Supplementary material

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§] Mariapia Possidente,[1] Silvia Cenci,[1] Jeffery L. Painter,[5] and Andrew Bate[3,6]

[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

[#]These authors contributed equally.

[§]Current affiliation: Daiichi Sankyo, Rome, Italy.

Corresponding authors:

Luciano Ciccarelli, GSK, Via Fiorentina 1, 53100 Siena, Italy; luciano.x.ciccarelli@gsk.com

Olivia Mahaux, GSK, Avenue Fleming 20, 1300 Wavre, Belgium; olivia.x.mahaux@gsk.com

Supplementary Table 1 Variable definitions and pre-processing for alert data, ICSR data, and product-level metadata

Variable	Value	Description	Included in models			
			Vx_init	Rx_init	Vx_ref	Rx_ref
Alert data						
Historical alert flag	Binary flag (1/0)	Tags alerts that were generated and triaged in the legacy signal management system.	✓			
Num DPA alerts Num Aberration alerts Num Trend alerts Vx only: Num TTO alerts Num Paediatric alerts Rx only: Num DME alerts	Numeric	Number of times the product-event pair triggered each alert rule during the study period (one count per active alert rule). Trend alert: spontaneous EB05≥2 (threshold lowered to 1.5 for newer products) and greater than EB95 calculated six months earlier. Paediatric alert: paediatric subgroup (age 0–<12 years) with new spontaneous reports and EB05≥2. For the refined models, the paediatric and trend alert rules were no longer active and were therefore not included.	✓	✓	✓	✓
PrevDecision	One of 14 possible decisions (Table 1) or "No previous decision"	Decision immediately preceding retained alert.	✓	✓	✓	✓
Spont_EB05	Numeric	EB05 (lower 90% credibility limit of EBGM) calculated at the most recent alert for the product-event pair during the study period.	✓	✓	✓	✓
N	Numeric	Total number of cases in the safety database reporting the alerting product and event.	✓	✓	✓	✓
Fatal_N	Numeric	Number of fatal cases in the safety database reporting the alerting product and event				✓
Serious_N	Numeric	Number of serious cases in the safety database reporting the alerting product and event.				✓
Medication error flag	Binary flag (1/0)	Indicates whether the alerting event matches the Medication error SMQ.	✓	✓	✓	✓
Serious flag	Binary flag (1/0)	Indicates whether the alerting event matches the auto-seriousness criteria.	✓	✓	✓	✓
DME flag	Binary flag (1/0)	Indicates whether the alerting event matches the DME CMQ.		✓	✓	✓
Listed flag	Binary flag (1/0)	Indicates whether the alerting event is listed in the product label.		✓	✓	✓
Percentage previous decision	Numeric (%)	For each decision class (except "Merits further investigation"), the percentage of prior decisions for the product-event pair that were of that class (excluding the last and penultimate decisions).			✓	✓
Num new cases	Numeric	New cases since the previous decision, computed as Num_cases(retained last alert) − Num_cases(previous alert). Negative values (e.g., due to case deletion or MedDRA coding changes) or missing previous counts were set to −1 by convention.			✓	✓

Variable	Value	Description	Included in models			
			Vx_init	Rx_init	Vx_ref	Rx_ref
Duration since last review	Numeric	Number of days between the period end of the retained alert and that of the immediately preceding retained alert.				✓
ICSR data						
Event outcome	Unknown Resolved Unresolved Fatal Not applicable Congenital anomaly For Rx refined model: Recovering	Categorical: grouped event outcome. Mappings used: 'Resolved' = {Resolved; Recovered/Resolved; Recovered; Recovered with Treatment; Resolved with Sequelae}; 'Unresolved' = {Improved; Recovering/Resolving; Unresolved; Not Recovered/Not Resolved; Unchanged; Lasting Damage; Worsened; Worse}; 'Fatal' = {Fatal; Foetal death; Abortion due to AE/Infection; Death due to AE/Infection; Death not due to AE/Infection}. 'Not applicable' and 'Congenital anomaly' are retained as separate categories. For the refined Rx model, an additional 'Recovering' category was created to separate 'Recovering/Resolving' and 'Improved' from 'Unresolved'.	✓	✓	✓	✓
Event duration	Binned categorical variable: Hours (0–<24 h) Days (24 h–<7 days) Weeks (7–<30 days) Months (30–<365 days) Years (≥365 days) Unknown (<0/missing)	Event duration is calculated from onset and stop dates when both are available.	✓	✓	✓	✓
Num product's indication Num unknown indication	Numeric	Number of distinct primary product indications at the product-event alert level (exclude null/unknown), and count of unknown indication instances. Indications are MedDRA coded; values recorded as 'Unknown' or 'Product used for unknown indication' were counted as unknown.	✓	✓		✓
Num product lot number Num unknown lot number	Numeric	Number of distinct lot numbers at the product-event alert level (exclude null/unknown), and count of unknown lot number instances. 'Unknown' values include missing entries or common placeholders (e.g. 'asku' or '%not %', '%unspec%', '%known%', '%available%', '%bekannt%', '%no batch%', '%invalid%', '%unk%', etc.) For the refined models, number of unknown lot number was included in Vx only.	✓	✓	✓	✓
Dose flag	Binary flag (1/0)	Indicates whether the dosage information is available (dose number >0, or dose/daily dose/total regimen dose > 0 with units present).	✓	✓	✓	✓
Action taken	Yes No Not applicable Unknown	Yes = action reported as 'Discontinued', 'Dose Adjusted', 'Interrupted', 'Dose Decreased', 'Dose Increased' or 'Withdrawn'. No = action reported as 'No Action Taken', 'No Change'. Not applicable = when not applicable or unknown or when vaccine as primary suspect product. For the refined Rx models, 'Unknown' was included as a separate value.	✓	✓	✓	✓
Dechallenge	Pos Neg	For Vx refined models, unknown dechallenge was set to 'Not applicable'; unknown included for Rx only.	✓	✓	✓	✓

Variable	Value	Description	Included in models			
			Vx_init	Rx_init	Vx_ref	Rx_ref
	Not applicable Unknown					
Rechallenge	Pos Neg Not applicable Unknown	For Vx refined models, unknown rechallenge was set to 'Not applicable', unknown was included for Rx only.	✓	✓	✓	✓
TTO	Hours (0–<24 h) Days (24h–<7 days) Weeks (7–<30 days) Months (30–< 365 days) Years (>365 days) Unknown (<0/missing)	Onset latency derived from dates and binned as hours/days/weeks/months/years (see bin edges above). Negative or missing values are coded as Unknown. In the refined models, free-text timing values (e.g. 'immediately') were incorporated where possible; for the Rx refined model, TTO was represented as a binary flag (A/0) indicating presence/absence of timing information.	✓	✓	✓	✓
Event listed flag	Binary flag (1/0)	Indicates whether the event is listed in the product label.	✓			
Age	Paediatric (0–11 years) Adult (12–64 years) Elderly (≥65 years) Unknown	Patient age in years enriched by the patient age group when available.	✓	✓		✓
Sex	Male Female Unknown	Patient sex as recorded in the safety database.	✓	✓		✓
Fatal flag	Binary flag (1/0)	Derived from customised fatal subpopulation in the company's safety database (i.e. there is a death date or a fatal outcome).	✓	✓		
Case seriousness flag	Binary flag (1/0)	Case reporting death, life-threatening, hospitalisation, disability, congenital anomaly or other significant medical event.	✓	✓	✓	✓
Num country of reporter	Numeric	Number of distinct reporting countries at the product-event alert level.	✓	✓	✓	✓
Long narrative flag	Binary flag (1/0)	Narrative length greater than one standard deviation above the mean narrative length across all cases.	✓	✓		
Verbatim flag	Binary flag (1/0)	Indicates presence of reporter's verbatim text within the narrative.	✓	✓		
HCP flag	Binary flag (1/0)	Indicates whether at least one report in the series was submitted by a HCP.	✓	✓		✓
Laboratory results flag	Binary flag (1/0)	Indicates presence of at least one laboratory or procedure result in the series. Valid result defined as non-missing and not equal to common placeholders (e.g., 'ND', 'UNK', '-', '+', 'UNKNOWN', 'NOT PROVIDED', 'NOT REPORTED', 'NOT DONE', 'AWAITED').	✓	✓	✓	✓
Medical history flag	Binary flag (1/0)	Indicates if at least one patient historical condition is reported.	✓	✓	✓	✓
Product history flag	Binary flag (1/0)	Indicates if at least one historical drug is reported.	✓	✓	✓	✓

Variable	Value	Description	Included in models			
			Vx_init	Rx_init	Vx_ref	Rx_ref
Comorbidity flag	Binary flag (1/0)	Indicates if at least one allergy, current condition or illness at vaccination is reported.	✓	✓	✓	✓
Concomitant products flag	Binary flag (1/0)	Indicates if at least one concomitant product is coded.	✓	✓	✓	✓
Other suspect products flag	Binary flag (1/0)	Indicates if at least one more suspect product to the one of interest is coded.	✓	✓	✓	✓
DME flag	Binary flag (1/0)	Indicates whether the alerting event matches the DME CMQ	✓			
Pregnancy event flag	Binary flag (1/0)	Indicates whether the alerting event matches the narrow SMQs 'Normal pregnancy conditions and outcomes', 'Pregnancy, labour and delivery complications and risk factors (excl. abortions and stillbirth)' and 'Termination of pregnancy and risk of abortion'.			✓	
Reactogenicity event flag	Binary flag (1/0)	Indicates whether the alerting event matches the SOC “General disorders and administration site conditions”.			✓	✓
Off-label use flag	Binary flag (1/0)	Indicates whether the alerting product was used off label.				✓
Pregnancy subpopulation	Binary flag (1/0)	Indicates whether the case concerned a pregnant patient.			✓	✓
Lack of efficacy flag	Binary flag (1/0)	Indicates whether the alerting event matches the vaccine specific “Lack of efficacy” CMQs and the “Lack of efficacy” SMQ.			✓	
Abuse subpopulation	Binary flag (1/0)	Indicates whether the case reports a drug abuse.				✓
Interaction subpopulation	Binary flag (1/0)	Indicates whether the case reports a drug interaction.				✓
Overdose subpopulation	Binary flag (1/0)	Indicates whether the case reports a drug overdose.				✓
Product-level metadata						
Product years in market	Vx: Seasonal; ≥10 years; <10 years Rx: Recent launch; Established	Vx: calculated from product international birth date and alert-period end as a proxy for novelty vs established safety profile. Rx: categorised based on current review frequency (less frequent review cycle was considered more established).	✓	✓	✓	✓
Vaccine category	Meningitis; DTP; Influenza; HAB; HZ; HPV; Malaria; MMRV; Polio; Rotavirus; RSV	Categories indicating vaccine groups; defined by shared antigens, exposed population or target disease.			✓	
Vaccine type	Adult; Paediatric; General	Category defined by vaccine indication: Adult (indicated for adults or elderly), Paediatric (indicated for paediatric populations), or General (indicated for both paediatric and adult/elderly populations).			✓	
Therapeutic area	Specialty medicines; Oncology; HIV products	High level therapeutic area.				✓

Variable	Value	Description	Included in models			
			Vx_init	Rx_init	Vx_ref	Rx_ref
Product route	Device; Inhaler; Injectable; Oral; Topical; Other	High level route of administration.				✓
Biologics flag	Yes No	Indicates whether the product is a biologic.				✓
Subtherapeutic area	HIV; Pain; Skin; Antihistamine; Anti-infective; Respiratory; Other; Supplements; Auto-immune diseases; Oncology; Gastroenterology; Neurology; Cardiology; Anti-inflammatory	More granular therapeutic area.				✓

AE, adverse event; DME, designated medical event; DPA, disproportionality analysis; DTP, diphtheria, tetanus, pertussis; EB05, lower bound (5th percentile) of the 90% credibility interval of the Empirical Bayes Geometric Mean ; CMQ, company MedDRA query; HAB, hepatitis A/B; HCP, healthcare provider; HZ, herpes zoster; ICSR, individual case safety report; MedDRA, Medical Dictionary for Regulatory Activities; MMRV, measles, mumps, rubella, and varicella; RSV, respiratory syncytial virus; Rx, Drugs; SMQ, standard MedDRA query; SOC, MedDRA system organ class; TTO, time to onset; Vx, Vaccines; ✓, included in the model.

Supplementary Table 2 Alerts with "Merits further investigation" decision

		Investigations, n						
	Period	Total	Predicted (top-ranked) - true positives	Top 3	Not in top 3	False positives - predicted investigation (top-ranked) but not investigated, n	Recall	Precision
Vaccines retrospective testing								
Refined model - test data	Up to July 2024	5	4	1 (non-validated)	0	8	0.8	0.33
Refined model - unseen data	Aug 2024–Oct 2024	3	0	1 (non-validated)	2 (refuted signals but triggered by HA request, context not available to MLIT)	-	-	-
Vaccines prospective validation[a]	Nov 2024–Feb 2025	2	1	1 (non-validated)	0	8	0.5	0.2
Drugs retrospective testing								
Refined model - test data	Up to January 2025	30	21	7 (non-validated)	4 (2 non-validated, 2 set in "Merits further investigation" by mistake and corrected in the signal management system)	18	0.7	0.54
Refined model - unseen data	Jan 2025–Mar 2025	5	2	1 (non-validated)	2 (non-validated)	-	-	-
Drugs prospective validation[a]	Apr 2025–Jul 2025	2	0	1 (non-validated)	1 (non-validated but term of interest for instream medical review)	4	0	0.33

HA, Health Authority; MLIT, Machine Learning for Intelligent Triage.
[a]Conducted on a subset of products.

Supplementary Fig. 1 Comparison of XGBoost classification reports for the initial and refined Vaccine models

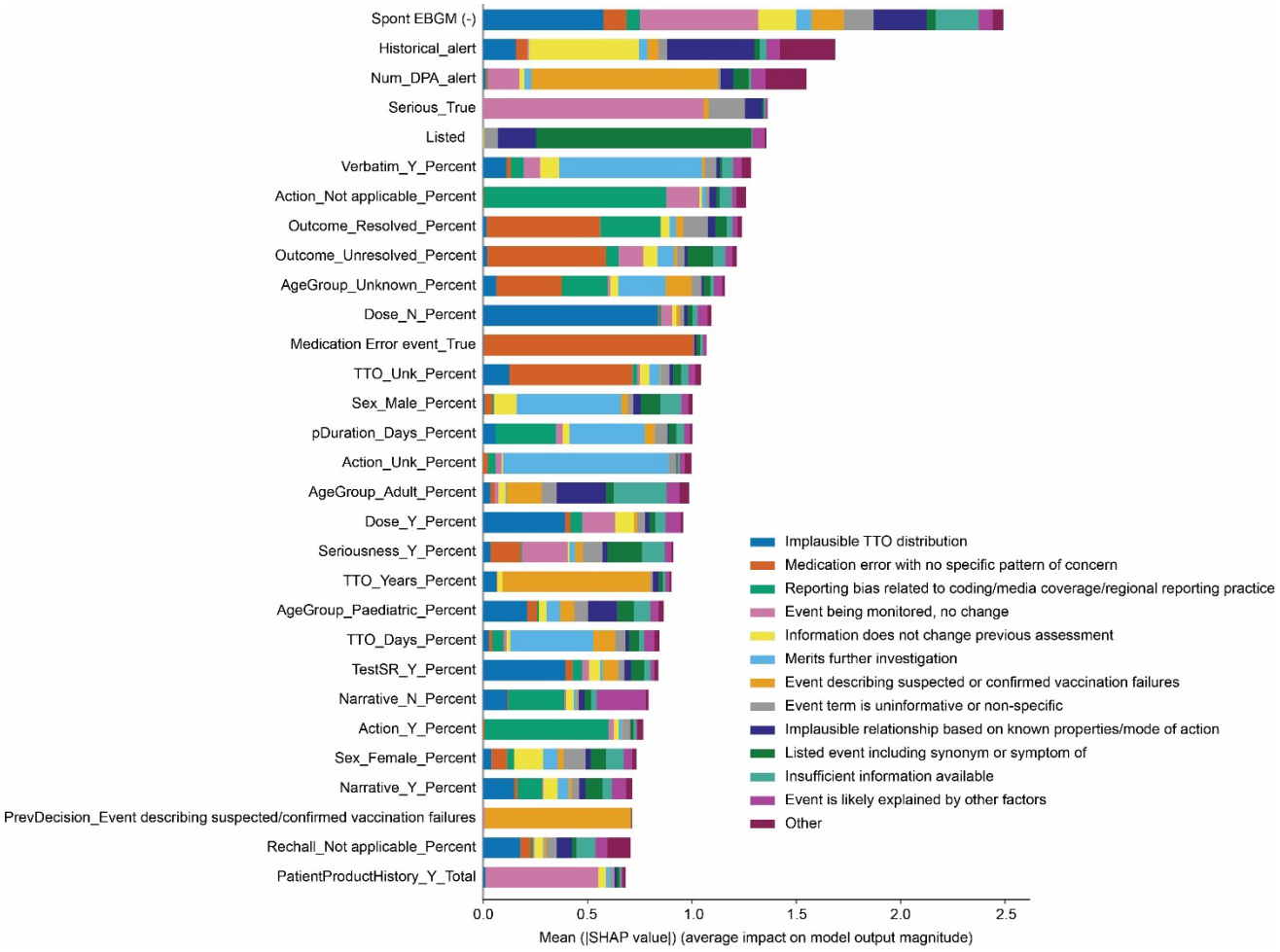
		Precision		Recall		F1-score		Support	
Predicted decision class	Model:	Initial	Refined	Initial	Refined	Initial	Refined	Initial	Refined
1	Event being monitored, no change	0.50	0.56	0.50	0.60	0.50	0.58	2	15
2	Event describing suspected or confirmed vaccination failures	0.83	0.42	0.62	0.77	0.71	0.54	8	13
3	Event is likely explained by other factors	0.63	0.86	0.57	0.70	0.59	0.77	136	179
4	Event term is uninformative or non-specific	0.11	0.55	0.05	0.62	0.07	0.58	19	39
5	Implausible time to onset distribution	0.00	NA	0.00	NA	0.00	NA	2	NA
6	Implausible relationship based on known properties/mode of action	0.82	0.23	0.89	0.35	0.85	0.28	252	17
7	Information does not change previous assessment	0.69	0.41	0.48	0.61	0.56	0.49	23	33
8	Insufficient information available	0.46	0.82	0.48	0.76	0.47	0.79	77	123
9	Listed event including synonym or symptom of	0.81	0.97	0.84	0.88	0.83	0.93	249	413
10	Medication error with no specific pattern of concern	0.77	0.94	0.87	0.96	0.82	0.95	39	76
11	Merits further investigation	0.67	0.33	0.67	0.80	0.67	0.47	3	5
12	Other	0.59	0.43	0.58	0.68	0.59	0.52	262	47
13	Reporting bias*	1.00	0.67	0.50	0.67	0.67	0.67	2	3
	Accuracy					**0.70**	**0.79**	**1074**	**963**
	Macro average	**0.61**	**0.60**	**0.54**	**0.70**	**0.56**	**0.63**	**1074**	**963**
	Weighted average	**0.69**	**0.83**	**0.70**	**0.79**	**0.70**	**0.81**	**1074**	**963**

NA, not applicable

*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.

Supplementary Fig. 2 Feature importance and contribution (SHAP analysis) for the initial Vaccine model - model-level explanation



DPA, disproportionality analysis; EBGM, Empirical Bayes Geometric Mean; SHAP, Shapley Additive Explanations; TTO, time to onset.

Supplementary methods and results – Drug models

Methods

The Drug models were developed following the same overall methodology as for Vaccines, with adaptations to accommodate differences in data, feature sets and triage considerations. For the initial Drug model, cumulative alerts up to November 2023 and individual case safety reports (ICSRs) extracted from the company's safety database on 8 March 2024 were used; ICSRs were mapped to the alerts using the same time window approach as for Vaccine models. As with the Vaccine models, Shapley Additive Explanations (SHAP)-based explainability and iterative expert review cycles were used to evaluate feature importance and refine the models.

The refined Drug model was retrained using an updated dataset that included cumulative alerts and ICSRs up to 1 January 2025. The principal differences from Vaccine models were in preprocessing and feature selection (**Supplementary Table 1**). For example, the initial Drug model included num_DME_alerts (designated medical event [DME] alerts are run for pharmaceutical products but not for vaccines). Correlation analyses performed for both the initial and refined models led to the exclusion of a different set of features due to high correlation (e.g. Action_yes_total, Dechall_not_app_total, Dose_yes_total, Comorb_yes_total, Subtherapeutic_area_oncology, Subtherapeutic_area_HIV, Seriousness_yes_total, Fatal_yes_total, NumUNKlot, Agegroup_adult_total, Action_yes_percent).

Preprocessing and feature representation were adjusted for the Drug model: values coded as "Unknown" and "Not applicable" were preserved as separate categories for action taken, dechallenge and rechallenge; time to onset (TTO) was represented as a presence/absence indicator in the refined Drug model; and duration since last review was added because it had greater influence on drug than vaccine triage decisions. Features retained for the Drug refined models but not for the Vaccine ones included age, sex, and counts of distinct and unknown product indications, as well as a healthcare provider-report flag. Drug-level metadata (therapeutic and subtherapeutic area, route of administration, biologics flag), and drug-specific features (subpopulation flags for abuse, interaction and overdose, and an off-label use flag) were also added. Alerts for events tagged as listed in the product label were not retained for training in the "Listed event including synonym or symptom of" decision category.

Certain decision categories with insufficient samples were excluded to avoid sparse outcome classes: "Implausible relationship based on known properties/mode of action" and "Reporting bias related to coding/media coverage/regional reporting practice".

These adaptations were made to reflect drug-specific reporting patterns and clinical considerations, align model inputs with expert decision-making, and optimise predictive performance across diverse therapeutic areas.

The four-month prospective validation study was conducted from April 2025 to July 2025.

Results

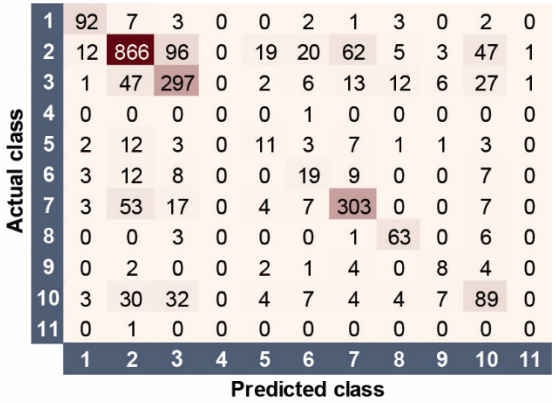
The refined Drug model was trained and tested on 2,890,733 ICSRs and 15,393 alert decisions covering all drugs. It improved on the initial model's F1 of 0.72 (**Supplementary Figure 3a**) to achieve a weighted average F1-score of 0.79 (**Supplementary Figure 4a**).

Supplementary Fig. 3 XGBoost classification report (a) and confusion matrix (b) for the initial Drug model

a

	Predicted decision class	Precision	Recall	F1-score	Support
1	Event being monitored, no change	0.79	0.84	0.81	110
2	Event is likely explained by other factors	0.84	0.77	0.80	1131
3	Event term is uninformative or non-specific	0.65	0.72	0.68	412
4	Implausible relationship based on known properties/mode of action	0.00	0.00	0.00	1
5	Information does not change previous assessment	0.26	0.26	0.26	43
6	Insufficient information available	0.29	0.33	0.31	58
7	Listed event including synonym or symptom of	0.75	0.77	0.76	394
8	Medication error with no specific pattern of concern	0.72	0.86	0.78	73
9	Merits further investigation	0.32	0.38	0.35	21
10	Other	0.46	0.49	0.48	180
11	Reporting bias*	0.00	0.00	0.00	1
	Accuracy			**0.72**	**2424**
	Macro average	**0.46**	**0.49**	**0.48**	**2424**
	Weighted average	**0.73**	**0.72**	**0.72**	**2424**

b



*Related to coding/media coverage/regional reporting practice. All drugs – train/test dataset of 9,693 alerts and 1,854,220 ICSRs.

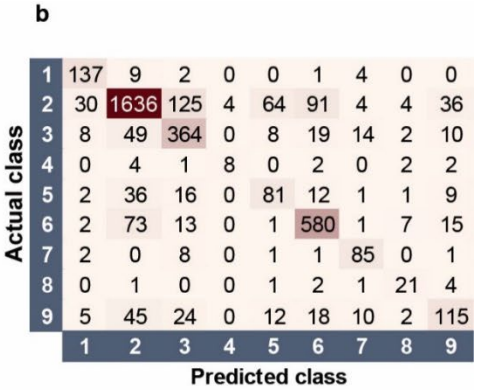
Notes: The classification report compares the model's prediction against the actual decision of the safety reviewer on the hold-out test set.
For the confusion matrix, the diagonal shows counts of correct predictions for each decision (i.e. the matches between safety reviewer's decisions and the highest probability predictions). Cells off the diagonal show misclassifications (which true decision was confused with which incorrect predicted decision).

Supplementary Fig. 4 XGBoost classification report (a) and confusion matrix (b) for the refined Drug model

a

	Predicted decision class	Precision	Recall	F1-score	Support
1	Event being monitored, no change	0.74	0.90	0.81	153
2	Event is likely explained by other factors	0.88	0.82	0.85	1994
3	Event term is uninformative or non-specific	0.66	0.77	0.71	474
4	Information does not change previous assessment	0.67	0.42	0.52	19
5	Insufficient information available	0.48	0.51	0.50	158
6	Listed event including synonym or symptom of	0.80	0.84	0.82	692
7	Medication error with no specific pattern of concern	0.71	0.87	0.78	98
8	Merits further investigation	0.54	0.70	0.61	30
9	Other	0.60	0.50	0.54	231
	Accuracy			**0.79**	**3849**
	Macro average	**0.67**	**0.70**	**0.68**	**3849**
	Weighted average	**0.79**	**0.79**	**0.79**	**3849**

b



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).

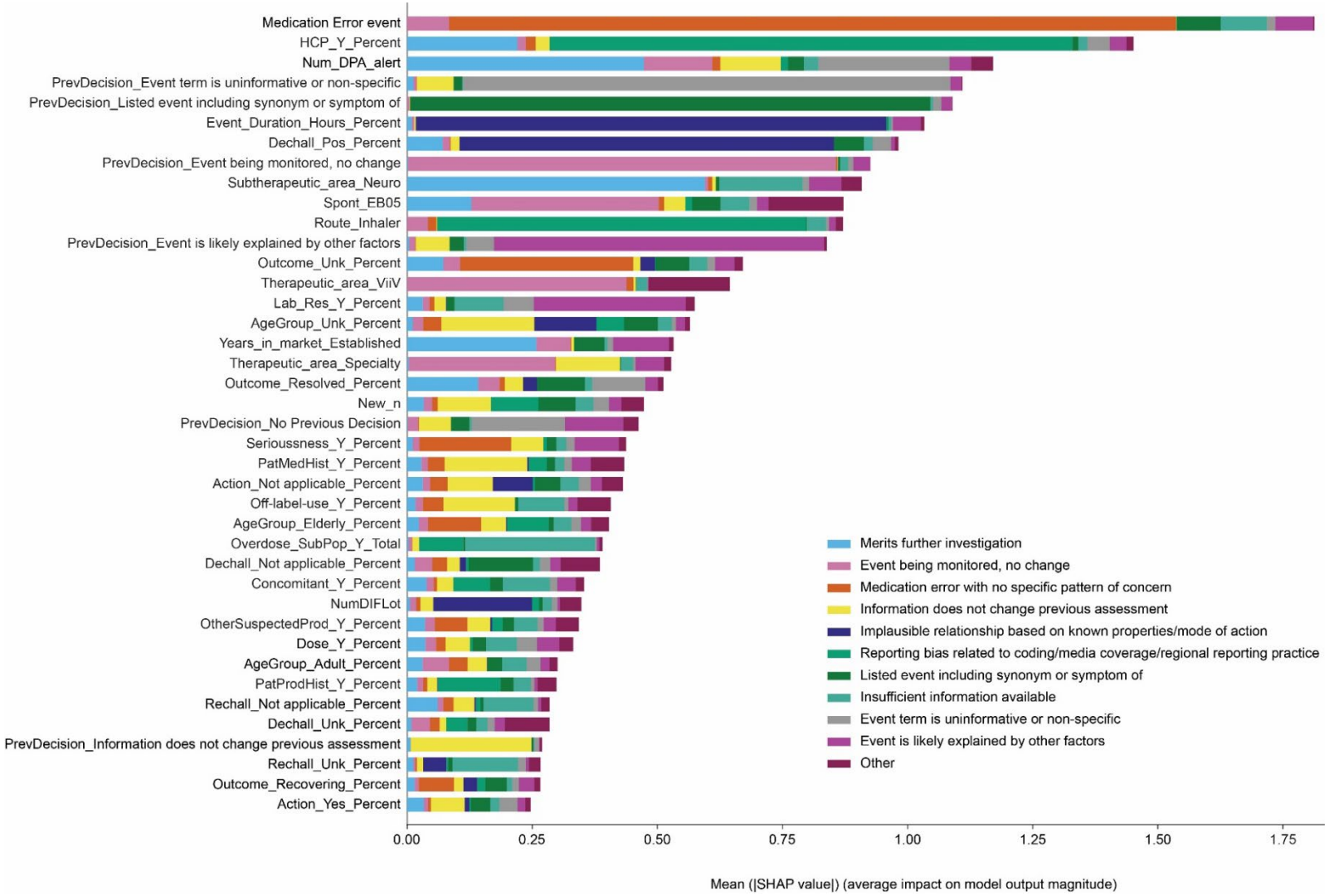
"Event is likely explained by other factors" (F1=0.85; support=1994) and "Listed event including synonym or symptom of" (F1=0.82; support=692) showed the strongest performance, reflecting high precision and recall for these common, well-represented closure decisions. "Event being monitored, no change" showed good performance with smaller support (153), with a strong F1-score (0.81) driven by high recall despite more modest precision, suggesting the model effectively predicts these events when present. Several classes showed lower F1-scores and/or limited support. Notably, "Insufficient information available" (F1=0.50; support=158) and "Other" (F1=0.54; support=231) had modest performance, reflecting heterogeneity or ambiguity in these decisions (**Supplementary Figure 4a**). When misclassified, "Insufficient information available" is mostly predicted as "Event is likely explained by other factors", "Event term is uninformative or non-specific", or "Listed event including synonym or symptom of" (**Supplementary Figure 4b**). In practice, these decisions were sometimes applied together when multiple decisions were used to document alert closure, illustrating the overlapping definitions and ambiguous case patterns that likely contribute to classification difficulty. Similarly to the Vaccine model, the Drug "Other" category showed substantial heterogeneity and inconsistent application: many selected "Other" alerts were predicted as a variety of other decision classes, and many predictions labelled "Other" were in fact other decisions, as illustrated in the confusion matrix (**Supplementary Figure 4b**).

"Merits further investigation" (F1=0.61; support=30) and "Information does not change previous assessment" (F1=0.52; support=19) were also predicted with lower confidence, largely due to small sample sizes. Importantly, the critical "Merits further investigation" decision achieved an acceptable recall of 0.70 (21/30 correctly predicted). Examination of the confusion matrix (**Supplementary Figure 4b**) showed that of the nine misclassified "Merits further investigation" alerts, four were predicted as "Other," two as "Listed event including synonym or symptom of" and one each as "Event is likely explained by other factors", "Insufficient information available", and "Medication error with no specific pattern of concern". Review and explainability analysis revealed that two of these nine were investigated as signals by mistake and documented as such; the remaining seven were concluded as non-validated signals. Three of these seven had relatively high probabilities ("Medication error with no specific pattern of concern": 79% 1st-ranked probability with "Merits further investigation as 3rd-ranked probability; "Other": 69% 1st-ranked probability, "Merits further investigation" not in top 3 and "Other": 68% 1st-ranked probability with "Merits further investigation" as 2nd-ranked probability), while the other four had low top prediction probabilities, and in three of those four the model placed "Merits further investigation" as the 2ndranked prediction (more details are provided in **Supplementary Table 2**).

Explainability insights

For the initial Drug model, the top 5 contributors were the medication error event flag, the percent of cases reported by healthcare providers, the cumulative number of disproportionality analysis alerts, "Event term is uninformative or non-specific" as previous decision, and "Listed event including synonym or symptom of" as previous decision (**Supplementary Figure 5**).

Supplementary Fig. 5 Feature importance and contribution (SHAP analysis) for the initial Drug model - model-level explanation

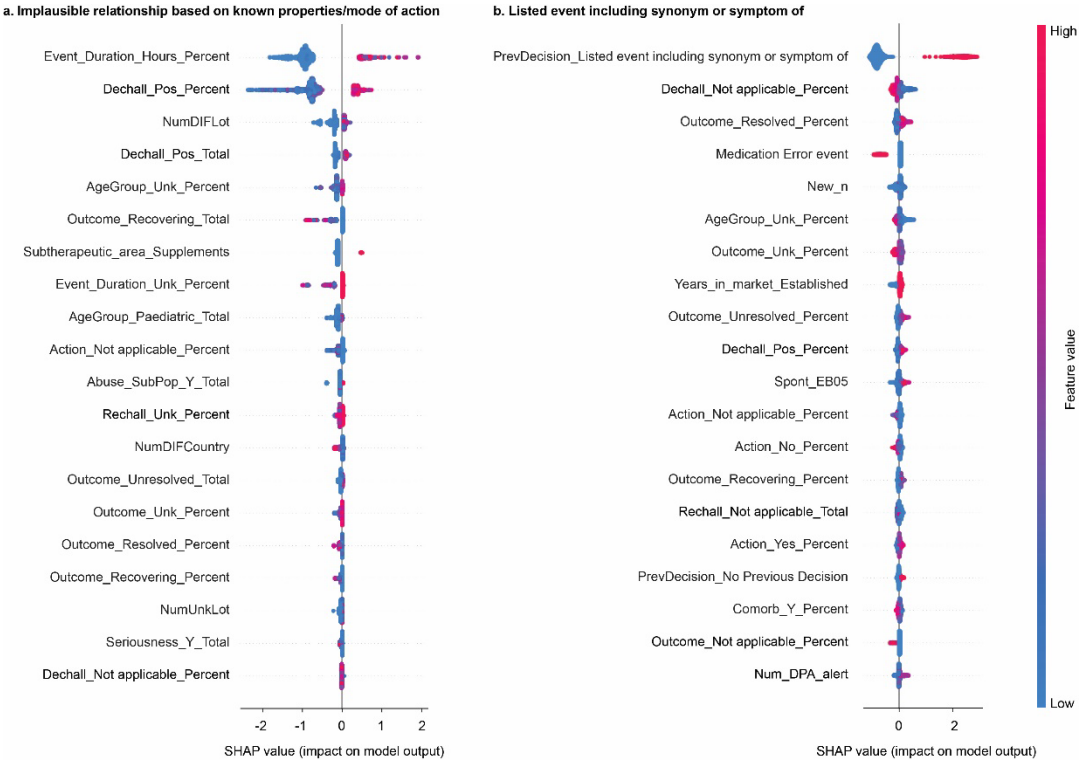


DPA, disproportionality analysis; EB05, lower bound (5th percentile) of the 90% credibility interval of the Empirical Bayes Geometric Mean; HCP, healthcare provider; SHAP, Shapley Additive Explanations.

Previous decisions were found to significantly influence predictions, particularly for "Event term is uninformative or non-specific" and "Listed event including synonym or symptom of" (class-level explanation, **Supplementary Figure 6b**). This is particularly relevant for these decisions, as they are closely tied to the alert preferred terms (PTs) and rely less on additional clinical context from the case series compared to other decisions. To enhance predictions for decisions where previous decisions are less consistently reused and where changes in decisions occur over time, the percentage of previous decisions for all decision classes was incorporated into subsequent model iterations.

As was the case for the Vaccine model, class-level explanations for the initial Drug model highlighted distinct feature contributions across decision categories. For the decision category "Implausible relationship based on known properties or mode of action", the most impactful features were the percentage of cases with event resolved in hours, and percentage of cases with positive dechallenge, with high values for both features driving predictions for this decision (**Supplementary Figure 6a**). This aligns with the clinical rationale that when an event resolves quickly, it increases the likelihood of dechallenge being observed, making it logical for both features to drive predictions in the same direction. Positive dechallenge further supports the potential causal role for the product, unless additional evidence contradicts this based on the product's known properties or mode of action.

Supplementary Fig. 6 Feature importance and contribution for decision classes in the initial Drug model

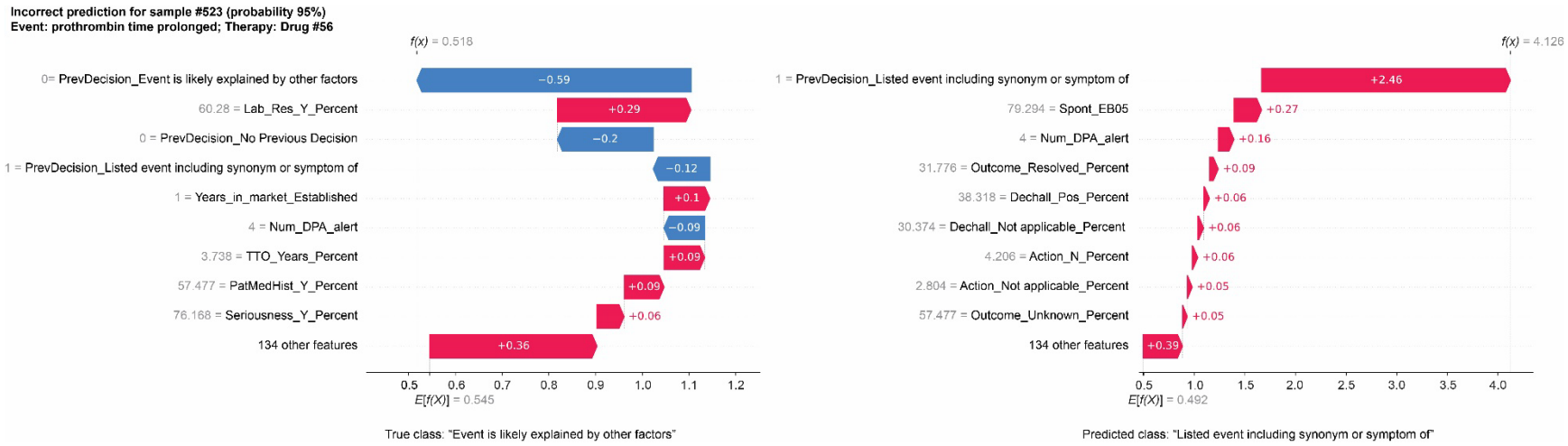


DPA, disproportionality analysis; EB05, lower bound (5th percentile) of the 90% credibility interval of the Empirical Bayes Geometric Mean; SHAP, Shapley Additive Explanations.

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.

During the refinement of the Drug model, the confusion matrix indicated, among other patterns, that a proportion of alerts closed with the decision “Event is likely explained by other factors” were incorrectly predicted as “Listed event including synonym or symptom of”. Sample-level explanations for several of these examples were analysed and one such instance is presented in **Supplementary Figure 7**. The explanations suggested that both decisions might be reasonable; however, discussions with the safety reviewer revealed that “Event is likely explained by other factors” had been erroneously selected. The event in question was, in fact, a synonym of a listed event for the drug, which justified the model’s prediction of “Listed event including synonym or symptom of”.

Supplementary Fig. 7 Feature importance and contribution for individual predictions for the initial Drug model



DPA, disproportionality analysis; EB05, lower bound (5th percentile) of the 90% credibility interval of the Empirical Bayes Geometric Mean; TTO, time to onset.

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.

Prospective use

The prospective validation included 17 marketed drugs from multiple divisions, spanning a range of market maturities.

Top 3 probability predictions performance

During the prospective validation study, 1488 alerts were received, including 75 first-time occurrences (**Supplementary Table 3**). Of these, 1250 (84%) were closed in agreement with the model's 1st-ranked prediction, 123 (8%) with the 2nd-ranked prediction and 48 (3%) with the 3rd-ranked prediction. For 1290/1488 alerts (87%), the top-ranked prediction had a probability ≥0.60, and in most of these cases the reviewers' final decisions agreed with the model's 1st-ranked prediction, with only a small number of exceptions. Only 67 PTs (5%) were closed with decisions outside the model's top 3 predictions. In 23 of these alerts, reviewers considered at least one of the top 3 model predictions acceptable but selected an alternative rationale, and thus a different decision in practice. After accounting for those reviewer-judged acceptable options, only 44 of 1488 (3%) drug alerts were ultimately closed with decisions not represented in the model's top 3 predictions.

Overall, 66/75 (88%) first-time alert occurrences were closed with a decision among the model's top 3 predictions.

Supplementary Table 3 Concordance between top 3 model predictions and safety reviewer decisions during prospective performance testing of the Drug refined model

		Alert PTs closed with each prediction, n (%)			
	Alert PTs, n	1st-ranked	2nd-ranked	3rd-ranked	Non-top 3
All alert types	1488	1250 (84%)	123 (8%)	48 (3%)	67 (5%)
First-time occurrence	75	43 (57%)	16 (21%)	7 (9%)	9 (12%)

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

Inconsistencies and errors detected in the current process

The Drug model also enabled identification of user-related errors and inconsistencies during the prospective validation. In several instances, safety reviewers selected decisions that differed from the previous decision and fell outside the model's top 3 predictions, even though the model's 1stranked prediction had a high probability and matched the prior decision. For example, three PTs with a previous decision of "Event is likely explained by other factors" for which the model's 1stranked prediction was also "Event is likely explained by other factors" (*Platelet count increased*, 87%; *Blood potassium increased*, 87%; *Ascites*, 96%) were closed as "Insufficient information available," a class that did not appear in the model's top 3. During the review period, these PTs had respectively only two, two, and one new ICSR added to cumulative case series of ≥94, 84, and 204 ICSRs, making a change in overall assessment unlikely. Subsequent investigation with the safety teams confirmed these discrepancies were due to human error.

Investigations performed during the prospective validation

Two investigations were performed. The first concerned PT *Rosacea,* for which the model's 2nd-ranked prediction was "Merits further investigation" (43% probability) and was concluded as non-validated signal. The second involved PT *Spontaneous penile erection,* where the model's top prediction was "Listed event including synonym or symptom of" with a low probability (37%) and "Merits further investigation" was not among the top 3 predictions. This alert was a first-time occurrence and considered a term of interest for

instream medical review; notably, the investigation did not lead to a label update and it was not added to the Risk Management Plan.

Time saving

Participants in the prospective validation of the Drug model reported an estimated collective time saving of about 10 hours attributable to the Machine Learning for Intelligent Triage tool, corresponding to approximately 27% of their triage time during the prospective validation study.