Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Computational Pharmacovigilance of Newly Launched Oral GLP-1 Receptor Agonists: Multi-Measure Signal Detection and Temporally Validated Machine Learning on FAERS

Tahsien Al-Quraishi

tquraishi@mit.edu.au

Melbourne Institute of Technology; Al-Ayen Iraqi University https://orcid.org/0000-0003-3814-0431

Research Article

Keywords: computational pharmacovigilance, disproportionality analysis, machine learning, conformal prediction, dataset shift, GLP-1 receptor agonists

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11792528/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Computational Pharmacovigilance of Newly Launched Oral GLP-1 Receptor Agonists: Multi-Measure Signal Detection and Temporally Validated Machine Learning on FAERS

Tahsien Al-Quraishi[1,2*]

[1*]School of Business, Melbourne Institute of Technology, Melbourne, Victoria, Australia.
[2]College of Science, Al-Ayen Iraqi University, Nasiriyah, Iraq.

Corresponding author(s). E-mail(s): tquraishi@mit.edu.au;

Abstract

Purpose Orforglipron and oral semaglutide 25 mg entered the US obesity market in 2026 without post-marketing safety data. This study aimed to detect early adverse-event reporting signals for both agents and to test whether machine-learning models that triage serious reports remain reliable after a product launch.

Methods We built a computational pharmacovigilance pipeline on deduplicated reports from the US Food and Drug Administration Adverse Event Reporting System (FAERS) received from January 2024 to September 2026. Each oral product was compared with injectable incretins reported over the same period using four disproportionality measures with false-discovery-rate control, sensitivity analyses and positive and negative controls. Four classifiers trained on 2024–2025 reports were validated on 2026 reports against simple baselines, with calibration and conformal prediction.

Results Of 212,284 reports, 3043 concerned orforglipron and 1543 oral semaglutide 25 mg. Gastrointestinal intolerance and perceived non-response were disproportionately reported for both agents (reporting odds ratios 1.4–4.0), consistently across analyses, and no pancreatic, biliary, hepatic or renal signal emerged. Gradient boosting reached an area under the curve of 0.86, but a simple reporting-process baseline reached 0.84, performance fell to 0.74 for orforglipron, and only class-conditional conformal prediction restored target coverage of serious reports.

Conclusion Early reports for both oral agents are dominated by gastrointestinal intolerance and perceived non-response; these hypothesis-generating signals cannot establish incidence or causality. Report-triage models largely learn how reports are filed and need product-specific monitoring after launch.

Keywords: computational pharmacovigilance, disproportionality analysis, machine learning, conformal prediction, dataset shift, GLP-1 receptor agonists

1 Introduction

Glucagon-like peptide-1 receptor agonists (GLP-1RAs) have reshaped the pharmacological management of obesity. In a nationally representative US survey conducted in late 2025, 18% of adults reported having used a GLP-1RA and 12% reported current use [1]. Randomised trials of injectable semaglutide and the dual glucose-dependent insulinotropic polypeptide/GLP-1 receptor agonist tirzepatide showed substantial weight reduction with recognised gastrointestinal

adverse effects [2–4]. Two oral agents have since been approved in the USA: oral semaglutide 25 mg, studied in OASIS 4 and approved in December 2025, and orforglipron, a small-molecule non-peptide agonist evaluated in ATTAIN-1 and approved on 1 April 2026 [5–8].

These trials had limited follow-up and samples too small to detect uncommon events, and use after launch extends to populations and settings the trials did not represent. Post-marketing surveillance has repeatedly extended the safety profile of the injectable agents, but no study has yet examined either oral obesity agent after launch (Sect. 2). Spontaneous-reporting analyses are also vulnerable to confounding by indication, notoriety and the reporting channel [9], and reports of perceived therapeutic non-response, such as "drug ineffective" or "weight loss poor", are usually treated as noise rather than as an outcome.

Machine learning has been proposed across the pharmacovigilance lifecycle [10], including classifiers that prioritise incoming reports for review. Such models are useful only if they add to simple rules and remain reliable when the reporting stream changes, yet published models have been evaluated on reports from the period in which they were developed (Sect. 2.3). The launch of a new product, which changes who reports, what is reported and how often reports are serious, is a natural test of this reliability.

We therefore built a computational pharmacovigilance pipeline on FAERS with three objectives: (1) to detect disproportionate reporting of prespecified event groups and individual preferred terms for orforglipron and oral semaglutide 25 mg relative to contemporaneously reported injectable incretins, using four agreeing measures with false-discovery-rate control; (2) to test whether perceived therapeutic non-response is disproportionately reported and robust to alternative analytic choices; and (3) to develop serious-report classifiers before these products were marketed and validate them temporally against simple baselines, with calibration, conformal uncertainty under covariate shift and subgroup analyses. The third objective concerns administrative report classification, not the probability that a treated patient will be harmed.

The remainder of the paper is organised as follows. Sect. 2 reviews related work from the past five years and identifies the gaps this study addresses. Sect. 3 describes the data, signal-detection methods and machine-learning pipeline, Sect. 4 reports the results, Sect. 5 interprets them and states the limitations, and Sect. 6 concludes.

2 Related work

This section reviews studies published between 2021 and 2026 in three areas that the present study connects: post-marketing safety studies of GLP-1RAs, methodological work on disproportionality analysis, and machine learning for pharmacovigilance. For each study, we summarise the problem addressed, the data and method used, the main results, and the strengths and limitations relevant to the present work. Table 1 compares these studies with the present study.

2.1 Post-marketing safety studies of GLP-1 receptor agonists

Most real-world safety evidence for GLP-1RAs comes from disproportionality analyses of spontaneous reports. Shu et al. [11] addressed the incomplete evidence on the gastrointestinal tolerability of semaglutide by applying the ROR to 5442 FAERS cases from 2018 to 2022; they detected 45 gastrointestinal signals, 17 of them unexpected, ranked them with a clinical-priority scale and estimated time to onset. Liu et al. [12] compared five marketed GLP-1RAs in 81,752 FAERS reports and found that semaglutide had the strongest association with nausea (ROR 7.41) and vomiting (ROR 6.67) and liraglutide with pancreatitis (ROR 32.67). Both studies characterised class effects with large samples, but each relied on a single measure, the ROR, and both preceded the obesity-dose products.

Several studies examined route of administration. Xiong et al. [13] analysed 2398 FAERS reports of semaglutide tablets for type 2 diabetes from 2019 to 2023 with four agreeing measures (ROR, PRR, the Bayesian confidence propagation neural network and the multi-item gamma–Poisson shrinker) and identified 93 preferred-term signals, including unexpected signals for acute cholecystitis and abducens nerve paralysis. Niu et al. [14] contrasted 2496 oral and 16,346 subcutaneous semaglutide records from 2017 to 2023 and reported more gastrointestinal events and a faster onset with oral administration, and Zhang et al. [15] combined

FAERS with online patient reviews and reported new route-specific signals. These studies are the closest antecedents of the present work, but all concerned the 3–14 mg diabetes formulation and preceded the obesity-dose products, and, to our knowledge, none analysed perceived non-response as a prespecified outcome.

Studies of specific safety concerns have used global databases and cohorts. Schoretsanitis et al. [16] used the World Health Organization database and found disproportionate reporting of suicidal ideation for semaglutide (ROR 1.45, 95% CI 1.18–1.77), robust to three active comparators. Kim et al. [17] compared liraglutide, semaglutide and tirzepatide in VigiBase with adjusted RORs and found associations of semaglutide with suicidal ideation, hair loss and vision loss, and Arslan [18] compared 60,423 semaglutide and 143,126 tirzepatide FAERS reports from 2022 to 2026, finding higher reporting odds with semaglutide for impaired gastric emptying, optic ischaemic neuropathy and acute kidney injury. Outside spontaneous reporting, Sodhi et al. [19] used United States claims data and associated GLP-1RAs prescribed for weight loss with pancreatitis, bowel obstruction and gastroparesis relative to bupropion–naltrexone, and Hathaway et al. [20] associated semaglutide with nonarteritic anterior ischaemic optic neuropathy in a single-centre matched cohort (hazard ratios 4.28 and 7.64). The cohort designs provide incidence-based estimates, but their small event counts and single-source populations limit generalisability, and none of these studies included an oral obesity agent.

2.2 Methodological challenges in disproportionality analysis

Recent methodological work has clarified why disproportionality findings require careful design. The READUS-PV guideline sets out reporting standards and lists confounding by indication, notoriety and reporting channel as threats to validity [9]. Gravel et al. [21] showed that the choice of reference set, whether the full database, a therapeutic class or an active comparator, changes the direction and magnitude of these biases. New products attract stimulated reporting in their first years [22, 23], so comparisons with established drugs reported over different periods can produce spurious signals, and negative controls [24] can expose residual bias. To our knowledge, none of the GLP-1RA studies above combined active comparators restricted to the same calendar period with positive and negative controls, multiple disproportionality measures and FDR control.

2.3 Machine learning for pharmacovigilance and reliability under shift

Machine learning has been applied to several pharmacovigilance tasks [10]. For report triage, Routray et al. [29] trained neural networks on a stratified random sample of manufacturer reports and determined seriousness with 83.0% accuracy in post-marketing reports, and Lieber et al. [25] combined structured fields with natural-language features in a bagging classifier that prioritised Dutch individual case safety reports for clinical review with an AUROC of 0.921. For signal prioritisation, Battini et al. [26] used the time of co-exposure and lasso regression in FAERS to prioritise temporally plausible drug–drug interaction signals. For GLP-1RAs specifically, Moon et al. [28] combined disproportionality analysis of 211,195 FAERS cases with XGBoost and SHAP, reaching an AUROC of 0.816 for psychiatric-event reporting, and Kim et al. [27] showed that large language models outperformed fine-tuned transformers in detecting adverse reactions in GLP-1RA posts on social media (F1 0.948 versus 0.729). These studies show that report-level models can discriminate well, but, to our knowledge, none evaluated a model across a product launch, compared it with a simple baseline built from reporting-process fields, or quantified its uncertainty with conformal prediction.

Evidence from clinical prediction shows why these omissions matter. Dataset shift can silently degrade deployed models [30], and calibration can drift while discrimination is maintained [31, 32]. Conformal prediction provides distribution-free coverage guarantees [33], but only under exchangeability; weighted [34], adaptive [35] and class-conditional [36] variants relax this assumption in different ways, yet none has been evaluated on adverse-event reports.

Table 1 Studies published from 2021 to 2026 are compared with the present study.

Study	Problem	Data and method	Main results	Advantages	Disadvantages	Addressed in the present study
Post-marketing safety studies of GLP-1 receptor agonists						
Shu et al. 2022 [11]	Gastrointestinal safety of semaglutide	FAERS 2018–2022, 5442 cases; ROR, clinical-priority scale, time to onset	45 gastrointestinal signals, 17 unexpected; median onset 6–23 days	Signal prioritisation; onset analysis	Single measure (ROR); all indications pooled	Calendar-aligned active comparator; four measures with FDR control
Liu et al. 2022 [12]	Gastrointestinal events across five GLP-1RAs	FAERS 2018–2022, 81,752 reports; ROR by drug	Semaglutide strongest for nausea (ROR 7.41); liraglutide for pancreatitis (ROR 32.67)	Head-to-head class comparison; large sample	Single measure (ROR); no oral obesity products	Oral obesity agents; positive and negative controls
Sodhi et al. 2023 [19]	Gastrointestinal risk of GLP-1RAs used for weight loss	US claims cohort; GLP-1RAs versus bupropion–naltrexone; Cox models	Higher risk of pancreatitis, bowel obstruction and gastroparesis	Incidence-based estimates with an active comparator	Injectables only; claims-based outcome coding	Early signal detection for the new oral agents
Xiong et al. 2024 [13]	Safety of oral semaglutide for type 2 diabetes	FAERS 2019–2023, 2398 reports; ROR, PRR, BCPNN, MGPS	93 signals, including acute cholecystitis and abducens nerve paralysis	Four agreeing measures	Diabetes formulation only; no machine learning	Obesity-dose oral products; 16 sensitivity analyses
Niu et al. 2024 [14]	Route-specific adverse events of semaglutide	FAERS 2017–2023; 2496 oral versus 16,346 subcutaneous records; ROR, stratification	More gastrointestinal events and faster onset with oral route	Direct route comparison; onset analysis	Routes accrued reports over different periods; single measure (ROR)	Comparators restricted to the same calendar period
Zhang et al. 2024 [15]	Oral versus subcutaneous semaglutide reactions	FAERS 2018–2023 plus 422 online patient reviews; ROR	Gastrointestinal events predominant; new route-specific signals	Adds patient-generated data	Review data unverified; single measure (ROR)	FDR control; route-ambiguity audit
Schoretsanitis et al. 2024 [16]	Suicidality with semaglutide and liraglutide	WHO VigiBase; ROR and IC versus all drugs and three active comparators	Suicidal ideation signal for semaglutide (ROR 1.45)	Global data; several comparators	Single outcome; injectables only	Broad prespecified outcome set; psychiatric group monitored
Hathaway et al. 2024 [20]	Optic neuropathy risk with semaglutide	Single-centre matched cohort; Cox models	Hazard ratios 4.28 (diabetes) and 7.64 (obesity)	Incidence-based design with matching	Single centre; few events; injectables only	Ocular events screened for the oral agents
Kim et al. 2025 [17]	Comparative safety of three obesity-approved GLP-1RAs	VigiBase; adjusted ROR and IC	Semaglutide associated with suicidal ideation, hair loss and vision loss	Covariate adjustment; three agents	Injectables only; outcomes limited to label warnings	Covariate-adjusted ROR with calendar-aligned comparators
Arslan 2026 [18]	Semaglutide versus tirzepatide reporting	FAERS 2022–2026, 215,337 reports; ROR, logistic regression, FDR	Higher reporting odds with semaglutide for gastric emptying, NAION and kidney injury	Large contemporary sample; FDR control	Injectables only; no machine learning	Oral agents; signal detection combined with machine learning
Machine learning for pharmacovigilance						
Lieber et al. 2023 [25]	Triage of case reports for clinical review	Dutch spontaneous reports; structured and text features; bagging classifier	AUROC 0.921	Operational triage target; uses free text	Single national centre; depends on free-text fields	Temporal validation across a launch; conformal uncertainty
Battini et al. 2024 [26]	False-positive drug–drug interaction signals	FAERS; time of co-exposure, lasso for confounders	61 of 122 interaction signals prioritised	Temporal plausibility built into signals	Interactions only; positive-control reference set	Negative controls; launch-aligned design
Kim et al. 2025 [27]	Adverse-reaction detection in GLP-1RA social-media posts	396 tweets; fine-tuned transformers versus large language models	F1 0.948 for best language model versus 0.729 for best transformer	Cross-drug transfer; benchmarking	Small test set; social media only	Regulatory reports; simple baselines
Moon et al. 2026 [28]	Case features linked to psychiatric reporting with GLP-1RAs	FAERS 2021–2025, 211,195 cases; disproportionality, XGBoost, SHAP	16 signals; AUROC 0.816	Combines signal detection with explainable machine learning	Explains reporting patterns rather than triaging reports; data precede the oral launches	Process-field baselines; calibration; class-conditional coverage
Present study	Early signals for oral obesity agents; reliability of report-triage models after launch	FAERS 2024–2026, 212,284 reports; four measures with FDR, 16 sensitivity analyses, controls; four classifiers validated on 2026 reports; weighted, adaptive and class-conditional conformal prediction	Gastrointestinal and non-response signals for both agents; AUROC 0.862 versus 0.838 baseline; only class-conditional sets restored serious-report coverage	First post-launch analysis of both oral agents; first triage model tested across a product launch	Spontaneous reports only; no exposure denominator; one database	–

2.4 Research gaps and contributions

Table 1 shows four gaps. First, no study has examined orforglipron or oral semaglutide 25 mg after launch. Second, to our knowledge, no earlier GLP-1RA study combined calendar-aligned active comparators, four agreeing measures with FDR control, and positive and negative controls, or analysed perceived non-response as a prespecified outcome. Third, report-level machine-learning models have not been validated temporally across a product launch or compared with simple baselines. Fourth, no study has assessed calibration and conformal coverage of such models under the covariate shift that a launch produces. The present study addresses all four within a single computational pipeline.

3 Methods

Fig. 1 shows the pipeline. Reports were acquired from FAERS through openFDA and preprocessed by excluding ineligible reports and reports naming more than one incretin product and by removing duplicates (Sect. 3.1). Two analyses followed: active-comparator signal detection on prespecified outcomes (Sect. 3.2) and temporal validation of serious-report classifiers trained on 2024–2025 reports and tested on 2026 reports (Sects. 3.4 and 3.5). Robustness was assessed through sensitivity, control and subgroup analyses (Sects. 3.3 and 3.5).

3.1 Data source, study design and cohort construction

This retrospective pharmacovigilance study used the openFDA drug adverse-event application programming interface, which exposes structured reports from the US Food and Drug Administration Adverse Event Reporting System (FAERS) [37]. Reports received from 1 January 2024 to 30 September 2026 were retrieved for semaglutide, tirzepatide and orforglipron in 66 half-monthly windows; the download log records 269 result pages and agreement between reported and retrieved counts in every window (openFDA data version of 28 October 2026). We used openFDA rather than the quarterly FAERS files because it provides the most recent reports with harmonised substance fields; the consequences of this choice for deduplication are addressed below. For each oral product, the index window was its marketed period: 1 April to 30 September 2026 for orforglipron and 1 January to 30 September 2026 for oral semaglutide 25 mg (Wegovy tablets). Injectable semaglutide and tirzepatide reports received in the same window formed the active comparator. The study follows READUS-PV and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement [9, 38].

Reports were retained if an incretin was coded as the primary suspect drug. Products were classified from substance, brand, route and dosage form into orforglipron, Wegovy tablets, oral semaglutide for type 2 diabetes (Rybelsus), oral semaglutide with unspecified product, injectable semaglutide and tirzepatide. A semaglutide entry was classified as an oral obesity product only with positive evidence of oral administration (oral route code, tablet dosage form or tablet product name) and a Wegovy product name; other semaglutide entries entered the injectable stratum. Because this rule places entries without route or dosage-form information in the comparator, we audited them separately (Sect. 3.3). Reports naming more than one incretin group as primary suspect were excluded. Duplicates were removed first by manufacturer case number, keeping the latest receipt, and then by an informative matching profile (sex, age, weight, country, suspect start date, full preferred-term set, full drug set and product); a looser profile was flagged for sensitivity analysis. Age was converted to years, weights outside 30–350 kg were set to missing, indication text was mapped to obesity, type 2 diabetes, or other or unknown, and reporter qualification to physician, other health professional, consumer, lawyer or unknown. Missing values were retained as explicit categories (Fig. S12 in Online Resource 1).

Analyses used Python 3.13 (pandas, statsmodels, scikit-learn, XGBoost, SHAP and InterpretML) with random seed 20261104. The READUS-PV and TRIPOD+AI checklists [39] (Online Resource 2) and a mapping of reporting topics (Table S21 in Online Resource 1) are provided as Supplementary Information.

3.2 Outcomes and disproportionality analysis

The primary outcomes were 13 prespecified event groups built from Medical Dictionary for Regulatory Activities (MedDRA) preferred-term lists, including gastrointestinal intolerance, pancreatitis, gallbladder and biliary disorders, hepatic injury, acute kidney injury, ocular events and psychiatric symptoms (complete term lists in Table S2 in Online Resource 1). A ten-term composite captured perceived non-response (including "drug ineffective", "weight loss poor" and "weight increased"), and a separate composite captured glycaemic non-response. Secondary outcomes were all individual preferred terms reported at least three times with an index product.

For each index product and outcome, we constructed a 2×2 table of index and comparator reports with and without the outcome and estimated the reporting odds ratio (ROR) and proportional reporting ratio (PRR) with 95% confidence intervals (CIs), the Bayesian information

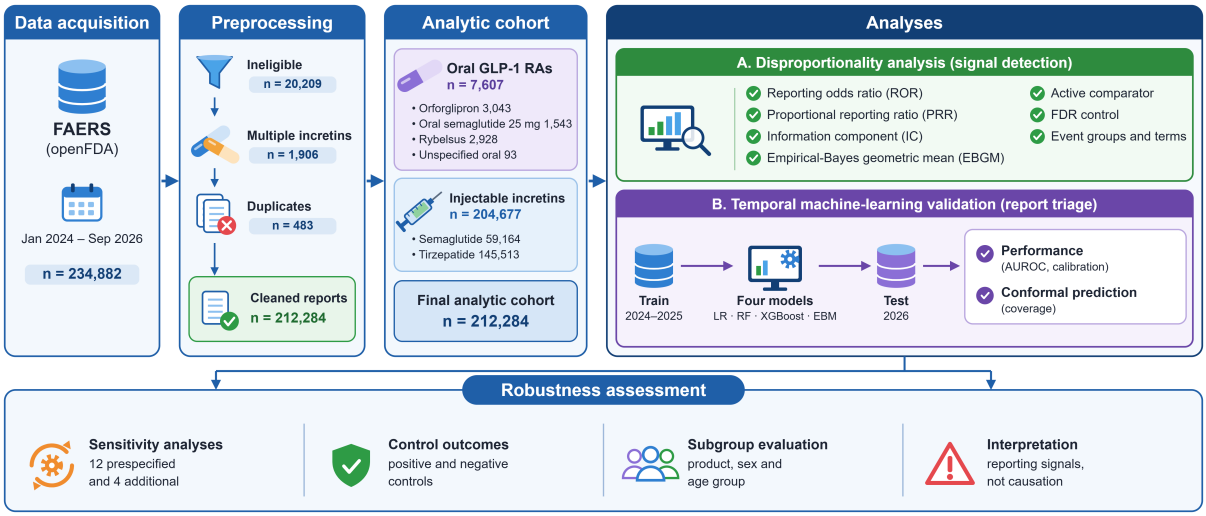


Fig. 1 The computational pharmacovigilance pipeline runs from data acquisition to robustness assessment. EBGM empirical-Bayes geometric mean, EBM explainable boosting machine, FAERS US Food and Drug Administration Adverse Event Reporting System, FDR false-discovery rate, GLP-1 RA glucagon-like peptide-1 receptor agonist, IC information component, LR logistic regression, PRR proportional reporting ratio, RF random forest, ROR reporting odds ratio, XGBoost extreme gradient boosting

component (IC) with its lower credibility bound (IC_{025}) [40], and the empirical-Bayes geometric mean (EBGM) with its lower 5th percentile (EB05) from a two-component gamma–Poisson mixture fitted to all product–term pairs [41]. Let a and b denote index-product reports with and without the outcome, c and d the corresponding comparator reports, and $N = a+b+c+d$. The frequentist measures are

$$\mathrm{ROR} = \frac{a/b}{c/d} = \frac{ad}{bc}, \tag{1}$$

$$\mathrm{PRR} = \frac{a/(a+b)}{c/(c+d)}, \tag{2}$$

with 95% CIs $\exp\{\ln \mathrm{ROR} \pm 1.96\,(a^{-1}+b^{-1}+c^{-1}+d^{-1})^{1/2}\}$ and the analogous log-scale interval for the PRR; 0.5 was added to every cell of a table containing a zero. With expected count $E = (a+b)(a+c)/N$, the information component and its approximate lower 95% credibility bound are

$$\mathrm{IC} = \log_2 \frac{a+0.5}{E+0.5}, \tag{3}$$

$$\mathrm{IC}_{025} \approx \mathrm{IC} - 3.3\,(a+0.5)^{-1/2} - 2\,(a+0.5)^{-3/2}. \tag{4}$$

For the empirical-Bayes measure, each observed count n with expected count E is modelled as $n \mid \lambda \sim \mathrm{Poisson}(\lambda E)$ with prior $\lambda \sim \pi\,\mathrm{Gamma}(\alpha_1, \beta_1) + (1-\pi)\,\mathrm{Gamma}(\alpha_2, \beta_2)$, whose hyperparameters are estimated by maximum marginal likelihood over all product–term pairs; then $\mathrm{EBGM} = \exp\{\mathbb{E}(\ln \lambda \mid n)\}$, and EB05 is the 5th percentile of the posterior distribution of λ. A signal required at least three index reports, a lower ROR bound above 1 and $IC_{025} > 0$ [42]. One-sided Fisher exact p values were adjusted with the Benjamini–Hochberg procedure to control the false-discovery rate (FDR) [43] across all 30 event-group tests (15 outcomes for each product) and, separately, across the preferred-term screen; process and product-use terms (18 for orforglipron and 22 for Wegovy tablets) were excluded from the clinical screen. Labelled status was assigned by matching against current US prescribing information [44, 45] and checked manually for signalling terms.

3.3 Adjusted, sensitivity and control analyses

The covariate-adjusted ROR for perceived non-response was estimated by logistic regression with sex, age band, reporter group, indication, US occurrence and receipt quarter, keeping unknown

values as categories, and was repeated in complete cases. A descriptive head-to-head contrast compared orforglipron with Wegovy-tablet reports from April to September 2026. Twelve prespecified sensitivity analyses (S1–S12) varied reporter source, indication, country, comparator (injectable semaglutide alone; tirzepatide alone; a 2024–2026 window; a launch-phase tirzepatide comparator for the early-marketing effect), masking (gastrointestinal reports removed), deduplication rule, report type (spontaneous only) and batch submissions (days with ≥50 index reports removed). Four further analyses were added after the prespecified set: exclusion from the comparator of semaglutide reports without positive evidence of injection; narrow event definitions approximating narrow-scope Standardised MedDRA Queries for pancreatitis, gallbladder disease, drug-related hepatic disorders and acute renal failure; exclusion of reports with administration- or medication-error terms or co-reported injectable incretins (a proxy for switching); and time from recorded therapy start to report receipt, because openFDA carries no reaction onset date.

Positive controls were labelled reactions expected to signal under a valid design: the gastrointestinal reactions in both labels and the sensory-disturbance cluster (dysaesthesia, allodynia, skin burning sensation, paraesthesia) labelled for Wegovy tablets [45]. Negative controls were six event groups chosen before analysis because no biological pathway links them to incretin therapy: herpes zoster, tendon disorders, ear infection, nail disorders, carpal tunnel syndrome and contact dermatitis. For each, we report the expected count and the probability of detecting a true ROR of 2.

3.4 Machine-learning models

The prediction outcome was a report carrying any regulatory seriousness criterion; seriousness indicates neither causality nor adjudicated severity. The intended use is to order incoming reports for pharmacovigilance review. Twenty-six fields available on the submitted report described age and its absence, weight and its absence, sex, formulation and drug family, indication, reporter type, US occurrence and co-reported medication classes; reported preferred terms were excluded to prevent outcome leakage. Reports received in 2024–2025 formed the development set ($n = 126{,}816$; 28.0% serious) and reports received from January to September 2026 formed the temporal validation set ($n = 85{,}468$; 26.8% serious); the development set contained no reports of orforglipron or Wegovy tablets. All eligible reports were used, and a learning curve assessed the sufficiency of the sample. No resampling or class weighting was applied because the outcome was common. The modelling pipeline is shown in Fig. S13 in Online Resource 1.

Four models spanning the trade-off between interpretability and flexibility were compared: L2-penalised logistic regression (LR), which is linear and additive on the log-odds scale; an explainable boosting machine (EBM), a generalised additive model with a few pairwise interactions whose shape functions can each be plotted [46]; random forest (RF), which averages deep, decorrelated trees [47]; and extreme gradient boosting (XGBoost), which sequentially fits shallow trees to the residual gradient [48]. RF and XGBoost were explained with SHapley Additive exPlanations (SHAP) [49], whereas LR and the EBM are explained by their coefficients and terms. XGBoost, RF and EBM hyperparameters were tuned by stratified cross-validation over 20 random configurations within the development set (Table S19 in Online Resource 1); LR used a fixed penalty. Model specifications are given in the Supplementary Methods of Online Resource 1.

Because seriousness is assigned partly through the regulatory route by which a report arrives, reporting-process fields (US occurrence, reporter type and missing sex, age and weight) may encode how the label is generated rather than case content. We therefore also fitted XGBoost and LR without these seven fields, and seven simple LR baselines on subsets of fields, from US occurrence alone to reporting process plus drug family (nine fields) and patient content alone.

3.5 Model evaluation and uncertainty quantification

All models were evaluated unchanged on the 2026 reports. Discrimination was measured by the area under the receiver operating characteristic curve (AUROC) with DeLong CIs and by the area under the precision–recall curve [50]. Because very small differences reach statistical significance with 85,468 reports, we report each model's AUROC difference from XGBoost with its paired DeLong 95%

CI [51, 52]; Holm-adjusted p values for all 55 pairwise comparisons [53] are in Table S20 in Online Resource 1. Calibration was assessed by the Brier score and by the recalibration model $\text{logit}\,\Pr(Y=1) = a + b\,\text{logit}\,\hat{p}(\mathbf{x})$, in which the calibration slope is b and calibration-in-the-large (CITL) is a with b fixed at 1. Decision curves report net benefit for hypothetical triage thresholds [54]. Discrimination and calibration were audited by sex and age group and for the two products absent from development.

For uncertainty, we used the non-conformity score $s(\mathbf{x}, y) = 1 - \hat{p}_y(\mathbf{x})$, where $\hat{p}_y$ is the predicted probability of class y [33]. Split conformal prediction computes, on a held-out 20% of development reports (n_c = 25,364), the empirical quantile $\hat{q}$ of the calibration scores at level $\lceil (n_c+1)(1-\alpha) \rceil / n_c$ and returns the prediction set

$$\mathcal{C}(\mathbf{x}) = \{y \in \{0,1\} : s(\mathbf{x}, y) \leq \hat{q}\}, \tag{5}$$

which satisfies

$$\Pr\{Y \in \mathcal{C}(\mathbf{X})\} \geq 1 - \alpha, \tag{6}$$

where the coverage guarantee holds under exchangeability; we used $\alpha = 0.1$. Class-conditional (Mondrian) sets compute a separate quantile $\hat{q}_k$ from calibration reports of class k, targeting coverage separately for serious and non-serious reports [36]. Because product launches violate exchangeability, we added two shift-aware methods. Weighted conformal prediction [34] replaces $\hat{q}$ by a weighted quantile in which calibration report i receives weight $w(\mathbf{x}_i)$ and the test report a point mass $w(\mathbf{x})$ at $+\infty$, with $w(\mathbf{x}) = \{\hat{\pi}(\mathbf{x})/(1-\hat{\pi}(\mathbf{x}))\}(n_c/n_v)$ estimated by a classifier $\hat{\pi}$ distinguishing 2026 reports (n_v) from calibration reports. Adaptive conformal inference [35] processes 2026 reports in order of receipt and updates the working miscoverage level after each report,

$$\alpha_{t+1} = \alpha_t + \gamma\,(\alpha - \text{err}_t), \tag{7}$$

with

$$\text{err}_t = \mathbf{1}\{y_t \notin \mathcal{C}_t(\mathbf{x}_t)\}, \tag{8}$$

with step size $\gamma = 0.005$, in both marginal and class-conditional versions. Logistic recalibration was fitted on reports from January–March 2026 and evaluated on those from April–September 2026, and a learning curve retrained models on 10%, 25% and 50% of the development data. Prediction modelling is reported with reference to TRIPOD+AI [39].

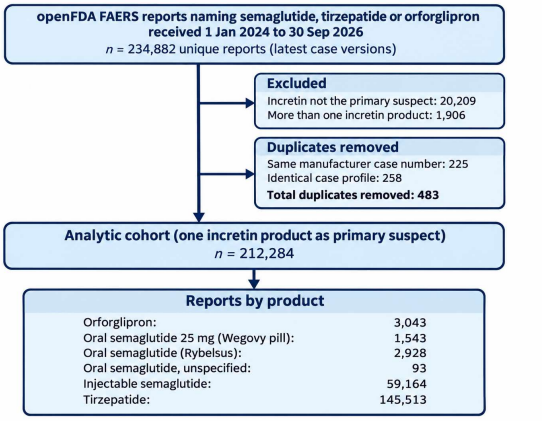


Fig. 2 Three exclusion steps produced the analytic cohort. FAERS US Food and Drug Administration Adverse Event Reporting System

4 Results

We first describe the analytic cohort (Sect. 4.1), then the disproportionality signals for event groups and preferred terms (Sect. 4.2) and the robustness of the perceived non-response signal (Sect. 4.3), and finally the temporal validation of the serious-report classifiers, including calibration and conformal coverage (Sect. 4.4).

4.1 Study population

Of 234,882 downloaded records, 20,209 had no incretin as primary suspect, 1906 named more than one incretin group and 483 were duplicates, leaving 212,284 reports (Fig. 2; Table 2). Orforglipron (3043 reports) and Wegovy tablets (1543) were compared with 58,789 and 80,035 comparator reports, respectively. The reporting mix differed between products: women accounted for 73.7% and 70.5% of oral-product reports, consumers for 81.4% and 79.6%, and 98.9% of orforglipron reports occurred in the USA. Age was missing in 79.8% and 50.2% of oral-product reports.

Table 2 Report characteristics differed between products.

Characteristic	Orforglipron	Wegovy tablet	Rybelsus	Inj. semaglutide	Tirzepatide
Reports, n	3,043	1,543	2,928	59,164	145,513
Women, %	73.7	70.5	54.1	60.5	60.7
Age missing, %	79.8	50.2	38.7	47.6	32.7
Median age, years[a]	63	58	63	60	57
Obesity indication, %	83.0	52.6	11.4	29.2	33.8
Consumer reporter, %	81.4	79.6	56.7	75.1	92.2
US occurrence, %	98.9	91.8	51.6	78.8	85.3
Serious, %	4.5	16.1	66.1	50.7	17.9
Non-response term, %	29.7	31.8	5.9	9.4	7.1

4.2 Signals for event groups and preferred terms

Orforglipron reports were disproportionately associated with gastrointestinal intolerance (1477 reports; ROR 2.41, 95% CI 2.24–2.59), dizziness or headache (290; 1.95, 1.72–2.21) and fatigue or asthenia (311; 1.52, 1.34–1.71). Wegovy-tablet reports signalled for gastrointestinal intolerance (543; 1.41, 1.27–1.57), hypersensitivity (96; 1.98, 1.61–2.44) and dizziness or headache (117; 1.54, 1.27–1.87) (Table 3; Fig. 3). All eight signals, including perceived non-response, had EB05 above 1 and survived FDR control across the 30 event-group tests. No signal was observed for pancreatitis, gallbladder and biliary disorders, hepatic injury, acute kidney injury, ocular events or delayed gastric emptying, and none emerged under narrow definitions, although counts were small (1–11 index reports per narrow group; Table S18 in Online Resource 1). For psychiatric symptoms with Wegovy tablets, the ROR was 1.33 (1.03–1.72) but IC_{025} was not above zero and the FDR-adjusted q was 0.06. RORs below one should not be read as protective given short marketing time and sparse counts.

Of 180 preferred terms screened for orforglipron, 24 met the signal rule, 20 of them clinical; 19 of these passed FDR control and 13 were labelled. For Wegovy tablets, 50 of 193 terms met the rule, 42 were clinical and all passed FDR control (Table S3 in Online Resource 1). The term "nightmare", absent from both labels, signalled for orforglipron (8 reports; ROR 6.45, 2.90–14.38; IC_{025} 0.82; EB05 1.25; $q = 0.001$) and Wegovy tablets (4 reports; 5.47, 1.95–15.35).

4.3 Perceived non-response, robustness and controls

A non-response term appeared in 905 of 3043 orforglipron reports (29.7%) and 491 of 1543 Wegovy-tablet reports (31.8%). Adjusted RORs were 2.89 (2.65–3.16) and 3.78 (3.37–4.24) with unknown values as categories, and 3.30 (2.72–4.02) and 2.55 (2.14–3.04) in complete cases (Table S13 in Online Resource 1). The signal persisted in all 13 analytic scenarios for both products (Fig. 4): against injectable semaglutide alone (2.95, 2.70–3.22; and 3.60, 3.21–4.03), against tirzepatide alone (3.78, 3.47–4.11; and 4.16, 3.73–4.65), in spontaneous reports only (3.44, 3.17–3.73; and 4.04, 3.61–4.51) and in health-professional reports only (orforglipron 2.21, 1.56–3.15). Administration- or medication-error terms were less common in oral-product than comparator reports (6.6% and 11.9% versus 22.4% and 22.1%), co-reported injectable incretins were rare (0.2%) and no compounded product was named; excluding these reports left the non-response ROR at 3.27 (3.01–3.56) and 3.52 (3.13–3.96) (Table S15 in Online Resource 1). The gastrointestinal signal persisted in 12 of 12 applicable scenarios for orforglipron and 11 of 12 for Wegovy tablets, and every signal in Table 3 persisted against each single comparator (Table S11 in Online Resource 1). In the head-to-head contrast, gastrointestinal intolerance was reported more often with orforglipron than with Wegovy tablets (ROR 1.73, 1.52–1.96), hypersensitivity less often (0.52, 0.39–0.70), and non-response did not differ clearly (0.88, 0.77–1.01).

Of 59,164 injectable-semaglutide reports, 6688 (11.3%) lacked positive evidence of injection, including 4659 Wegovy-named reports received in 2026. Their profile resembled confirmed injectable rather than tablet reports (perceived non-response 13.5% versus 10.5% and 31.8%; serious 49.1%

Table 3 Eight event-group signals were detected for the oral agents.

Event group	Index n	Comp. n	ROR (95% CI)	PRR	IC_{025}	EB05	q
Orforglipron (index $n = 3043$; comparator $n = 58{,}789$)							
Gastrointestinal intolerance	1477	16,545	2.41 (2.24–2.59)	1.72	0.65	1.59	<0.001
Dizziness/headache	290	3017	1.95 (1.72–2.21)	1.86	0.64	1.59	<0.001
Fatigue/asthenia	311	4102	1.52 (1.34–1.71)	1.46	0.33	1.30	<0.001
Perceived non-response	905	6464	3.43 (3.16–3.72)	2.70	1.21	2.34	<0.001
Wegovy tablet (index $n = 1543$; comparator $n = 80{,}035$)							
Gastrointestinal intolerance	543	22,208	1.41 (1.27–1.57)	1.27	0.19	1.17	<0.001
Hypersensitivity	96	2594	1.98 (1.61–2.44)	1.92	0.57	1.53	<0.001
Dizziness/headache	117	4045	1.54 (1.27–1.87)	1.50	0.26	1.26	<0.001
Perceived non-response	491	8418	3.97 (3.56–4.43)	3.03	1.39	2.68	<0.001

versus 44.8% and 16.1%), and excluding them from the comparator left every signal in place (orforglipron non-response 3.59, 3.31–3.90; Wegovy tablets 4.12, 3.69–4.60; Table S17 in Online Resource 1). In that analysis the borderline psychiatric estimate for Wegovy tablets crossed the signal threshold (1.38, 1.07–1.78). Among reports with a therapy start date (8.7% for orforglipron and 41.9% for Wegovy tablets), the median time from therapy start to report receipt was 26 and 53 days (Table S16 in Online Resource 1).

Labelled gastrointestinal reactions signalled for orforglipron (nausea ROR 2.06, 1.88–2.24; constipation 2.03, 1.80–2.28), and the labelled sensory-disturbance cluster signalled strongly for Wegovy tablets (dysaesthesia 30 reports, ROR 17.8, 11.7–27.0). No negative control signalled, but expected index counts were 0.2–2.8, so the probability of detecting a true ROR of 2 exceeded 0.5 only for herpes zoster with both products and tendon disorders with orforglipron (Table S14 in Online Resource 1).

4.4 Performance of the machine-learning models

XGBoost achieved an AUROC of 0.862 (0.859–0.865) on 2026 reports, followed by tuned RF (0.861), tuned EBM (0.855) and penalised LR (0.845) (Table 4); all differences from XGBoost were below 0.02. A nine-field LR model on reporting-process and drug-family fields reached 0.838 (difference −0.024, −0.026 to −0.022), whereas patient content alone reached 0.572. Removing the seven reporting-process fields reduced XGBoost to 0.733 (0.729–0.737) and orforglipron AUROC from 0.742 to 0.661. The EBM's leading terms (US occurrence, dual agonist, consumer reporter and an interaction between unknown sex and US occurrence) matched the top SHAP features of XGBoost, whose ranking was stable across refits (mean Spearman correlation 0.987). Curves are shown in Fig. 5, and EBM explanations of individual orforglipron reports in Fig. S15 in Online Resource 1.

All models overestimated the risk of a serious report: mean predicted risk was 0.298–0.316 against an observed 0.268 (CITL −0.27 to −0.42), and overestimation was severe for orforglipron, where serious reports made up 4.5% (CITL −2.31 to −2.53; Table S12 in Online Resource 1). For orforglipron, every model's accuracy at a 0.5 threshold (0.914–0.938) was below the 0.955 achieved by predicting every report non-serious. XGBoost AUROC was 0.742 (0.693–0.791) for orforglipron and 0.777 (0.740–0.809) for Wegovy tablets, and lower for reports of people aged 65 years or older (0.835) or of unknown age (0.815) than for those aged 45–64 years (0.903) (Fig. S11 in Online Resource 1). XGBoost trained on 10% of the development data came within 0.007 AUROC of the full model (Fig. S14 in Online Resource 1).

Split conformal sets covered 88.3% of 2026 reports but only 67.8% of serious reports, and only 41.3% of serious orforglipron reports (Table 5). A classifier separated conformal-calibration reports from 2026 reports with an AUROC of 0.80, confirming substantial covariate shift. Weighted conformal prediction raised serious-report coverage to 72.1% overall and 67.4% for orforglipron, and adaptive conformal inference restored overall coverage to the 90% target (90.0%) but not coverage of serious reports (74.4%). Only the class-conditional variants approached or reached nominal coverage of serious reports (89.5% unweighted, 88.4%

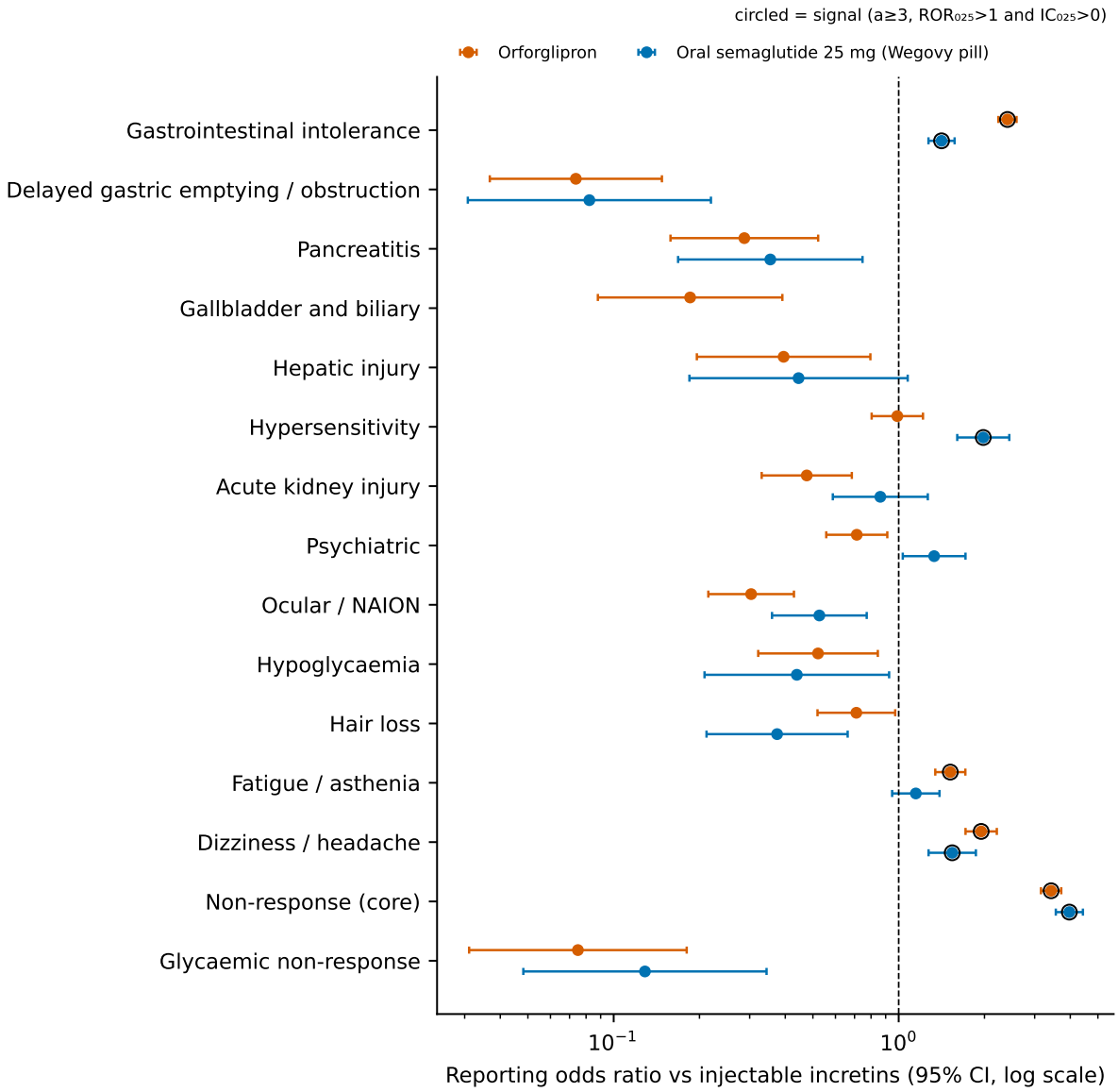


Fig. 3 Reporting odds ratios are shown for each event group. CI confidence interval, IC_{025} lower 95% credibility bound of the information component, NAION nonarteritic anterior ischaemic optic neuropathy, ROR reporting odds ratio

weighted and 90.0% adaptive), at the cost of larger prediction sets (mean 1.28–1.33 labels). Recalibrating XGBoost on January–March 2026 reports worsened April–September calibration (Brier score 0.122 to 0.142), because the share of serious reports fell from 29.0% to 24.5% and then rose to 27.8%.

5 Discussion

During the first six to nine months after launch, gastrointestinal intolerance and perceived non-response were the dominant disproportionate reporting patterns for orforglipron and oral semaglutide 25 mg relative to injectable incretins reported over the same periods. These signals agreed across four measures, survived FDR control, persisted against each comparator separately

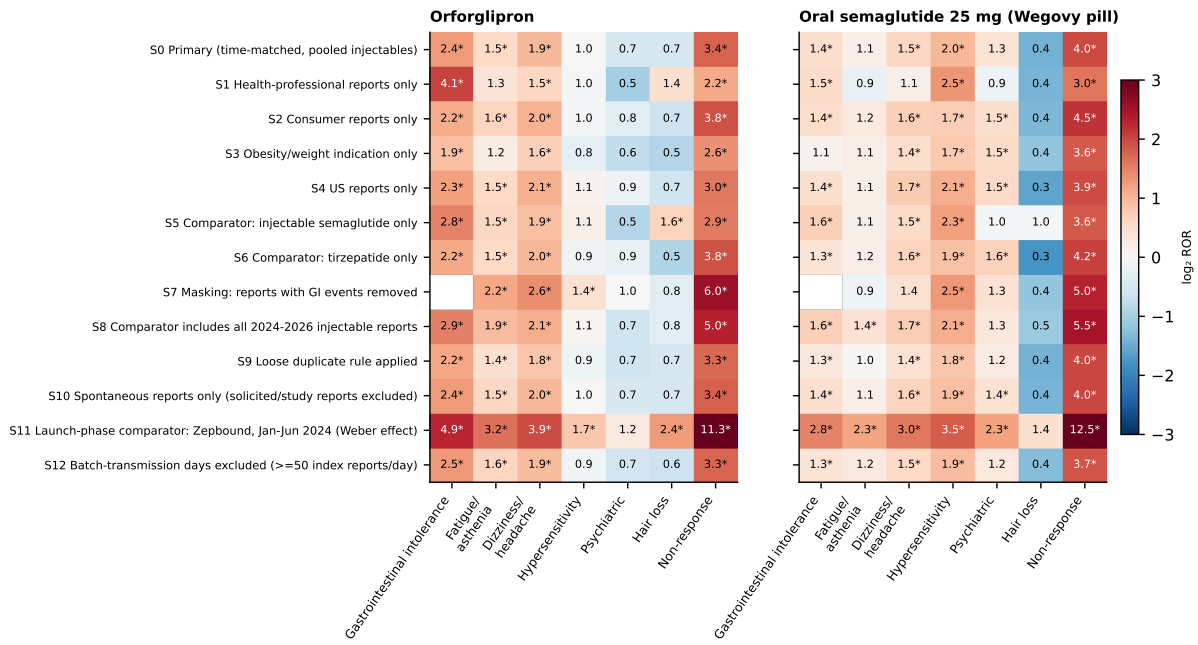


Fig. 4 Signals persisted across sensitivity analyses. Asterisks mark signals. GI gastrointestinal, ROR reporting odds ratio

Table 4 All four models discriminated well on 2026 reports, but most of their discrimination came from reporting-process fields.

Model	Fields	AUROC (95% CI)	Δ (95% CI)	Orforglipron
XGBoost	26	0.862 (0.859–0.865)	–	0.742 (0.695–0.790)
RF (tuned)	26	0.861 (0.858–0.864)	−0.002 (−0.002 to −0.001)	0.747 (0.699–0.795)
EBM (tuned)	26	0.855 (0.852–0.858)	−0.007 (−0.008 to −0.006)	0.737 (0.687–0.786)
Penalised LR	26	0.845 (0.842–0.849)	−0.017 (−0.018 to −0.015)	0.709 (0.655–0.763)
XGBoost without process fields	19	0.733 (0.729–0.737)	−0.129 (−0.132 to −0.126)	0.661 (0.609–0.712)
LR without process fields	19	0.661 (0.657–0.665)	−0.201 (−0.205 to −0.197)	0.585 (0.530–0.641)
Process + drug family (LR)	9	0.838 (0.835–0.842)	−0.024 (−0.026 to −0.022)	0.735 (0.688–0.782)
Patient content only (LR)	16	0.572 (0.567–0.576)	−0.290 (−0.296 to −0.285)	0.556 (0.503–0.609)

Table 5 Class-conditional conformal sets restored coverage of serious reports.

Method	Overall	Serious	Orfo. serious	Wegovy serious	Size
Split conformal	88.3%	67.8%	41.3%	56.5%	1.05
Weighted conformal	90.1%	72.1%	67.4%	65.3%	1.11
Adaptive conformal inference	90.0%	74.4%	64.5%	64.1%	1.12
Class-conditional (Mondrian)	86.1%	89.5%	100.0%	100.0%	1.28
Weighted class-conditional	88.3%	88.4%	100.0%	100.0%	1.28
Adaptive class-conditional	90.0%	90.0%	97.8%	97.2%	1.33

and in spontaneous reports only, and were not explained by misclassified route, administration errors or switching. No pancreatic, biliary, hepatic, renal or ocular signal met the prespecified criteria.

A classifier developed before either launch discriminated serious reports well overall, but most of its signal came from how and by whom reports were filed, and its calibration and uncertainty coverage deteriorated for the new products.

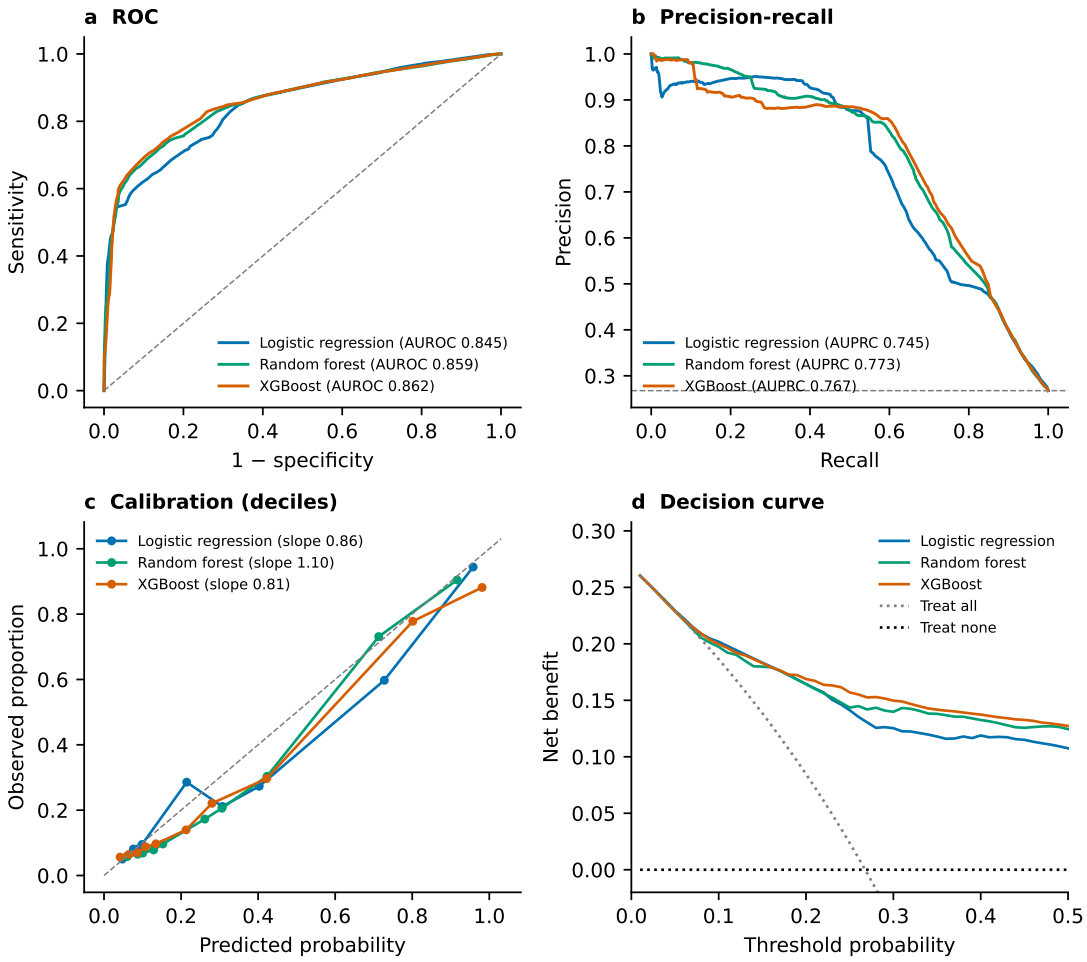


Fig. 5 Discrimination, calibration and net benefit are shown for 2026 reports. AUPRC area under the precision–recall curve, AUROC area under the receiver operating characteristic curve, ROC receiver operating characteristic

5.1 Clinical interpretation of the safety signals

The gastrointestinal signals agree with the trials and labels, in which nausea, vomiting and constipation were the most frequent adverse events of both agents [5, 7, 44, 45]. A disproportion relative to injectables may reflect daily oral dosing, titration, a treatment-naive population or stimulated reporting in the first year after launch [22]; FAERS cannot separate these. The short interval between therapy start and report receipt (median 26 days for orforglipron) is consistent with events during initiation and titration. The dizziness and fatigue signals may reflect reduced food and fluid intake, and the hypersensitivity signal for Wegovy tablets, with its lower relative share for orforglipron in the head-to-head contrast, may reflect formulation differences but needs confirmation.

Perceived non-response was the most robust finding, but it rests on a reporting-based denominator, and reports lack baseline weight, adherence, titration and duration. Oral semaglutide must be taken fasting with limited water and a waiting period before food, whereas orforglipron has no such restriction [44, 45]; incorrect administration is therefore a plausible contributor for Wegovy tablets. Administration-error terms were, however, less frequent in oral-product than comparator reports, and excluding them changed the estimates little. Patients who switch from injectables to an oral product at a lower equivalent exposure could also perceive loss of effect, but co-reported injectables were rare. The signal therefore

cannot identify clinical non-responders; its persistence in health-professional reports makes a purely consumer-driven explanation unlikely.

For clinical practice, these findings support counselling at the start of oral GLP-1RA therapy about gastrointestinal effects and their management during titration, the administration requirements specific to oral semaglutide, and realistic expectations of the time course of weight response. The nightmare signal, based on 8 and 4 reports, should be treated as a case-review priority rather than an adverse reaction; central effects of GLP-1 receptor agonism are biologically conceivable, but narratives, time to onset and dechallenge information are needed. The borderline psychiatric estimate for Wegovy tablets, which crossed the threshold only when route-ambiguous comparator reports were excluded, warrants monitoring in the context of earlier semaglutide signals [16].

5.2 Reliability of machine-learning models for report triage

Three observations follow from the benchmark. First, model choice mattered little: the glass-box EBM was within 0.007 AUROC of XGBoost, so interpretability came at almost no cost. Second, most of the predictive signal came from the reporting process: a nine-field model reached 0.838, removing these fields reduced XGBoost to 0.733, and patient content alone reached 0.572. Because seriousness in FAERS partly reflects the reporting route, these fields may encode how the label is generated rather than case severity; a report-triage model should be judged against such a baseline and against established screening rules, not against chance. Third, the limiting factor was the change in the reporting stream rather than sample size: performance saturated at 10% of the development data, and the pre-launch and post-launch reports were readily distinguishable.

The uncertainty analysis shows why overall discrimination is an insufficient criterion for deployment. Marginal conformal coverage concealed poor coverage of serious reports; weighted and adaptive methods designed for shift improved overall coverage but did not restore it for serious reports, which only class-conditional sets achieved [34–36]. A single recalibration chased a prevalence that had already moved, consistent with calibration drift reported in clinical models [31]. For computational pharmacovigilance, this implies that report-level models should be monitored with class-conditional coverage and product-specific calibration, not only with AUROC. Such models classify submitted reports, not patient risk; any operational use would require a defined review target, verification that each field is available at the point of triage, continuous monitoring and product-specific recalibration.

5.3 Comparison with previous studies

Earlier FAERS and VigiBase studies of incretins used other products, periods and comparators [11–15, 17, 18] (Table 1). The gastrointestinal predominance we observed is consistent with these studies and with the trials, but differences in comparator construction, windows and term definitions mean that RORs should not be pooled across studies. To our knowledge, no earlier report-level model [25, 28, 29] has been evaluated across a product launch, against simple baselines or with conformal uncertainty.

5.4 Strengths and limitations

Strengths include an audited, complete download; launch-aligned active comparators; primary-suspect restriction and two deduplication rules; four agreeing disproportionality measures with FDR control; 12 prespecified and 4 additional sensitivity analyses; positive and negative controls; and a temporally validated benchmark with tuned models, simple baselines, paired CIs and shift-aware uncertainty methods.

The study has several limitations. First, FAERS reports are selected, incompletely verified and lack exposure denominators, so neither incidence nor causality can be estimated [9, 55], and stimulated reporting is likely early after launch [22, 23]. Second, both oral agents had been marketed for less than a year, limiting detection of rare or delayed events and the power of the negative controls. Third, most oral-product reports were consumer reports transmitted by manufacturers, many with missing age and weight, and the pooled comparator spans two pharmacological classes, so residual confounding by indication and reporting channel is likely. Fourth, openFDA

provides no reaction onset date or narrative, precluding time-to-onset and case-level causality assessment; the deduplication was not validated against the quarterly FAERS files, so residual duplicates are possible. Fifth, the event groups are investigator-defined; the narrow definitions approximate but do not reproduce licensed Standardised MedDRA Queries. Sixth, all findings rest on one database; replication in VigiBase or EudraVigilance is needed. Seventh, with 85,468 validation reports, very small AUROC differences are statistically significant, and the DeLong CIs treat reports as independent. Finally, the classifier was validated in one data source, its prediction target is an administrative flag, and it has not been evaluated prospectively in a review workflow or against tabular foundation models [56].

6 Conclusion

Spontaneous reports for orforglipron and oral semaglutide 25 mg in their first months on the market were dominated by gastrointestinal intolerance and perceived non-response relative to contemporaneous injectable incretins, with consistent results across measures, comparators and sensitivity analyses. No organ-specific signal was detected, although this cannot exclude rare or delayed harms. These findings describe reporting patterns, not incidence or causality; they support counselling on gastrointestinal tolerability, administration and expected response at initiation, and case review of the nightmare reports. Machine-learning models that triage serious reports can discriminate well overall while relying mainly on reporting-process fields, and they lose calibration and uncertainty coverage when new products enter the reporting stream; they should be benchmarked against simple baselines, monitored with class-conditional coverage and recalibrated for each new product. Replication in independent databases and denominator-based studies of incidence are the next steps.

Supplementary Information. The supplementary material for this article is provided as Electronic Supplementary Material with the online version and is openly available at Zenodo: https://doi.org/10.5281/zenodo.24095237. Online Resource 1 (ESM_1.pdf) contains the Supplementary Methods, Tables S1–S22 and Figs. S1–S15. Online Resource 2 (ESM_2.pdf) contains the completed READUS-PV and TRIPOD+AI reporting checklists.

Statements and Declarations

Funding No specific funding was received for this research.

Competing interests The author has no relevant financial or non-financial interests to disclose.

Ethics approval Not applicable. This study is a secondary analysis of publicly available, de-identified adverse-event reports from FAERS. No participants were recruited or contacted, and no identifiable personal information was accessed.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material FAERS reports are publicly available through the openFDA drug adverse-event interface at https://open.fda.gov/apis/drug/event/. Derived analysis tables and numerical outputs are available from the author upon reasonable request.

Code availability The analysis code, term definitions and figure-generation scripts are available from the author upon reasonable request.

Protocol and registration No study protocol was registered. The prespecified outcomes, term lists, signal criteria and analyses are documented in the analysis code.

Clinical trial registration Not applicable.

Patient and public involvement Patients and the public were not involved in the design, conduct, reporting or dissemination of this research.

Author contributions T.A.: conceptualisation, methodology, data acquisition and curation, software, formal analysis, investigation, validation, visualisation, and writing (original draft, review and editing). The author read and approved the final manuscript and is accountable for all aspects of the work.

Use of generative artificial intelligence A large language model assisted with code drafting and language editing. The author reviewed and verified all code, analyses, results and text and takes full responsibility for the content.

References

[1] KFF. Kff health tracking poll: Prescription drug costs, views on trump administration actions, and glp-1 use, November 2025. URL https://www.kff.org/public-opinion/kff-health-tracking-poll-prescription-drug-costs-views-on-trump-administration-actions-and-glp-1-use/. Accessed 5 November 2026.

[2] John P. H. Wilding, Rachel L. Batterham, Salvatore Calanna, et al. Once-weekly semaglutide in adults with overweight or obesity. *New England Journal of Medicine*, 384:989–1002, 2021. https://doi.org/10.1056/NEJMoa2032183.

[3] Ania M. Jastreboff, Louis J. Aronne, Nadia N. Ahmad, et al. Tirzepatide once weekly for the treatment of obesity. *New England Journal of Medicine*, 387:205–216, 2022. https://doi.org/10.1056/NEJMoa2206038.

[4] A. Michael Lincoff, Kirstine Brown-Frandsen, Helen M. Colhoun, et al. Semaglutide and cardiovascular outcomes in obesity without diabetes. *New England Journal of Medicine*, 389:2221–2232, 2023. https://doi.org/10.1056/NEJMoa2307563.

[5] Sean Wharton, Ildiko Lingvay, Pawel Bogdanski, et al. Oral semaglutide at a dose of 25 mg in adults with overweight or obesity. *New England Journal of Medicine*, 393:1077–1087, 2025. https://doi.org/10.1056/NEJMoa2500969.

[6] U.S. Food and Drug Administration. WEGOVY (semaglutide) tablets, for oral use: Prescribing information, 2025. URL https://www.accessdata.fda.gov/drugsatfda_docs/label/2025/218316Orig1s000lbl.pdf. Accessed 5 November 2026.

[7] Sean Wharton, Louis J. Aronne, Adam Stefanski, et al. Orforglipron, an oral small-molecule GLP-1 receptor agonist for obesity treatment. *New England Journal of Medicine*, 393:1796–1806, 2025. https://doi.org/10.1056/NEJMoa2511774.

[8] U.S. Food and Drug Administration. FDA approves first new molecular entity under national priority voucher program, 2026. URL https://www.fda.gov/news-events/press-announcements/fda-approves-first-new-molecular-entity-under-national-priority-voucher-program. Press release, 1 April 2026; accessed 5 November 2026.

[9] Michele Fusaroli, Francesco Salvo, Bernard Begaud, et al. The reporting of a disproportionality analysis for drug safety signal detection using individual case safety reports in pharmacovigilance (READUS-PV): Development and statement. *Drug Safety*, 47:575–584, 2024. https://doi.org/10.1007/s40264-024-01421-9.

[10] Andrew Bate and Steve F. Hobbiger. Artificial intelligence, real-world automation and the safety of medicines. *Drug Safety*, 44(2):125–132, 2021. https://doi.org/10.1007/s40264-020-01001-7.

[11] Yamin Shu, Xucheng He, Pan Wu, Yanxin Liu, Yufeng Ding, and Qilin Zhang. Gastrointestinal adverse events associated with semaglutide: A pharmacovigilance study based on FDA adverse event reporting system. *Frontiers in Public Health*, 10:996179, 2022. https://doi.org/10.3389/fpubh.2022.996179.

[12] Lulu Liu, Jia Chen, Lei Wang, Chen Chen, and Li Chen. Association between different GLP-1 receptor agonists and gastrointestinal adverse reactions: A real-world disproportionality study based on FDA adverse event reporting system database. *Frontiers in Endocrinology*, 13:1043789, 2022. https://doi.org/10.3389/fendo.2022.1043789.

[13] Si Xiong, Ruoyu Gou, Xudong Liang, Hao Wu, Shuitao Qin, Bing Li, Changjun Luo, and Junan Chen. Adverse events of oral GLP-1 receptor agonist (semaglutide tablets): A real-world study based on FAERS from 2019 to 2023. *Diabetes Therapy*, 15(8):1717–1733, 2024. https://doi.org/10.1007/s13300-024-01594-7.

[14] Kaibin Niu, Maoxia Fan, Wulin Gao, Chen Chen, and Guohua Dai. Adverse events in different administration routes of semaglutide: a pharmacovigilance study based on the FDA adverse event reporting system. *Frontiers in Pharmacology*, 15:1414268, 2024. https://doi.org/10.3389/fphar.2024.1414268.

[15] Jing Zhang, Xiaofen Wang, and Yiting Zhou. Comparative analysis of semaglutide induced adverse reactions: Insights from FAERS database and social media reviews with a focus on oral vs subcutaneous administration. *Frontiers in Pharmacology*, 15:1471615, 2024.

https://doi.org/10.3389/fphar.2024.1471615.

[16] Georgios Schoretsanitis, Stefan Weiler, Corrado Barbui, Emanuel Raschi, and Chiara Gastaldon. Disproportionality analysis from World Health Organization data on semaglutide, liraglutide, and suicidality. *JAMA Network Open*, 7(8):e2423385, 2024. https://doi.org/10.1001/jamanetworkopen.2024.23385.

[17] Tae Hyeon Kim, Kyeongmin Lee, Seoyoung Park, et al. Adverse drug reaction patterns of GLP-1 receptor agonists approved for obesity treatment: Disproportionality analysis from global pharmacovigilance database. *Diabetes, Obesity and Metabolism*, 27(6):3490–3502, 2025. https://doi.org/10.1111/dom.16376.

[18] Turgay Arslan. Comparative adverse-event reporting patterns of semaglutide and tirzepatide: a disproportionality analysis of FAERS. *Endocrine*, 91(1):357, 2026. https://doi.org/10.1007/s12020-026-04840-z.

[19] Mohit Sodhi, Ramin Rezaeianzadeh, Abbas Kezouh, and Mahyar Etminan. Risk of gastrointestinal adverse events associated with glucagon-like peptide-1 receptor agonists for weight loss. *JAMA*, 330(18):1795–1797, 2023. https://doi.org/10.1001/jama.2023.19574.

[20] Jimena Tatiana Hathaway, Madhura P. Shah, David B. Hathaway, et al. Risk of nonarteritic anterior ischemic optic neuropathy in patients prescribed semaglutide. *JAMA Ophthalmology*, 142(8):732–739, 2024. https://doi.org/10.1001/jamaophthalmol.2024.2296.

[21] Christopher A. Gravel and Antonios Douros. Considerations on the use of different comparators in pharmacovigilance: A methodological review. *British Journal of Clinical Pharmacology*, 89(9):2671–2676, 2023. https://doi.org/10.1111/bcp.15802.

[22] Nicole R. Hartnell and James P. Wilson. Replication of the Weber effect using postmarketing adverse event reports voluntarily submitted to the United States Food and Drug Administration. *Pharmacotherapy*, 24(6):743–749, 2004. https://doi.org/10.1592/phco.24.8.743.36068.

[23] Antoine Pariente, Fleur Gregoire, Annie Fourrier-Reglat, Françoise Haramburu, and Nicholas Moore. Impact of safety alerts on measures of disproportionality in spontaneous reporting databases: the notoriety bias. *Drug Safety*, 30(10):891–898, 2007. https://doi.org/10.2165/00002018-200730100-00007.

[24] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010. https://doi.org/10.1097/EDE.0b013e3181d61eeb.

[25] Thomas Lieber, Helen R. Gosselt, Pelle C. Kools, Okko C. Kruijssen, Stijn N. C. Van Lierop, Linda Härmark, and Florence P. A. M. Van Hunsel. Natural language processing for automated triage and prioritization of individual case safety reports for case-by-case assessment. *Frontiers in Drug Safety and Regulation*, 3:1120135, 2023. https://doi.org/10.3389/fdsfr.2023.1120135.

[26] Vera Battini, Marianna Cocco, Maria Antonietta Barbieri, Greg Powell, Carla Carnovale, Emilio Clementi, Andrew Bate, and Maurizio Sessa. Timing matters: A machine learning method for the prioritization of drug-drug interactions through signal detection in the FDA adverse event reporting system and their relationship with time of co-exposure. *Drug Safety*, 47(9):895–907, 2024. https://doi.org/10.1007/s40264-024-01430-8.

[27] Minjung Kim, Kyoung Eun Kim, Jae-Hee Kwon, Ja-Young Han, Jae Hyun Kim, and Myeong Gyu Kim. Transformer-based models for ADR detection: cross-drug validation and benchmarking against large language models. *Therapeutic Advances in Drug Safety*, 16:20420986251405082, 2025. https://doi.org/10.1177/20420986251405082.

[28] Suhyeon Moon, EunJu Lee, Doyeon Kim, Kyung Hee Choi, Yeo Jin Choi, and Sooyoung Shin. Psychiatric safety signals of GLP-1 receptor agonists: A FAERS-based pharmacovigilance study with explainable machine learning. *Pharmaceuticals*, 19(6):953, 2026. https://doi.org/10.3390/ph19060953.

[29] Ramani Routray, Niki Tetarenko, Claire Abu-Assal, et al. Application of augmented intelligence for pharmacovigilance case seriousness determination. *Drug Safety*, 43(1):57–66, 2020. https://doi.org/10.1007/s40264-019-00869-4.

[30] Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, et al. The clinician and dataset shift in artificial intelligence. *New*

England Journal of Medicine, 385(3):283–286, 2021. https://doi.org/10.1056/NEJMc2104626.

[31] Sharon E. Davis, Thomas A. Lasko, Guanhua Chen, Edward D. Siew, and Michael E. Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*, 24(6):1052–1061, 2017. https://doi.org/10.1093/jamia/ocx030.

[32] Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019. https://doi.org/10.1186/s12916-019-1466-7.

[33] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: a gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. https://doi.org/10.1561/2200000101.

[34] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 2530–2540, 2019. URL https://mlanthology.org/neurips/2019/tibshirani2019neurips-conformal/.

[35] Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html.

[36] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. https://doi.org/10.1080/01621459.2017.1395341.

[37] U.S. Food and Drug Administration. openFDA: Drug adverse events api, 2026. URL https://open.fda.gov/apis/drug/event/. Accessed 28 October 2026.

[38] Erik von Elm, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and Jan P. Vandenbroucke. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*, 370(9596):1453–1457, 2007. https://doi.org/10.1016/S0140-6736(07)61602-X.

[39] Gary S. Collins, Karel G. M. Moons, Paula Dhiman, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. https://doi.org/10.1136/bmj-2023-078378.

[40] Andrew Bate, Marie Lindquist, I. Ralph Edwards, Sten Olsson, Roland Orre, Anders Lansner, and Ralph M. De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54:315–321, 1998. https://doi.org/10.1007/s002280050466.

[41] William DuMouchel. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician*, 53(3):177–190, 1999. https://doi.org/10.1080/00031305.1999.10474456.

[42] Eugene P. van Puijenbroek, Andrew Bate, Hubert G. M. Leufkens, Marie Lindquist, Roland Orre, and Antoine C. G. Egberts. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and Drug Safety*, 11(1):3–10, 2002. https://doi.org/10.1002/pds.668.

[43] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[44] U.S. Food and Drug Administration. FOUNDAYO (orforglipron) tablets, for oral use: Prescribing information, 2026. URL https://www.accessdata.fda.gov/drugsatfda_docs/label/2026/220934Orig1s000lbl.pdf. Initial U.S. approval 2026; accessed 5 November 2026.

[45] U.S. Food and Drug Administration. WEGOVY (semaglutide) injection and tablets: Prescribing information, 2026. URL https://www.accessdata.fda.gov/drugsatfda_docs/label/2026/218316s007lbl.pdf. Revised June 2026; accessed 5 November 2026.

[46] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 623–631, 2013. https://doi.org/10.1145/2487575.2487579.

[47] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. https://doi.org/10.1023/A:1010933404324.

[48] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. https://doi.org/10.1145/2939672.2939785.

[49] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[50] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10 (3):e0118432, 2015. https://doi.org/10.1371/journal.pone.0118432.

[51] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988. https://doi.org/10.2307/2531595.

[52] Xu Sun and Weichao Xu. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014. https://doi.org/10.1109/LSP.2014.2337313.

[53] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[54] Andrew J. Vickers and Elena B. Elkin. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006. https://doi.org/10.1177/0272989X06295361.

[55] U.S. Food and Drug Administration. FDA adverse event monitoring system (AEMS), 2026. URL https://www.fda.gov/safety/fda-adverse-event-monitoring-system-aems. Accessed 5 November 2026.

[56] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. https://doi.org/10.1038/s41586-024-08328-6.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Fig3.pdf
- Fig2.png
- Fig1.png
- Fig5.pdf
- Fig4.pdf