSK at the time of study. OM, CR, SC, JLP, and AB hold financial equities in GSK. AB is an Editorial Board member of *Drug Safety*. AB was not involved in the selection of peer reviewers for the manuscript nor in any of the subsequent editorial decisions. The authors declare no other financial or non-financial relationships and activities.

**Ethics Approval** Not applicable.

**Consent to Participate** Not applicable.

**Consent for Publication** Not applicable.

**Availability of Data and Material** The data that support the findings of this study are available from the corresponding author (LC), upon reasonable request.

**Code Availability** Not applicable.

**Authors' Contributions** LC, OM, CR, SC, JLP, and AB participated in study concept/design; OM was involved in data acquisition; LC, OM, CR, AF, AK, EO, and MP participated to data analysis. All authors took part in data interpretation. All authors read and approved the final version.



## References

1. Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJW, Yuen N. Novel statistical tools for monitoring the safety of marketed drugs. Clin Pharmacol Ther. 2007;82:157–66. https://doi.org/10.1038/sj.clpt.6100258.
2. Lindquist M, Edwards IR, Bate A, Fucik H, Nunes AM, Ståhl M. From association to alert-a revised approach to international signal analysis. Pharmacoepidemiol Drug Saf. 1999;8(Suppl 1):S15–25. https://doi.org/10.1002/(SICI)1099-1557(199904)8:1+%3CS15::AID-PDS402%3E3.0.CO;2-B.
3. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf. 2009;18:427–36. https://doi.org/10.1002/pds.1742.
4. 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;93:539–46. https://doi.org/10.1038/clpt.2013.24.
5. CIOMS Working Group VIII. Practical Aspects of Signal Detection in Pharmacovigilance. 2010. https://cioms.ch/wp-content/uploads/2018/03/WG8-Signal-Detection.pdf. Accessed 19 Jan 2026.
6. Hauben M, Aronson JK. Defining "signal" and its subtypes in pharmacovigilance based on a systematic review of previous definitions. Drug Saf. 2009;32:99–110. https://doi.org/10.2165/00002018-200932020-00003.

7. Bate A, Edwards IR. Data mining in spontaneous reports. Basic Clin Pharmacol Toxicol. 2006;98:324–30. https://doi.org/10.1111/j.1742-7843.2006.pto_232.x.
8. Cutroneo PM, Sartori D, Tuccori M, Crisafulli S, Battini V, Carnovale C, et al. Conducting and interpreting disproportionality analyses derived from spontaneous reporting systems. Front Drug Saf Regul. 2023;3:1323057. https://doi.org/10.3389/fdsfr.2023.1323057.
9. Painter JL, Kassekert R, Bate A. An industry perspective on the use of machine learning in drug and vaccine safety. Front Drug Saf Regul. 2023;3:1110498. https://doi.org/10.3389/fdsfr.2023.1110498.
10. Warner J, Prada Jardim A, Albera C. Artificial intelligence: applications in pharmacovigilance signal management. Pharmaceut Med. 2025;39:183–98. https://doi.org/10.1007/s40290-025-00561-2.
11. Bate A, Michael Tregunno P. How is AI developing in pharmacovigilance? Ther Adv Drug Saf. 2026;17:20420986251412773. https://doi.org/10.1177/20420986251412773.
12. Glaser M, Littlebury R. Governance of artificial intelligence and machine learning in pharmacovigilance: what works today and what more is needed? Ther Adv Drug Saf. 2024;15:20420986241293303. https://doi.org/10.1177/20420986241293303.
13. Bate A, Stegmann J-U. Artificial intelligence and pharmacovigilance: what is happening, what could happen and what should happen? Health Policy Technol. 2023;12:100743. https://doi.org/10.1016/j.hlpt.2023.100743.
14. 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;36:371–88. https://doi.org/10.1007/s40264-013-0053-7.
15. 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;3:197–208. https://doi.org/10.1002/sam.10078.
16. Al-Azzawi F, Mahmoud I, Haguinet F, Bate A, Sessa M. Developing an artificial intelligence-guided signal detection in the Food and Drug Administration Adverse Event Reporting System (FAERS): a proof-of-concept study using galcanezumab and simulated data. Drug Saf. 2023;46:743–51. https://doi.org/10.1007/s40264-023-01317-0.
17. Bae J-H, Baek Y-H, Lee J-E, Song I, Lee J-H, Shin J-Y. Machine learning for detection of safety signals from spontaneous reporting system data: example of Nivolumab and Docetaxel. Front Pharmacol. 2021;11:602365. https://doi.org/10.3389/fphar.2020.602365.
18. Dauner DG, Leal E, Adam TJ, Zhang R, Farley JF. Evaluation of four machine learning models for signal detection. Ther Adv Drug Saf. 2023;14:20420986231219472. https://doi.org/10.1177/20420986231219472.
19. Pham M, Cheng F, Ramachandran K. A comparison study of algorithms to detect drug-adverse event associations: frequentist, Bayesian, and machine-learning approaches. Drug Saf. 2019;42:743–50. https://doi.org/10.1007/s40264-018-00792-0.
20. Dong G, Bate A, Haguinet F, Westman G, Dürlich L, Hviid A, et al. Optimizing signal management in a vaccine adverse event reporting system: a proof-of-concept with COVID-19 vaccines using signs, symptoms, and natural language processing. Drug Saf. 2024;47:173–82. https://doi.org/10.1007/s40264-023-01381-6.
21. Imran M, Bhatti A, King DM, Lerch M, Dietrich J, Doron G, et al. Supervised machine learning-based decision support for signal validation classification. Drug Saf. 2022;45:583–96. https://doi.org/10.1007/s40264-022-01159-2.
22. Bate A, Hobbiger SF. Artificial intelligence, real-world automation and the safety of medicines. Drug Saf. 2021;44:125–32. https://doi.org/10.1007/s40264-020-01001-7.
23. Kara V, Powell G, Mahaux O, Jayachandra A, Nyako N, Golds C, et al. Finding needles in the haystack: clinical utility score for prioritisation (CUSP), an automated approach for identifying spontaneous reports with the highest clinical utility. Drug Saf. 2023;46:847–55. https://doi.org/10.1007/s40264-023-01327-y.
24. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf. 2002;25:381–92. https://doi.org/10.2165/00002018-200225060-00001.
25. 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;21:603–10. https://doi.org/10.1002/pds.3226.
26. Pinheiro LC, Candore G, Zaccaria C, Slattery J, Arlett P. An algorithm to detect unexpected increases in frequency of reports of adverse events in EudraVigilance. Pharmacoepidemiol Drug Saf. 2018;27:38–45. https://doi.org/10.1002/pds.4344.
27. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. Advances in neural information processing systems (NeurIPS 2015), Canada, 7–12 December 2015, 649–57.
28. Navada A, Ansari AN, Patil S, Sonkamble BA. Overview of use of decision tree algorithms in machine learning. In: IEEE Control and System Graduate Research Colloquium; IEEE, Malaysia, 2011; pp. 37–42.
29. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016;4:218. https://doi.org/10.21037/atm.2016.03.37.
30. Wang Q. Support vector machine algorithm in machine learning. In: IEEE International Conference on artificial intelligence and computer applications (ICAICA); IEEE, China, 2022; pp. 750–6.
31. Rigatti SJ. Random forest. J Insur Med. 2017;47:31–9. https://doi.org/10.17849/insm-47-01-31-39.1.
32. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting. R package version 04-2. 2015;1:1–4. https://github.com/dmlc/xgboost.
33. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery; 2016. pp. 785–94. https://doi.org/10.1145/2939672.785.
34. Prodregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
35. Liang J. Confusion matrix: Machine learning. POGIL Act Clear. 2022;3. Retrieved from https://pac.pogil.org/index.php/pac/article/view/304.
36. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin Transl Sci. 2024;17:e70056. https://doi.org/10.1111/cts.70056.
37. Painter JL, Haguinet F, Powell GE, Bate A. Ontology-based semantic similarity measures for clustering medical concepts in drug safety. AMIA Annu Symp Proc. 2024;2024:979–88. https://doi.org/10.48550/arXiv.2503.20737.
38. CIOMS Working Group XIV. Artificial Intelligence in Pharmacovigilance. 2025. https://cioms.ch/wp-content/uploads/2022/05/CIOMS-WG-XIV_Draft-report-for-Public-Consultation_1May2025.pdf. Accessed 19 Jan 2026.
39. Ramcharran D, Painter JL, Kara V, Glaser M, Vanini M, Chalamalasetti VR, et al. Orchestrating generative AI in pharmacovigilance: predicting and preempting the unpredictable. Ther Adv Drug Saf. 2025;16:20420986251396023. https://doi.org/10.1177/20420986251396023.

## Authors and Affiliations

**Luciano Ciccarelli[1] · Olivia Mahaux[2] · Christie Roshan[3] · Ami Fofana[1] · Anna Kawka[4] · Emilia Occhipinti[1,7] · Mariapia Possidente[1] · Silvia Cenci[1] · Jeffery L. Painter[5] · Andrew Bate[3,6]**

✉ Luciano Ciccarelli
luciano.x.ciccarelli@gsk.com

✉ Olivia Mahaux
olivia.x.mahaux@gsk.com

1 GSK, Siena, Italy

2 GSK, Wavre, Belgium

3 GSK, London, UK

4 GSK, Warsaw, Poland

5 GSK, Durham, NC, USA

6 London School of Hygiene and Tropical Medicine, London, UK

7 Present Address: Daiichi Sankyo, Rome, Italy