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.

Temporal shifts in pancreatitis reporting with GLP-1 receptor agonists: active-comparator disproportionality analyses and machine-learning report prioritization in FAERS

Festina Balidemaj

fe7247ba-s@student.lu.se

Lund University

Stephen Burleigh

Lund University

Research Article

Keywords: FAERS, pharmacovigilance, GLP-1 receptor agonists, DPP-4 inhibitors, SGLT2 inhibitors, pancreatitis, disproportionality analysis, active comparator, reporting bias, machine learning

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11441953/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Spontaneous reporting changes over time in volume, reporters, and drugs. Using pancreatitis, a labelled event, with glucagon-like peptide-1 receptor agonists (GLP-1 RAs) as a case study, we assessed how these changes affect disproportionality analyses and machine-learning report prioritization in the US Food and Drug Administration Adverse Event Reporting System (FAERS).

Methods

FAERS files (2015–2025) were deduplicated. Pancreatitis was defined by the narrow Acute pancreatitis Standardised MedDRA Query. Proportional reporting ratios, reporting odds ratios (RORs), information components, and empirical Bayes geometric means were calculated for seven GLP-1 RAs with standard thresholds, without causality assessment. Exploratory analyses used dipeptidyl peptidase-4 (DPP-4) and sodium-glucose cotransporter-2 (SGLT2) inhibitors as active comparators, crude and adjusted for calendar year, reporter type, age, sex, and region. As a methodological demonstration, machine-learning models were compared with simple baselines under temporal validation and annual retraining.

Results

Of 14,930,529 reports, 359,192 involved a GLP-1 RA, 7,321 with pancreatitis. All agents met all signal criteria against the full database (class ROR 8.00; 95% confidence interval 7.80–8.21). Against DPP-4 inhibitors, the crude class ROR was 0.99 (0.94–1.04), although GLP-1 RA reports included pancreatitis more often in every year. The year-adjusted (primary) ROR was 1.38 (1.31–1.45) and the fully adjusted ROR 1.85 (1.74–1.96), but among the minority of reports with a recorded diabetes indication it was 0.85 (0.80–0.91). Against SGLT2 inhibitors, the class ROR was 1.86 (1.76–1.97) crude and 2.12 (2.01–2.25) adjusted for year. Under temporal validation, gradient boosting outperformed the best baseline (area under the receiver operating characteristic curve 0.723 vs 0.627); a frozen model deteriorated (0.646 in 2025), whereas annually retrained gradient boosting remained at 0.73–0.80. Fields completed after case assessment inflated performance (0.938).

Conclusions

Changes in calendar time and reporter type influenced the crude comparison with DPP-4 inhibitors and eroded machine-learning prioritization, which annual retraining only partly mitigated. Pancreatitis was consistently reported more often with GLP-1 RAs than with SGLT2 inhibitors, whereas the comparison with DPP-4 inhibitors was inconclusive, and agent-level estimates varied with the comparator and model. These are reporting patterns for a labelled event, not estimates of risk or causation.

Background

Glucagon-like peptide-1 receptor agonists (GLP-1 RAs) are used to treat type 2 diabetes mellitus and obesity. In randomized trials, once-weekly semaglutide and tirzepatide produced substantial weight reduction in adults with overweight or obesity [1, 2], and liraglutide and semaglutide reduced major adverse cardiovascular events in patients at high cardiovascular risk [3, 4]. Their growing use increases the importance of post-marketing safety surveillance.

Pancreatitis has remained a recurring concern for GLP-1 RAs. An early analysis of the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) found that pancreatitis was reported more often with exenatide and sitagliptin than with other therapies [5], and in 2014 the FDA and the European Medicines Agency concluded that the available data did not support a causal association between incretin-based drugs and pancreatitis and considered the existing pancreatitis information in the product labelling adequate [6]. In January 2026, the UK Medicines and Healthcare products Regulatory Agency strengthened the product information of GLP-1 RAs and dual GLP-1/GIP receptor agonists to highlight reports of necrotising and fatal acute pancreatitis, while noting that its overall frequency remains uncommon [7]. Pancreatitis is thus a labelled event, and the question for newer analyses is how robust the reporting association is rather than whether a new signal exists. Later evidence has been inconsistent. Meta-analyses of randomized trials found either no increased risk of adjudicated acute pancreatitis [8] or a modest overall increase (risk ratio 1.44; 95% CI 1.09–1.89) that was not significant after stratification by background medication use [9]. A multinational population-based cohort study found no increased risk with incretin-based drugs compared with other oral antidiabetic drugs [10], whereas a claims-based research letter on GLP-1 RAs used for weight loss reported a higher risk than with bupropion–naltrexone [11]. Recent FAERS analyses have reported pancreatitis reporting signals for GLP-1 RAs as a class [12] and for each of seven individual agents [13].

FAERS is widely used for post-marketing signal detection but is affected by under-reporting, stimulated reporting, missing data, duplicate reports, and the lack of exposure denominators [14]. Disproportionality measures such as the proportional reporting ratio (PRR), reporting odds ratio (ROR), information component (IC), and empirical Bayes geometric mean (EBGM) compare the observed reporting of a drug–event combination with that expected from the reporting background [14–19]. Their results depend on the choice of reference group and on who reports. Previous FAERS studies of GLP-1 RAs compared them with all other drugs in the database [12, 13] or, for individual agents, with another GLP-1 RA [12]. One also analysed DPP-4 and SGLT2 inhibitors, but compared each class with all other drugs rather than with GLP-1 RAs [12], and, to our knowledge, none has examined a clinically relevant active comparator while accounting for changes in reporting over time.

Machine learning is increasingly studied in pharmacovigilance, including for classification, signal validation, triage, and report prioritization [20–25], and its value depends on the context of use, validation, calibration, interpretability, and governance [24–27]. Models trained on spontaneous reports inherit the same reporting patterns as disproportionality analyses, may degrade when the data to which they are applied differ from the training data [28, 29], and may appear to perform well if they use information about the outcome that would not be available at the time of prediction [30]. The rapid growth and changing composition of GLP-1 RA reports in recent years make them a realistic test of whether such models remain valid as the reporting population changes.

We used pancreatitis with GLP-1 RAs as a case study of how calendar time and the reporting population affect pharmacovigilance analyses. The aims were to assess pancreatitis reporting for seven GLP-1 RAs; to examine, in exploratory analyses without prespecified hypotheses, how the estimates change when the reference group, reporting population, drug role, indication, alternative causes, and calendar time are taken into account, and to describe the reported cases; and to evaluate whether machine-learning prioritization of GLP-1 RA reports is robust to changes in the reporting population and to the timing of the information used as predictors.

Methods

Study design and data source

This retrospective pharmacovigilance study used individual case safety reports (ICSRs) in FAERS, a public spontaneous reporting database maintained by the FDA. All 44 quarterly FAERS ASCII data files from the first quarter of 2015 to the fourth quarter of 2025 were downloaded on 1 October 2026 from the FDA website [31]. No study protocol was registered. The study is reported in accordance with the READUS-PV guideline [32, 33] (Additional file 2), and further methodological details are given in Additional file 1: Methods S1. A large language model was used for text editing and review during preparation of the manuscript; the authors reviewed all changes and take full responsibility for the content.

Data processing and definitions

Analyses were performed in Python 3.13.9 (pandas 2.3.3, NumPy 2.3.5, SciPy 1.16.3, scikit-learn 1.8.0, and matplotlib 3.10.6) and R 4.5.3 (openEBGM 0.10.0 [34]). Cases listed in the FDA deleted-case files were removed and only the latest version of each case was retained; reports were assigned to the year of initial FDA receipt, and cases first received from 2015 to 2025 formed the analytic dataset. GLP-1 RA exposure (exenatide, lixisenatide, liraglutide, albiglutide, semaglutide, dulaglutide, and tirzepatide) was identified from the active-ingredient and drug-name fields using predefined generic and brand names (Additional file 1: Table S1). Tirzepatide, a glucose-dependent insulinotropic polypeptide (GIP) and GLP-1 receptor agonist [2], was included with the GLP-1 RAs. In the main analysis, exposure in any reported role was counted.

The primary outcome was pancreatitis-related reporting, defined as at least one Medical Dictionary for Regulatory Activities (MedDRA) preferred term (PT) from the narrow-scope Acute pancreatitis Standardised MedDRA Query (SMQ) (MedDRA version 29.1; Additional file 1: Table S2a). Gallbladder/biliary events, defined by author-selected PTs (Additional file 1: Table S2b), were analysed as a positive control outcome and secondary exploratory outcome.

Disproportionality analysis

Report-level 2 × 2 tables compared reports with each GLP-1 RA with all other reports in the analytic dataset. The PRR and ROR were calculated with 95% confidence intervals (CIs) [15, 16] and the IC, without shrinkage, with an approximate 95% interval (lower bound IC025) [17]. The EBGM with its 5th and 95th percentiles (EB05 and EB95) was estimated with the gamma-Poisson shrinker [18] implemented in openEBGM [34], with the prior estimated from all active ingredient–PT pairs in the analytic dataset. Because this prior was estimated from

single ingredient–PT pairs and then applied to the composite pancreatitis outcome (reports with any PT in the SMQ), the EBGM values for pancreatitis are approximate (Additional file 1: Methods S1). The predefined signal criteria were PRR ≥ 2, lower 95% CI of the ROR > 1, IC025 > 0, and EB05 > 1, with at least three cases required for the PRR, ROR, and IC. These thresholds were screening criteria for statistical disproportionality, not evidence of causality, and case-level causality assessment was not performed.

Robustness analyses

The robustness analyses were developed after the main analysis and are exploratory and hypothesis-generating; no adjustment was made for multiple comparisons. First, RORs against all other reports were re-estimated with exposure restricted to the primary or secondary suspect drug, in reports from health professionals only, and after excluding reports with a co-reported PT indicating another common cause of acute pancreatitis (gallstone disease, alcohol-related terms, hypertriglyceridaemia, or hypercalcaemia; Additional file 1: Table S3).

Second, reports with a GLP-1 RA and no dipeptidyl peptidase-4 (DPP-4) inhibitor were compared with reports with a DPP-4 inhibitor and no GLP-1 RA. Because the two classes were marketed and reported in different periods, crude RORs were complemented by adjusted RORs from maximum-likelihood logistic regression including year of initial FDA receipt as a categorical variable (so that no particular time trend was assumed), age group, sex, reporter type, and reporting region (categories and reference levels in Additional file 1: Methods S1). The adjusted ROR was the exponentiated exposure coefficient, with a Wald 95% CI. Stepwise models with year only, reporter type only, and year and reporter type were also fitted, and the comparison was repeated in health-professional reports, with GLP-1 RA exposure restricted to diabetes products (defined below), with both restrictions combined, and in reports with a recorded diabetes indication. Year alone was regarded as the adjustment requiring the fewest assumptions, because reporter type may partly depend on the drug (for example, consumer reporting through manufacturer support programmes for newer agents) and on the seriousness of the event, and therefore may act as a mediator or collider rather than only as a confounder. The year-only model was therefore taken as the primary adjusted estimate for the active-comparator comparisons and their subgroups, with full adjustment as a sensitivity analysis; the other adjusted analyses (product-based indication, exclusion of lawyer-submitted reports, and positive control) are reported with full adjustment.

Because DPP-4 inhibitors have themselves been associated with pancreatitis reporting [5, 12] and were reported mainly in the earlier years, the comparison was repeated against a second active comparator, sodium-glucose cotransporter-2 (SGLT2) inhibitors (Additional file 1: Table S1), which are also used to treat type 2 diabetes and were reported throughout the study period. The comparison was constructed and adjusted in the same way, in all reports, in health-professional reports, with GLP-1 RA exposure restricted to diabetes products, and in reports in which diabetes was recorded as the indication for both the GLP-1 RA and the SGLT2 inhibitor. It was also repeated with exposure restricted to the primary or secondary suspect drug in both groups, and after excluding from both groups reports with a reaction term for diabetic ketoacidosis, Fournier's gangrene, or amputation, publicized events that dominate SGLT2 inhibitor reporting and could lower the proportion of SGLT2 inhibitor reports that include pancreatitis. For both comparators, the number of reports in each year was tabulated to confirm that both classes were represented in every year, year-specific crude RORs were estimated,

with a Mantel–Haenszel ROR stratified by year and a likelihood-ratio test for an exposure × year interaction, and the comparisons were repeated after excluding reports submitted by lawyers.

Third, because the recorded indication was missing or uninformative for most reports, the marketed product was used as a proxy for indication: products approved only for type 2 diabetes were classified as diabetes products, and the weight-management products of semaglutide (Wegovy), liraglutide (Saxenda), and tirzepatide (Zepbound) as weight products (Additional file 1: Methods S1). Within GLP-1 RA reports, pancreatitis reporting with weight products was compared with diabetes products of the same agent, crude and adjusted as above. Analyses based on the recorded indication are given in Additional file 1: Tables S4–S5.

Fourth, gallbladder/biliary events, associated with GLP-1 RA use in randomized trials [35], were analysed with the same comparisons as a positive control. Finally, reporter types and the proportion of GLP-1 RA reports including pancreatitis were tabulated by year.

Case characteristics and time to onset

Pancreatitis reports with GLP-1 RAs and with DPP-4 inhibitors were described by age, sex, reporter type, seriousness, indication, and co-reported pancreatitis causes. Time to onset, calculated as the event date minus the earliest therapy start date of the agent (for the class, of any GLP-1 RA in the report), for reports with complete dates and onset within 0–3,650 days, was summarized by median and interquartile range (IQR) and fitted with a Weibull distribution. Because only cases were analysed, without an exposed denominator, the Weibull shape parameter describes the distribution of reported onset times rather than a hazard in treated patients; a value below 1 indicates that reported onsets are concentrated soon after treatment initiation [36].

Machine-learning report prioritization

Because reaction terms are usually coded when a report is processed, the prioritization task was designed mainly as a test case for how report-ranking models behave when the reporting population changes, not as a tool for finding uncoded pancreatitis. Models were trained among GLP-1 RA reports to rank those containing a pancreatitis-related PT. The primary model used intake-type variables, that is, information of the kind available at report intake: age, body weight, number of reported drugs, sex, reporter type, reporting country, indicators for each GLP-1 RA and its suspect role, DPP-4 inhibitor co-exposure, multiple GLP-1 RA exposure, polypharmacy, and missingness indicators for age and weight. These variables were extracted from the latest version of each case, which may include follow-up information, so not all of them were necessarily available at initial receipt; a sensitivity analysis took them from the initial version of each case instead, to test the effect of these later updates. The number of reaction terms was not used, because a report whose only reaction term is the outcome term can only be outcome-positive. A sensitivity model added the seriousness criteria and report type, which are assigned after the reported event has been assessed and may therefore partly reflect the outcome.

Logistic regression with balanced class weights and histogram-based gradient boosting without class weighting were evaluated with a stratified random 80/20 split and with temporal validation (training on reports first received from 2015 to 2022, testing on 2023–2025), and compared with two simple baselines: ranking by the highest historical pancreatitis proportion among the GLP-1 RAs reported (historical agent rate), and a rule ranking health-professional reports first. Performance was assessed with the area under the receiver operating

characteristic curve (AUROC), the area under the precision–recall curve (AUPRC), calibration, and review workload, with 200 bootstrap resamples for uncertainty, including paired resamples for differences in AUROC; ties in ranking scores were broken at random. Models retrained each year on all earlier reports (expanding window) or on the three preceding years (sliding window) were compared with a frozen model trained on 2015–2017 reports (test years 2018–2025), performance was examined by reporter type and agent, and the shift in the reporting population was quantified with a classifier distinguishing 2023–2025 from 2015–2022 reports. Preprocessing, hyperparameters, calibration metrics, and interpretability methods are described in Additional file 1: Methods S1.

Results

Study population

After deduplication, 14,930,529 reports first received from 2015 to 2025 formed the analytic dataset (Additional file 1: Table S6). Of these, 359,192 involved a GLP-1 RA, and 7,321 (2.0%) of these included a pancreatitis-related PT.

Disproportionality findings

All seven GLP-1 RAs met all predefined signal criteria for pancreatitis against the full database (Table 1; Fig. 1). Liraglutide had the largest ROR and albiglutide the smallest, and the class ROR was 8.00 (95% CI 7.80–8.21). All five DPP-4 inhibitors also had EB05 values above 1 (Additional file 1: Table S7a).

Table 1
Disproportionality analysis for pancreatitis reports with GLP-1 receptor agonists in FAERS, 2015–2025

Exposure	Cases (n)	PRR (95% CI)	ROR (95% CI)	IC (IC025–IC975)	EBGM	EB05	EB95
Exenatide	619	6.33 (5.85–6.85)	6.43 (5.94–6.97)	2.65 (2.53–2.76)	6.22	5.82	6.65
Lixisenatide	55	4.79 (3.68–6.22)	4.84 (3.71–6.32)	2.26 (1.87–2.64)	4.58	3.65	5.68
Liraglutide	1,617	15.81 (15.06–16.60)	16.53 (15.71–17.39)	3.93 (3.86–4.01)	15.23	14.61	15.86
Albiglutide	78	2.93 (2.35–3.66)	2.95 (2.36–3.69)	1.55 (1.23–1.87)	2.87	2.38	3.45
Semaglutide	1,953	8.53 (8.16–8.92)	8.72 (8.33–9.13)	3.04 (2.97–3.10)	8.19	7.89	8.50
Dulaglutide	1,863	7.46 (7.13–7.82)	7.61 (7.26–7.97)	2.85 (2.78–2.92)	7.18	6.92	7.46
Tirzepatide	1,326	3.71 (3.51–3.92)	3.74 (3.54–3.95)	1.86 (1.78–1.94)	3.62	3.46	3.79
GLP-1 RA class	7,321	7.86 (7.67–8.06)	8.00 (7.80–8.21)	2.75 (2.72–2.79)	–	–	–

EBGM was estimated at the active-ingredient level, with the prior estimated from all active ingredient–preferred term pairs, and is therefore not reported for the class; because the prior was estimated from single preferred terms and applied to the composite outcome, EBGM values are approximate. CI confidence interval, EB05/EB95 5th/95th percentiles of the EBGM posterior distribution, EBGM empirical Bayes geometric mean, IC information component (without shrinkage), IC025/IC975 lower/upper bounds of the 95% interval, PRR proportional reporting ratio, ROR reporting odds ratio.

Robustness of the pancreatitis reporting estimates

Against all other reports, excluding reports with another co-reported pancreatitis cause changed the estimates little, restricting exposure to suspect drugs increased the class ROR only slightly but increased the RORs for liraglutide and semaglutide, and restricting to health-professional reports increased the class ROR to 14.74 (95% CI 14.22–15.29) (Additional file 1: Table S4).

Against DPP-4 inhibitors, the crude class ROR was 0.99 (95% CI 0.94–1.04). DPP-4 inhibitor reports were concentrated in earlier years (median year of initial receipt 2019 vs 2023 for GLP-1 RA reports), when the proportion of reports including pancreatitis was higher for both classes. Both classes were nevertheless reported in every year (5,352–12,249 DPP-4 inhibitor reports per year; Additional file 1: Table S8). In every year from 2015 to 2025, however, this proportion was higher for GLP-1 RAs than for DPP-4 inhibitors, although the year-specific RORs varied between years ($P < 0.001$ for interaction): 1.05–1.10 in 2015–2017 and 1.09–1.32 in 2023–2025, with confidence intervals mostly including 1, and 1.64–2.28 in 2018–2022, a period in which the DPP-4 inhibitor proportion fell steeply (Fig. 2a; Additional file 1: Tables S8–S9). The Mantel–Haenszel ROR stratified by year was 1.38, although, given the interaction, it summarizes heterogeneous year-specific associations; the same applies to the regression-adjusted estimates below. Adjustment for year alone, the primary adjusted estimate, gave a class ROR of 1.38 (1.31–1.45), and adjustment for year and reporter type 2.09 (1.98–2.21); with all covariates the class ROR was 1.85 (1.74–1.96) (Table 2; Fig. 2b; Additional file 1: Table S10). Year-adjusted RORs were above 1 for liraglutide, semaglutide, and dulaglutide, close to 1 for tirzepatide and lixisenatide, and below 1 for exenatide and albiglutide; fully adjusted RORs were above 1 for liraglutide, semaglutide, and dulaglutide, but not for tirzepatide, exenatide, or lixisenatide, and below 1 for albiglutide (Table 2; Additional file 1: Table S11).

The adjusted class estimate depended on the comparison. Adjusted for year, it was 3.25 (3.03–3.48) in health-professional reports, 1.35 (1.28–1.43) with GLP-1 RA exposure restricted to diabetes products, and 0.85 (0.80–0.91) among reports with a recorded diabetes indication, a selected minority of reports that may differ from those without a recorded indication. Fully adjusted, it was 2.62 (2.43–2.83) in health-professional reports and 1.78 (1.68–1.89) with GLP-1 RA exposure restricted to diabetes products, but 1.07 (0.99–1.16) among reports with a recorded diabetes indication (crude 0.73; Fig. 2b). Adjusted RORs for liraglutide and semaglutide were above 1 in all of these comparisons, for semaglutide only marginally so among reports with a recorded diabetes indication after full adjustment (1.16; 1.01–1.33; adjusted for year 1.48; 1.30–1.68); those for dulaglutide ranged from 0.81 to 2.17; and none of those for tirzepatide was above 1, and they were below 1 among reports with a recorded diabetes indication with either adjustment and in health-professional reports with full adjustment (Table 2; Additional file 1: Table S11).

Against SGLT2 inhibitors, the proportion of reports including pancreatitis was higher for GLP-1 RAs in every year, with year-specific RORs of 1.60–2.96 whose confidence intervals excluded 1 in every year but which varied

between years (P < 0.001 for interaction; Mantel–Haenszel ROR 2.11; Fig. 2a; Additional file 1: Table S9), and the class ROR was 1.86 (1.76–1.97) crude, 2.12 (2.01–2.25) adjusted for year, and 3.66 (3.44–3.89) fully adjusted; in health-professional reports, it was 4.04 (3.77–4.34) adjusted for year and 3.91 (3.63–4.20) fully adjusted (Table 2; Fig. 2b); with diabetes products, it was 2.12 (2.00–2.25) adjusted for year and 3.69 (3.46–3.93) fully adjusted (Additional file 1: Table S12). A diabetes indication was recorded in a similar proportion of SGLT2 inhibitor and GLP-1 RA reports (36.7% and 32.6%), and among these reports the class ROR was 1.63 (1.50–1.76) crude, 1.78 (1.64–1.93) adjusted for year, and 3.46 (3.16–3.78) fully adjusted (Table 2; Additional file 1: Table S12). RORs for liraglutide, semaglutide, and dulaglutide were above 1 in all SGLT2 inhibitor comparisons, and those for exenatide in all except the comparison restricted to a recorded diabetes indication (year-adjusted 1.08; 0.94–1.24). The tirzepatide ROR was not above 1 crude (except with diabetes products) or in health-professional reports but was above 1 after adjustment for year (1.31; 1.18–1.45) and after full adjustment (2.37; 2.07–2.70); among reports with a recorded diabetes indication, it was 1.06 (0.89–1.26) adjusted for year and 1.47 (1.18–1.84) fully adjusted, and with suspect drugs only, 1.10 (0.98–1.24) adjusted for year. Lawyers submitted less than 1% of SGLT2 inhibitor reports in every year (Additional file 1: Table S8), and excluding all lawyer-submitted reports changed the estimates little (fully adjusted class ROR 1.84 against DPP-4 inhibitors and 3.66 against SGLT2 inhibitors; Additional file 1: Table S13). Diabetic ketoacidosis, Fournier's gangrene, or amputation was reported in 15.2% of SGLT2 inhibitor reports and 0.3% of GLP-1 RA reports; excluding these reports from both groups changed the class ROR little (year-adjusted 2.04; 1.92–2.17), as did restricting both groups to suspect drugs (2.13; 2.00–2.27) (Additional file 1: Table S12).

The product-based indication classified 97.3% of GLP-1 RA reports (Additional file 1: Table S14). Within GLP-1 RA reports, pancreatitis was reported less often with weight products than with diabetes products (fully adjusted ROR 0.71; 0.66–0.78), for semaglutide (0.72) and tirzepatide (0.43) but not for liraglutide (1.05).

For the gallbladder/biliary positive control, the class ROR against DPP-4 inhibitors increased from 1.32 (1.20–1.44) crude to 2.67 (2.41–2.97) fully adjusted. Adjustment for year alone, however, decreased the gallbladder/biliary estimate (1.15; 1.05–1.27), unlike that for pancreatitis (Additional file 1: Tables S7b, S10 and S15). Against SGLT2 inhibitors, the gallbladder/biliary ROR increased from 2.88 (2.58–3.21) crude to 5.85 (5.20–6.58) fully adjusted (Additional file 1: Table S12).

Table 2
Pancreatitis RORs for GLP-1 receptor agonists under alternative comparisons

Reference group	Estimate	Class cases (n)	Class ROR (95% CI)	Liraglutide	Semaglutide	Dulaglutide	Tirzepatide
All other reports	Crude	7,321	8.00 (7.80–8.21)	16.53 (15.71–17.39)	8.72 (8.33–9.13)	7.61 (7.26–7.97)	3.74 (3.54–3.95)
DPP-4i	Crude	7,051	0.99 (0.94–1.04)	2.30 (2.15–2.46)	1.22 (1.15–1.30)	1.06 (0.99–1.13)	0.54 (0.50–0.57)
	Year-adjusted		1.38 (1.31–1.45)	2.18 (2.03–2.33)	2.36 (2.15–2.59)	1.36 (1.27–1.46)	0.94 (0.82–1.08)
	Fully adjusted		1.85 (1.74–1.96)	1.97 (1.83–2.12)	2.13 (1.93–2.36)	1.34 (1.23–1.45)	0.98 (0.84–1.15)
DPP-4i, health-professional reports	Year-adjusted	3,355	3.25 (3.03–3.48)	3.98 (3.65–4.33)	4.07 (3.66–4.53)	2.17 (1.93–2.43)	1.09 (0.87–1.37)
	Fully adjusted		2.62 (2.43–2.83)	3.06 (2.79–3.35)	2.95 (2.63–3.30)	1.71 (1.51–1.94)	0.69 (0.54–0.89)
DPP-4i, recorded diabetes indication	Crude	2,905	0.73 (0.68–0.78)	1.87 (1.70–2.05)	0.83 (0.75–0.91)	0.71 (0.65–0.78)	0.50 (0.45–0.56)
	Year-adjusted		0.85 (0.80–0.91)	1.76 (1.61–1.94)	1.48 (1.30–1.68)	0.86 (0.78–0.94)	0.82 (0.67–0.99)
	Fully adjusted		1.07 (0.99–1.16)	1.45 (1.31–1.61)	1.16 (1.01–1.33)	0.81 (0.72–0.90)	0.55 (0.44–0.69)
SGLT2i	Crude	6,775	1.86 (1.76–1.97)	4.54 (4.23–4.89)	2.27 (2.12–2.44)	2.02 (1.88–2.16)	0.98 (0.91–1.05)
	Year-adjusted		2.12 (2.01–2.25)	3.73 (3.46–4.03)	3.02 (2.78–3.28)	1.96 (1.83–2.11)	1.31 (1.18–1.45)
	Fully adjusted		3.66 (3.44–3.89)	4.67 (4.32–5.05)	3.62 (3.32–3.96)	3.16 (2.90–3.44)	2.37 (2.07–2.70)
SGLT2i, health-professional reports	Year-adjusted	3,192	4.04 (3.77–4.34)	5.45 (4.98–5.97)	4.39 (4.00–4.83)	2.56 (2.28–2.88)	1.21 (0.98–1.49)

Reference group	Estimate	Class cases (n)	Class ROR (95% CI)	Liraglutide	Semaglutide	Dulaglutide	Tirzepatide
	Fully adjusted		3.91 (3.63–4.20)	5.40 (4.92–5.93)	3.92 (3.54–4.33)	2.57 (2.27–2.91)	1.00 (0.79–1.25)
SGLT2i, recorded diabetes indication	Crude	2,722	1.63 (1.50–1.76)	4.47 (4.03–4.96)	1.72 (1.55–1.92)	1.63 (1.47–1.80)	1.06 (0.93–1.21)
	Year-adjusted		1.78 (1.64–1.93)	4.13 (3.71–4.59)	2.08 (1.83–2.38)	1.69 (1.52–1.87)	1.06 (0.89–1.26)
	Fully adjusted		3.46 (3.16–3.78)	5.40 (4.83–6.04)	2.70 (2.34–3.10)	3.20 (2.82–3.64)	1.47 (1.18–1.84)

Agent columns show ROR (95% CI). Year-adjusted: logistic regression with year of initial FDA receipt as a categorical variable. Fully adjusted: logistic regression adjusted for year of initial FDA receipt, age group, sex, reporter type, and reporting region. All other reports: all reports without the GLP-1 RA in question. DPP-4i: reports with a DPP-4 inhibitor and no GLP-1 RA (exposed: GLP-1 RA and no DPP-4 inhibitor). SGLT2i: reports with an SGLT2 inhibitor and no GLP-1 RA (exposed: GLP-1 RA and no SGLT2 inhibitor). Recorded diabetes indication: diabetes recorded as the indication for the GLP-1 RA in the exposed group (class: any GLP-1 RA) and for the DPP-4 or SGLT2 inhibitor in the reference group. These reports were a minority and may differ from reports without a recorded indication. Class cases: GLP-1 RA reports with pancreatitis in the comparison. Results for all agents and further comparisons are given in Additional file 1: Tables S4 and S10–S12. DPP-4i dipeptidyl peptidase-4 inhibitor, ROR reporting odds ratio, SGLT2i sodium-glucose cotransporter-2 inhibitor.

Changes in the reporting population over time

GLP-1 RA reports increased from 14,248 in 2015 to 95,898 in 2025, mostly from 2023 onward. Over the same period, the share submitted by consumers increased from 61.2% to 84.2%, the share from health professionals decreased from 21.9% to 12.6%, and the proportion including pancreatitis fell from 4.2% to 1.4% (Fig. 3; Additional file 1: Table S16).

Case characteristics and time to onset

Compared with DPP-4 inhibitor pancreatitis reports, GLP-1 RA pancreatitis reports concerned younger patients and less often reported hospitalization or death (Table 3). Health professionals submitted most semaglutide and liraglutide pancreatitis reports (71.2% and 68.8%) but only 9.4% of tirzepatide reports. Time to onset could be calculated for 1,592 GLP-1 RA pancreatitis reports (21.7%): the median was 68 days, 55.5% of events occurred within 90 days, and the Weibull shape parameter was 0.53, with point estimates below 1 for all agents with sufficient data (Table 3; Fig. 4).

Table 3
Characteristics and time to onset of pancreatitis reports

Characteristic	GLP-1 RA class	Liraglutide	Semaglutide	Dulaglutide	Tirzepatide	Exenatide	DPP-4i
Pancreatitis reports (n)	7,321	1,617	1,953	1,863	1,326	619	2,114
Age, years, median (IQR)	57 (49–65)	57 (48–65)	57 (48–66)	59 (50–66)	52 (41–61)	60 (53–68)	65 (56–73)
Female, % of known sex	55.0	57.7	54.9	48.9	66.2	49.7	41.4
Health-professional reporter, %	47.7	68.8	71.2	27.0	9.4	56.5	59.3
Consumer reporter, %	49.0	27.8	26.7	70.6	88.2	22.9	38.6
Hospitalization, %	48.9	49.7	53.8	45.9	46.9	45.9	64.2
Life-threatening, %	5.4	4.4	5.8	3.4	9.0	5.3	7.9
Death, %	3.6	3.8	3.8	2.8	2.2	10.2	7.8
Diabetes indication, %	42.1	47.9	37.5	46.6	28.4	55.4	65.3
Weight/obesity indication, %	11.4	11.4	15.6	1.8	24.7	1.0	0.0
Other pancreatitis cause co-reported, %	6.8	6.9	7.3	5.9	9.3	4.8	4.2
Time to onset available (n)	1,592	345	485	355	274	143	n/a
Time to onset, days, median (IQR)	68 (14–228)	75 (17–280)	62 (8–203)	79 (12–277)	56 (14–137)	188 (25–961)	n/a
Onset within 90 days, %	55.5	53.3	57.9	52.1	62.0	42.0	n/a
Weibull shape (95% CI)	0.53 (0.51–0.55)	0.54 (0.51–0.59)	0.52 (0.48–0.55)	0.54 (0.50–0.58)	0.64 (0.59–0.72)	0.53 (0.48–0.60)	n/a

DPP-4i: pancreatitis reports with a DPP-4 inhibitor and no GLP-1 RA. Lixisenatide (55 reports) and albiglutide (78 reports) are not shown; their time to onset is shown in Fig. 4 where at least 10 reports had complete dates. Percentages are of all reports in the column; indication: recorded for the agent in the column (class: any GLP-1

RA), and most reports had no informative indication. Time to onset: from the earliest therapy start date of the agent (see Methods). DPP-4i dipeptidyl peptidase-4 inhibitor, IQR interquartile range, n/a not assessed.

Machine-learning report prioritization

In the stratified random split, the AUROC was 0.810 for gradient boosting, 0.772 for logistic regression, 0.685 for the health-professional rule, and 0.635 for the historical agent rate (Table 4; Fig. 5a). Under temporal validation, performance declined for all methods: the AUROC was 0.723 (95% CI 0.714–0.733) for gradient boosting, 0.691 (0.681–0.701) for logistic regression, and 0.627 and 0.591 for the health-professional rule and historical agent rate, respectively (Table 4; Fig. 5b; Additional file 1: Table S17). Gradient boosting exceeded the health-professional rule by 0.096 (0.088–0.103) and logistic regression by 0.032 (0.025–0.039) (Additional file 1: Table S18). The temporal AUPRC of gradient boosting (0.048) was more than three times the prevalence of 1.4%, which is the AUPRC expected from random ranking. Within the top 1% of ranked reports, gradient boosting required 12.2 reports per pancreatitis case found, compared with 28.3 for the health-professional rule and 69.1 for random review, and it found half of all cases after review of 19.4% of reports (Table 5; Additional file 1: Table S19).

A frozen gradient boosting model trained on 2015–2017 reports deteriorated from an AUROC of 0.797 in 2018 to 0.646 in 2025, whereas models retrained each year had AUROCs between 0.729 and 0.797 with either an expanding or a 3-year sliding window (Fig. 6; Additional file 1: Table S20). The AUPRC nevertheless declined over time with all schemes as the prevalence of pancreatitis fell. A classifier distinguished 2023–2025 from 2015–2022 reports with an AUROC of 0.924, driven mainly by tirzepatide and semaglutide exposure and reporter type (Additional file 1: Table S21). Under temporal validation, performance was lower in consumer than in health-professional reports (AUROC 0.643 vs 0.745) and for tirzepatide than for semaglutide (0.670 vs 0.773) (Additional file 1: Table S22). Of the GLP-1 RA cases, 8.9% had follow-up versions; with predictors from the initial version of each case, the temporal AUROC of gradient boosting was 0.693 (0.685–0.703) (Table 4).

Under temporal validation, logistic regression overpredicted risk, as expected with balanced class weights, whereas the mean predicted risk of gradient boosting was close to the observed prevalence, although its calibration slope (0.81) indicated that its predictions were too extreme; recalibration corrected average risk without improving discrimination (Additional file 1: Tables S23–S24 and Fig. S1). Adding seriousness criteria and report type increased the AUROC of gradient boosting to 0.961 in the random split and 0.938 under temporal validation (Table 4; Additional file 1: Table S25); pancreatitis-positive reports were more often coded as hospitalization (48.9% vs 11.4%) and were more often expedited reports (44.6% vs 28.9%) than pancreatitis-negative reports. Reporter type and the suspect-role indicators contributed most to the rankings of both primary models, and reporting country and weight missingness also contributed to gradient boosting (Additional file 1: Figs. S2–S3).

Table 4
Discrimination of pancreatitis report prioritization methods

Predictors	Method	Random split AUROC	Random split AUPRC	Temporal AUROC	Temporal AUPRC
Baseline	Historical agent rate	0.635	0.031	0.591	0.019
Baseline	Health-professional rule	0.685	0.046	0.627	0.028
Intake (primary)	Logistic regression	0.772	0.076	0.691 (0.681–0.701)	0.040
Intake (primary)	Gradient boosting	0.810	0.153	0.723 (0.714–0.733)	0.048
Intake, initial case version	Logistic regression	0.768	0.074	0.687 (0.678–0.698)	0.039
Intake, initial case version	Gradient boosting	0.806	0.153	0.693 (0.685–0.703)	0.045
+ Seriousness, report type	Logistic regression	0.927	0.153	0.918 (0.914–0.921)	0.101
+ Seriousness, report type	Gradient boosting	0.961	0.478	0.938 (0.935–0.941)	0.207

Random split: stratified 80/20 split (test n = 71,839; 1,464 positives). Temporal: training on reports first received 2015–2022 (n = 156,379), testing on 2023–2025 (n = 202,813; 2,934 positives); temporal AUROCs of the models are shown with 95% CIs from 200 bootstrap resamples of the test set. Health-professional rule: health-professional reports ranked first, ties broken by the historical agent rate. Intake (primary): intake-type variables extracted from the latest case version (see Methods); initial case version: the same predictors taken from the first version of each case received by the FDA (Additional file 1: Methods S1). AUPRC area under the precision–recall curve, AUROC area under the receiver operating characteristic curve.

Table 5
Prioritization and review workload under temporal validation

Method	Recall, top 1% (%)	Recall, top 5% (%)	Recall, top 10% (%)	Reports per case, top 1%	Reports per case, top 10%	Reports reviewed to find 50% of cases (%)
Gradient boosting	5.7	25.2	38.6	12.2	17.9	19.4
Logistic regression	4.7	22.5	35.8	14.6	19.3	22.0
Baseline: health-professional rule	2.4	15.7	32.5	28.3	21.3	29.0
Baseline: historical agent rate	1.7	8.3	16.1	39.9	43.0	32.2
Random review	1.0	5.0	10.0	69.1	69.1	50.0

Recall is the percentage of all pancreatitis-positive test reports found within the top k% of ranked reports. Reports per case: number of reports reviewed per pancreatitis-positive report found. Ties in ranking scores were broken at random (expected values). Top-k results for both validation designs are given in Additional file 1: Table S19. Workload figures describe relative ranking performance in retrospective data and are not estimates of operational benefit, because pancreatitis terms are normally coded at intake.

Discussion

Principal findings

Using pancreatitis with GLP-1 RAs as a case study, this study showed how calendar time and the reporting population shape pharmacovigilance analyses. All seven GLP-1 RAs showed pancreatitis reporting signals against the full database, consistent with previous FAERS-based studies [12, 13]. Against DPP-4 inhibitors, the crude comparison showed no difference in reporting, partly because the classes differed in calendar time and reporter type; adjusted analyses showed higher reporting for GLP-1 RAs, but not among reports with a recorded diabetes indication (year-adjusted ROR 0.85; a selected minority of reports), the comparison most comparable with respect to recorded indication, so the class-level result against DPP-4 inhibitors was inconclusive. Against SGLT2 inhibitors, class-level reporting was higher without adjustment and more so after it, including among reports with a recorded diabetes indication. Agent-level RORs varied in size with the comparator and model: those for liraglutide and semaglutide were above 1 in every adjusted comparison and those for tirzepatide were not consistently above 1, differences that may reflect time on the market, reporter type, and indication rather than differences between the agents themselves. The same shift in the reporting population degraded machine-learning prioritization, which annual retraining only partly mitigated. These analyses were exploratory and post hoc, and all estimates describe reporting, not the clinical risk of pancreatitis or causation.

Interpretation of the disproportionality findings

The crude active-comparator estimate was likely confounded by the timing and source of reports, although spontaneous reporting data cannot identify every source of bias. DPP-4 inhibitor reports were concentrated in earlier years and GLP-1 RA reports in later years, when the number of reports increased several-fold, consumer reports grew, and the proportion involving pancreatitis fell for both classes. Pooling over years therefore compared GLP-1 RA reports from a period with low pancreatitis proportions with DPP-4 inhibitor reports from a period with higher proportions, and the difference between periods offset the higher point estimates seen in every year, a form of Simpson's paradox. Adjustment for year alone increased the class estimate from 0.99 to 1.38, and additional adjustment for reporter type, which captured the growth in consumer reports, increased it to 2.09. Reporter type may, however, partly depend on the drug and on the seriousness of the event, so adjustment for it may introduce bias as well as remove it, and the year-only estimate, which requires the fewest modelling assumptions, was therefore taken as the primary adjusted estimate. The within-year difference from DPP-4 inhibitors was also not uniform: it was largest in 2018–2022, when DPP-4 inhibitor pancreatitis reporting fell steeply for reasons that cannot be determined from FAERS. The year-adjusted and fully adjusted RORs are therefore summaries across heterogeneous years rather than single stable associations, and they should be read alongside the year-specific estimates (Fig. 2a; Additional file 1: Table S9). Active-comparator analyses in FAERS should therefore account for calendar time and examine the influence of reporter characteristics, particularly when the compared drugs were marketed in different periods, recognizing that adjustment for the variables recorded in FAERS cannot remove all confounding by indication or all differences in reporting behaviour between the compared classes.

The second active comparator illustrates how strongly the result depends on the choice of comparator. SGLT2 inhibitors were reported over the same years as GLP-1 RAs, so the crude comparison was less affected by calendar time, and the GLP-1 RA proportion was higher in every year, although the size of the difference varied. Because DPP-4 inhibitors have themselves been associated with pancreatitis reporting in FAERS [5, 12] and in the present data (Additional file 1: Table S7a), that comparator would tend to reduce, rather than inflate, the RORs estimated for GLP-1 RAs, consistent with the larger estimates against SGLT2 inhibitors, although differences between the patients and reporters of the two comparator classes may also contribute. In particular, SGLT2 inhibitors have been approved for heart failure and chronic kidney disease since 2020, so later SGLT2 inhibitor reports increasingly concern patients without diabetes, whose baseline pancreatitis risk and reporters may differ from those of GLP-1 RA users; however, restricting both classes to reports with a recorded diabetes indication changed the estimate little (year-adjusted ROR 1.78 vs 2.12 in all reports). SGLT2 inhibitors also showed a pancreatitis reporting signal in a previous FAERS analysis, although a weaker one than DPP-4 inhibitors [12]. In addition, SGLT2 inhibitor reports are dominated by other publicized events, such as diabetic ketoacidosis, Fournier's gangrene, and, for canagliflozin, amputation; these may lower the proportion of SGLT2 inhibitor reports that include pancreatitis and so inflate RORs against this comparator, a form of competition bias [37]. Excluding reports with these events from both groups, however, changed the class estimate little (year-adjusted ROR 2.04 vs 2.12), so competition bias from these events is unlikely to explain the result. Litigation can influence FAERS reporting, but lawyers submitted few SGLT2 inhibitor reports, and excluding lawyer-submitted reports changed the estimates little (Additional file 1: Table S13).

The positive control is of limited help in judging whether the adjusted estimates are unbiased. GLP-1 RA use was associated with gallbladder and biliary diseases in a meta-analysis of randomized trials, especially at higher doses, with longer duration, and when used for weight loss [35], and the gallbladder/biliary ROR against

DPP-4 inhibitors increased after full adjustment, but adjustment for year alone moved it in the opposite direction to pancreatitis, and the large shifts between crude and adjusted estimates indicate that the adjusted estimates are sensitive to model choice. Gallstones are also a cause of pancreatitis, so the positive control is not independent of the primary outcome, and excluding gallstone-related reports may remove cases on a possible pathway rather than only confounded cases.

Higher class-level reporting than with DPP-4 inhibitors was not seen among reports with a recorded diabetes indication. Because DPP-4 inhibitors are used only for type 2 diabetes, this restriction addressed confounding by indication most directly, although comparability of recorded indication does not imply comparable disease severity, prescribing patterns, or baseline pancreatitis risk. These reports were, however, a minority, and a diabetes indication was recorded more often for DPP-4 inhibitor than for GLP-1 RA pancreatitis reports (65.3% vs 42.1%; Table 3), so the restriction may also introduce selection. In this subgroup, the class estimate was below 1 after adjustment for year (0.85; 0.80–0.91) and close to 1 after full adjustment (1.07; 0.99–1.16), so even its direction depended on the model. The difference between the year-adjusted estimates with diabetes products (1.35) and with a recorded diabetes indication (0.85) may partly reflect incomplete exclusion of use for weight management by the product-based proxy, for example off-label use of the diabetes products of semaglutide and tirzepatide, although FAERS cannot show how much of the difference this explains. Higher reporting persisted with diabetes products, in health-professional reports, and against SGLT2 inhibitors, including among reports with a recorded diabetes indication (year-adjusted ROR 1.78). That the same restriction gave different results against the two comparators suggests that the lower estimate against DPP-4 inhibitors in this subgroup may partly reflect pancreatitis reporting with DPP-4 inhibitors themselves rather than indication alone, but confounding by indication cannot be excluded, and the class-level result against DPP-4 inhibitors should be regarded as inconclusive. Adjusted RORs for liraglutide and semaglutide were above 1 in all comparisons, and liraglutide also had the highest ROR among GLP-1 RAs in previous FAERS analyses [12, 13], in which tirzepatide had the lowest [13]. Differences between agents in spontaneous reporting cannot be read as differences in risk, because the agents differ in time on the market, in the years in which most of their reports were received, and in reporter type, indication, dose, and publicity; liraglutide and exenatide, for example, were marketed earlier than semaglutide and tirzepatide, so more of their reports came from the earlier years, when pancreatitis was reported more often for all classes (Fig. 2a). Tirzepatide reports were recent and mostly submitted by consumers, so its estimates may be affected by stimulated reporting [14] and limited follow-up, and its results were inconsistent: its RORs were not above 1 against DPP-4 inhibitors and were above 1 against SGLT2 inhibitors after adjustment, but not in health-professional reports or, after adjustment for year, among reports with a recorded diabetes indication or with suspect drugs only. Lower pancreatitis reporting with weight products than with diabetes products may reflect differences in patients, baseline risk, or reporting, and off-label use of diabetes products for weight management would bias this comparison toward the null.

Most reported onsets occurred within the first months of treatment, consistent with a previous FAERS analysis (median 92 days) [13]. Onset times were also truncated by time on the market and by the study period: tirzepatide was approved in 2022, so its reports could not include long onsets, which shortens its reported time to onset compared with older agents. The early concentration of onsets may reflect a tendency to report events soon after a drug is started, closer attention to new symptoms early in treatment, or an early-onset reaction, and dates were available for only about one fifth of reports; the pattern warrants further evaluation in data sources with reliable exposure and event timing.

These reporting findings should be considered alongside clinical studies that have estimated pancreatitis risk directly and have given inconsistent results, as summarized in the Background [8–11]. As pancreatitis is a labelled event [6, 7], disproportionality analyses here measure reporting rather than incidence and cannot establish whether GLP-1 RAs cause pancreatitis.

Implications for machine-learning prioritization

Models using intake-type variables extracted from the latest case version outperformed simple rules, with moderate discrimination and a lower review workload in retrospective evaluation; workload measures are more interpretable than the AUROC alone. Because pancreatitis terms are normally coded at intake, however, these results illustrate how report-ranking models behave under reporting shift rather than a practical gain for this outcome. A frozen model deteriorated as the reporting population shifted toward consumer reports and newer agents, an example of dataset shift [28], whereas annual retraining partly mitigated this deterioration and generally preserved better discrimination than the frozen model, although performance still varied between years and the AUPRC fell substantially as pancreatitis became rarer. Using only the first version of each case reduced the temporal AUROC of gradient boosting, indicating that part of its performance came from follow-up information, although it still exceeded the best baseline. Seriousness criteria and report type produced much higher apparent performance, but they are completed after the event has been assessed and partly re-encode the outcome, a form of leakage [30]. Predictors should therefore match the information available at the intended point in the workflow, and deployed models require monitoring and scheduled updating, as shown for clinical prediction models whose calibration deteriorated over time [29].

These results support a conservative, risk-based approach to artificial intelligence in pharmacovigilance, with emphasis on validity, robustness, transparency, human oversight, and continuous monitoring [24–27]; the models here were intended only to rank reports for human review, not to assess causality, replace medical review, or trigger regulatory action.

Strengths and limitations

Strengths include the use of all quarterly FAERS files over 11 years with case-level deduplication, frequentist and Bayesian disproportionality measures, two active comparators with stepwise adjustment, a product-based indication proxy covering almost all reports, and a machine-learning evaluation with simple baselines, temporal validation, annual retraining, and attention to predictor timing.

FAERS lacks exposure denominators and is affected by under-reporting, duplicate and stimulated reporting, notoriety effects and litigation-related reporting, missing information, confounding by indication, co-medication, and differences in reporting across countries and over time; the results are therefore reporting signals rather than causal or incidence estimates. Reports naming only the generic semaglutide or tirzepatide may include compounded products. Competition from publicized events other than those examined may have inflated the estimates against SGLT2 inhibitors, and the extension of SGLT2 inhibitor indications to heart failure and chronic kidney disease changed the comparator population over time, although restriction to reports with a recorded diabetes indication changed the SGLT2 inhibitor estimate little. Adjustment was limited to variables recorded in FAERS, residual confounding is likely, adjustment for reporter type may itself introduce bias, and the positive control could not validate the adjustment. The EBGM for the composite outcome was approximate because its prior was estimated from single preferred terms. The product-based indication does not capture off-label use,

the recorded indication was available for a selected minority of reports, alternative pancreatitis causes were identified only from co-reported terms, and time to onset could be calculated for a minority of reports, was based on cases only, and was truncated by time on the market. The robustness analyses were not prespecified, multiple comparisons were not adjusted for, case-level causality assessment was not performed, and deduplication relied on case identifiers without probabilistic matching. Part of the machine-learning performance depended on follow-up information in later case versions, the features were limited to structured report-level variables, and the results may not generalize to other outcomes, databases, or prospective settings.

Conclusions

In FAERS from 2015 to 2025, changes in calendar time and in who reported substantially influenced the crude comparison of pancreatitis reporting between GLP-1 RAs and DPP-4 inhibitors, and they eroded machine-learning prioritization. In exploratory analyses accounting for these changes, the class-level result depended on the choice of comparator: pancreatitis was reported more often than with SGLT2 inhibitors, but the comparison with DPP-4 inhibitors was inconclusive, and agent-level estimates also varied with the comparator and model. These are reporting patterns for a labelled event; they do not estimate clinical risk, establish causation, or allow the agents to be ranked by risk. Disproportionality analyses and pharmacovigilance machine-learning tools should account for calendar time and the reporting population, and deployed models should be monitored and retrained regularly.

Abbreviations

AUPRC
Area under the precision–recall curve
AUROC
Area under the receiver operating characteristic curve
CI
Confidence interval
DPP-4i
Dipeptidyl peptidase-4 inhibitor
EB05/EB95
5th/95th percentiles of the EBGM posterior distribution
EBGM
Empirical Bayes geometric mean
FAERS
FDA Adverse Event Reporting System
FDA
US Food and Drug Administration
GIP
Glucose-dependent insulinotropic polypeptide
GLP-1 RA
Glucagon-like peptide-1 receptor agonist
IC

Information component
IC025/IC975
Lower/upper bounds of the 95% interval of the IC
ICSR
Individual case safety report
IQR
Interquartile range
MedDRA
Medical Dictionary for Regulatory Activities
PRR
Proportional reporting ratio
PT
Preferred term
READUS-PV
REporting of A Disproportionality Analysis for DrUg Safety signal detection using individual case safety reports in PharmacoVigilance
ROC
Receiver operating characteristic
ROR
Reporting odds ratio
SGLT2
Sodium-glucose cotransporter-2
SMQ
Standardised MedDRA Query

Declarations

Ethics approval and consent to participate

This study used publicly available, de-identified spontaneous reporting data from FAERS and did not involve direct interaction with human participants. Therefore, formal ethics approval and informed consent were not required.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed during the current study are available from the US Food and Drug Administration FAERS Quarterly Data Extract Files [31]. Raw FAERS files and processed analytic datasets are not redistributed but can

be regenerated from the public FAERS files with the analysis code [38]. Supplementary tables and figures are provided in Additional file 1, and the completed READUS-PV checklist in Additional file 2.

Project name: spontaneous-reporting-temporal-shift

Project home page: https://github.com/Festinabp/spontaneous-reporting-temporal-shift

Archived version: https://doi.org/10.5281/zenodo.23319956

Operating system(s): Platform independent

Programming language: Python 3.13 and R 4.5

Other requirements: pandas 2.3.3, NumPy 2.3.5, SciPy 1.16.3, scikit-learn 1.8.0, matplotlib 3.10.6, openEBGM 0.10.0

License: MIT

Any restrictions to use by non-academics: None

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for conducting this study.

Authors' contributions

FB contributed to study conception, data extraction, methodology, analysis, interpretation, and writing of the manuscript. SB contributed to supervision and critical review of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Clinical trial number

Not applicable.

References

1. 1. Wilding JPH, Batterham RL, Calanna S, Davies M, Van Gaal LF, Lingvay I, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med. 2021;384(11):989–1002. https://doi.org/10.1056/NEJMoa2032183.
2. 2. Jastreboff AM, Aronne LJ, Ahmad NN, Wharton S, Connery L, Alves B, et al. Tirzepatide once weekly for the treatment of obesity. N Engl J Med. 2022;387(3):205–16. https://doi.org/10.1056/NEJMoa2206038.
3. 3. Marso SP, Daniels GH, Brown-Frandsen K, Kristensen P, Mann JFE, Nauck MA, et al. Liraglutide and cardiovascular outcomes in type 2 diabetes. N Engl J Med. 2016;375(4):311–22. https://doi.org/10.1056/NEJMoa1603827.
4. 4. Lincoff AM, Brown-Frandsen K, Colhoun HM, Deanfield J, Emerson SS, Esbjerg S, et al. Semaglutide and cardiovascular outcomes in obesity without diabetes. N Engl J Med. 2023;389(24):2221–32. https://doi.org/10.1056/NEJMoa2307563.
5. 5. Elashoff M, Matveyenko AV, Gier B, Elashoff R, Butler PC. Pancreatitis, pancreatic, and thyroid cancer with glucagon-like peptide-1-based therapies. Gastroenterology. 2011;141(1):150–6. https://doi.org/10.1053/j.gastro.2011.02.018.
6. 6. Egan AG, Blind E, Dunder K, de Graeff PA, Hummer BT, Bourcier T, et al. Pancreatic safety of incretin-based drugs—FDA and EMA assessment. N Engl J Med. 2014;370(9):794–7. https://doi.org/10.1056/NEJMp1314078.
7. 7. Medicines and Healthcare products Regulatory Agency. GLP-1 receptor agonists and dual GLP-1/GIP receptor agonists: strengthened warnings on acute pancreatitis, including necrotising and fatal cases. Drug Safety Update. 2026;19(6). https://www.gov.uk/drug-safety-update/glp-1-receptor-agonists-and-dual-glp-1-slash-gip-receptor-agonists-strengthened-warnings-on-acute-pancreatitis-including-necrotising-and-fatal-cases. Accessed 12 Oct 2026.
8. 8. Storgaard H, Cold F, Gluud LL, Vilsbøll T, Knop FK. Glucagon-like peptide-1 receptor agonists and risk of acute pancreatitis in patients with type 2 diabetes. Diabetes Obes Metab. 2017;19(6):906–8. https://doi.org/10.1111/dom.12885.
9. 9. Wen J, Nadora D, Bernstein E, How-Volkman C, Truong A, Joy B, et al. Evaluating the rates of pancreatitis and pancreatic cancer among GLP-1 receptor agonists: a systematic review and meta-analysis of randomised controlled trials. Endocrinol Diabetes Metab. 2025;8(5):e70113. https://doi.org/10.1002/edm2.70113.
10. 10. Azoulay L, Filion KB, Platt RW, Dahl M, Dormuth CR, Clemens KK, et al. Association between incretin-based drugs and the risk of acute pancreatitis. JAMA Intern Med. 2016;176(10):1464–73. https://doi.org/10.1001/jamainternmed.2016.1522.
11. 11. Sodhi M, Rezaeianzadeh R, Kezouh A, Etminan M. Risk of gastrointestinal adverse events associated with glucagon-like peptide-1 receptor agonists for weight loss. JAMA. 2023;330(18):1795–7. https://doi.org/10.1001/jama.2023.19574.
12. 12. Alenzi KA, Alsuhaibani D, Batarfi B, Alshammari TM. Pancreatitis with use of new diabetic medications: a real-world data study using the post-marketing FDA adverse event reporting system (FAERS) database. Front Pharmacol. 2024;15:1364110. https://doi.org/10.3389/fphar.2024.1364110.
13. 13. Guo H, Guo Q, Li Z, Wang Z. Association between different GLP-1 receptor agonists and acute pancreatitis: case series and real-world pharmacovigilance analysis. Front Pharmacol. 2024;15:1461398. https://doi.org/10.3389/fphar.2024.1461398.

14. 14. Sakaeda T, Tamon A, Kadoyama K, Okuno Y. Data mining of the public version of the FDA Adverse Event Reporting System. Int J Med Sci. 2013;10(7):796–803. https://doi.org/10.7150/ijms.6048.
15. 15. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10(6):483–6. https://doi.org/10.1002/pds.677.
16. 16. van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11(1):3–10. https://doi.org/10.1002/pds.668.
17. 17. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54(4):315–21. https://doi.org/10.1007/s002280050466.
18. 18. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat. 1999;53(3):177–90. https://doi.org/10.1080/00031305.1999.10474456.
19. 19. Wisniewski AFZ, Bate A, Bousquet C, Brueckner A, Candore G, Juhlin K, et al. Good signal detection practices: evidence from IMI PROTECT. Drug Saf. 2016;39(6):469–90. https://doi.org/10.1007/s40264-016-0405-1.
20. 20. Warner J, Prada Jardim A, Albera C. Artificial intelligence: applications in pharmacovigilance signal management. Pharmaceut Med. 2025;39(3):183–98. https://doi.org/10.1007/s40290-025-00561-2.
21. 21. Schreier T, Tropmann-Frick M, Böhm R. Integration of FAERS, DrugBank and SIDER data for machine learning-based detection of adverse drug reactions. Datenbank-Spektrum. 2024;24(3):233–42. https://doi.org/10.1007/s13222-024-00486-1.
22. 22. Imran M, Bhatti A, King DM, Lerch M, Dietrich J, Doron G, et al. Supervised machine learning-based decision support for signal validation classification. Drug Saf. 2022;45(5):583–96. https://doi.org/10.1007/s40264-022-01159-2.
23. 23. Dimitsaki S, Natsiavas P, Jaulent MC. Applying AI to structured real-world data for pharmacovigilance purposes: scoping review. J Med Internet Res. 2024;26:e57824. https://doi.org/10.2196/57824.
24. 24. Ball R, Dal Pan G. "Artificial intelligence" for pharmacovigilance: ready for prime time? Drug Saf. 2022;45(5):429–38. https://doi.org/10.1007/s40264-022-01157-4.
25. 25. Pinheiro LC, Kurz X. Artificial intelligence in pharmacovigilance: a regulatory perspective on explainability. Pharmacoepidemiol Drug Saf. 2022;31(12):1308–10. https://doi.org/10.1002/pds.5524.
26. 26. Council for International Organizations of Medical Sciences (CIOMS). Artificial intelligence in pharmacovigilance. CIOMS Working Group XIV report. Geneva: CIOMS; 2025. https://doi.org/10.56759/cdob6397.
27. 27. Nagar A, Gobburu JVS, Chakravarty A. Artificial intelligence in pharmacovigilance: advancing drug safety monitoring and regulatory integration. Ther Adv Drug Saf. 2025;16:20420986251361435. https://doi.org/10.1177/20420986251361435.
28. 28. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283–6. https://doi.org/10.1056/NEJMc2104626.

29. 29. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052–61. https://doi.org/10.1093/jamia/ocx030.
30. 30. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data. 2012;6(4):15. https://doi.org/10.1145/2382577.2382579.
31. 31. US Food and Drug Administration. FDA Adverse Event Reporting System (FAERS): quarterly data extract files. https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html. Accessed 1 Oct 2026.
32. 32. Fusaroli M, Salvo F, Bégaud B, AlShammari TM, Bate A, Battini V, 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 Saf. 2024;47(6):575–84. https://doi.org/10.1007/s40264-024-01421-9.
33. 33. Fusaroli M, Salvo F, Bégaud B, AlShammari TM, Bate A, Battini V, et al. The REporting of A Disproportionality Analysis for DrUg Safety Signal Detection Using Individual Case Safety Reports in PharmacoVigilance (READUS-PV): explanation and elaboration. Drug Saf. 2024;47(6):585–99. https://doi.org/10.1007/s40264-024-01423-7.
34. 34. Canida T, Ihrie J. openEBGM: an R implementation of the Gamma-Poisson Shrinker data mining model. R J. 2017;9(2):499–519. https://doi.org/10.32614/RJ-2017-063.
35. 35. He L, Wang J, Ping F, Yang N, Huang J, Li Y, et al. Association of glucagon-like peptide-1 receptor agonist use with risk of gallbladder and biliary diseases: a systematic review and meta-analysis of randomized clinical trials. JAMA Intern Med. 2022;182(5):513–9. https://doi.org/10.1001/jamainternmed.2022.0338.
36. 36. Sauzet O, Carvajal A, Escudero A, Molokhia M, Cornelius VR. Illustration of the Weibull shape parameter signal detection tool using electronic healthcare record data. Drug Saf. 2013;36(10):995–1006. https://doi.org/10.1007/s40264-013-0061-7.
37. 37. Arnaud M, Salvo F, Ahmed I, Robinson P, Moore N, Bégaud B, et al. A method for the minimization of competition bias in signal detection from spontaneous reporting databases. Drug Saf. 2016;39(3):251–60. https://doi.org/10.1007/s40264-015-0375-8.
38. 38. Balidemaj F, Burleigh S. Temporal reporting shifts in pharmacovigilance: analysis code. Version 1.0. Zenodo. 2026. https://doi.org/10.5281/zenodo.23319956.

Figures

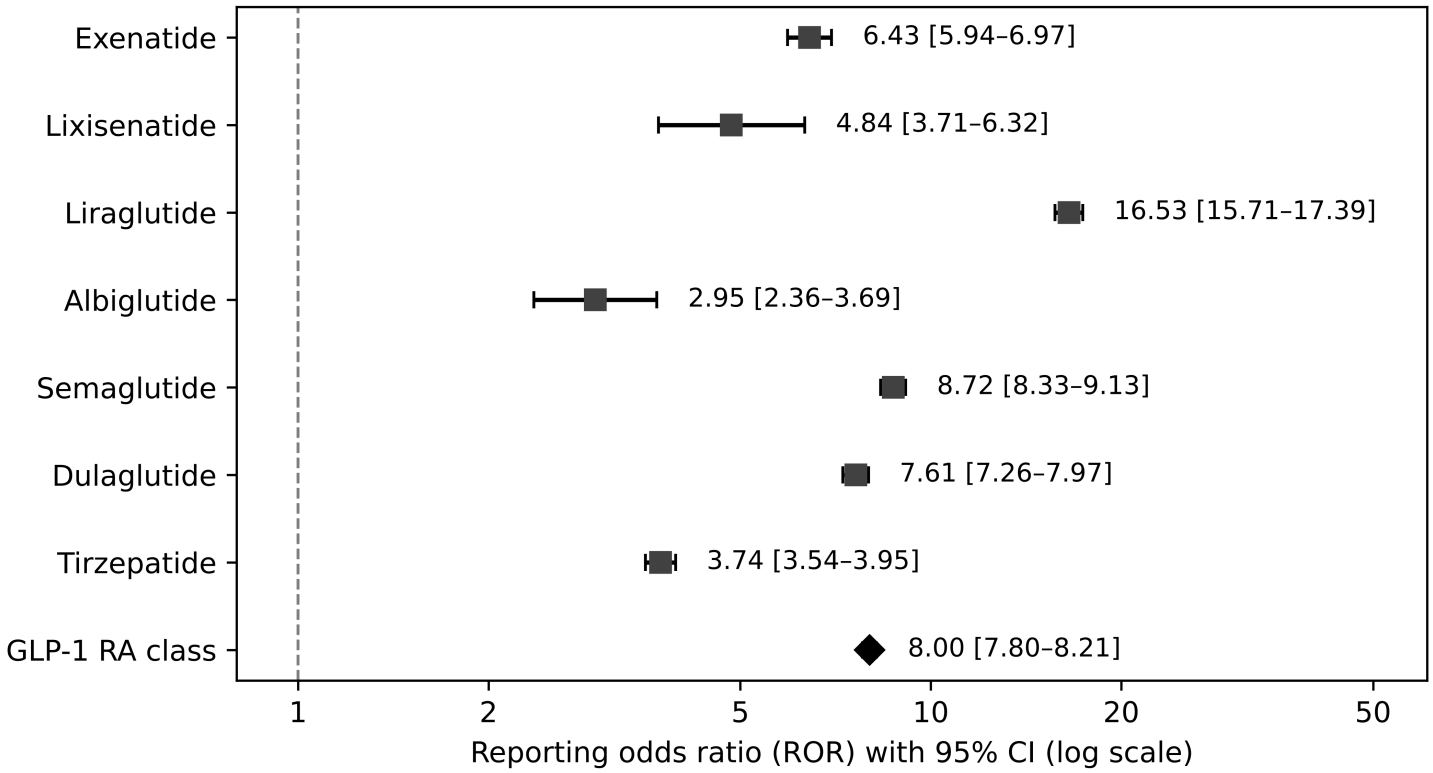


Figure 1

Forest plot of pancreatitis reporting odds ratios by GLP-1 receptor agonist

Reporting odds ratios with 95% CIs for each agent and the class against all other reports in FAERS, 2015–2025 (logarithmic scale). CI confidence interval, FAERS FDA Adverse Event Reporting System, ROR reporting odds ratio.

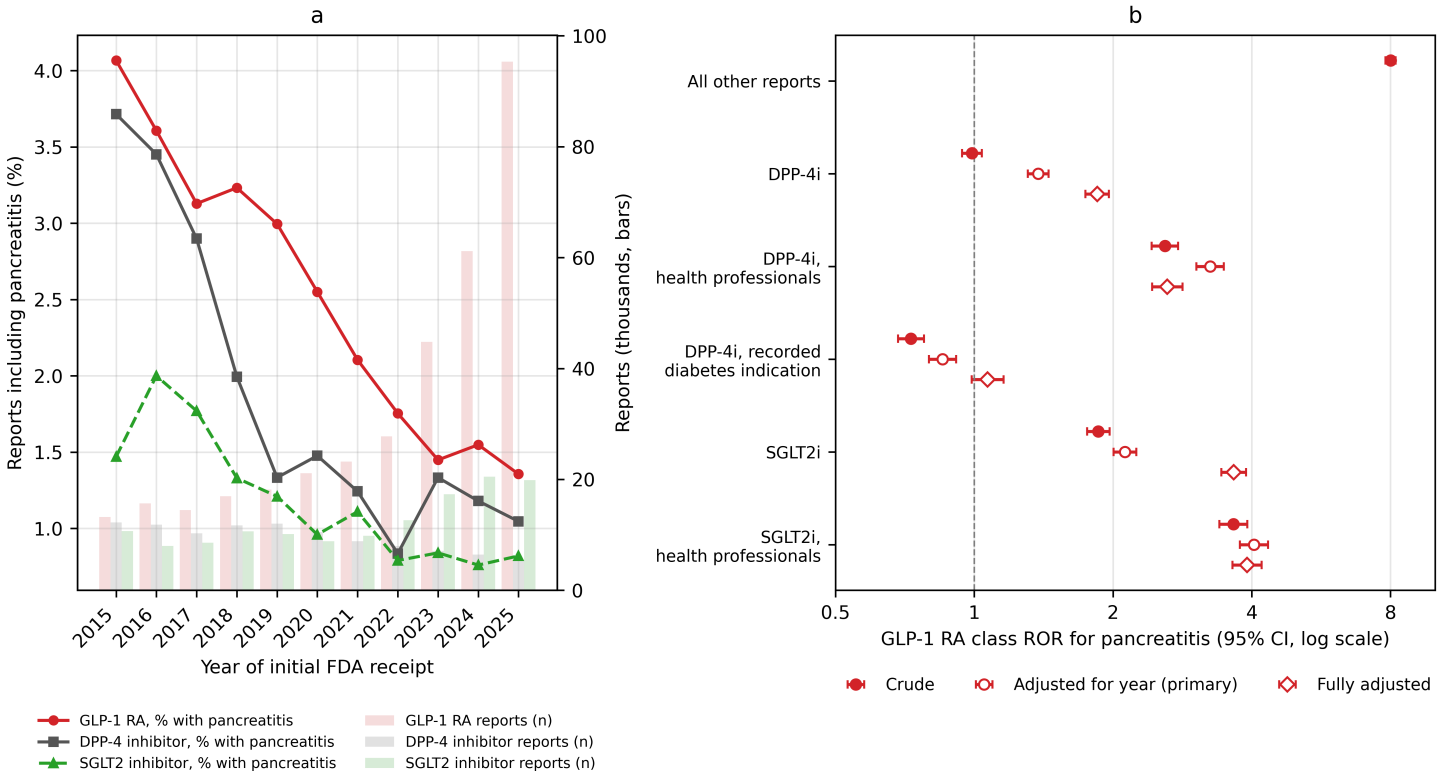


Figure 2

Pancreatitis reporting with GLP-1 receptor agonists versus DPP-4 and SGLT2 inhibitors

(a) Percentage of reports including pancreatitis by year of initial FDA receipt (lines; left axis) and number of reports (bars; right axis) for reports with a GLP-1 RA and no DPP-4 inhibitor, reports with a DPP-4 inhibitor and no GLP-1 RA, and reports with an SGLT2 inhibitor and no GLP-1 RA. (b) GLP-1 RA class RORs for pancreatitis (95% CI, logarithmic scale) against all other reports, against DPP-4 inhibitors (all reports, health-professional reports, and reports with a recorded diabetes indication), and against SGLT2 inhibitors (all reports and health-professional reports), crude (filled circles), adjusted for year (open circles), and fully adjusted for year, age group, sex, reporter type, and region (open diamonds). Results for the gallbladder/biliary positive control are given in Additional file 1. DPP-4i dipeptidyl peptidase-4 inhibitor, ROR reporting odds ratio, SGLT2i sodium-glucose cotransporter-2 inhibitor.

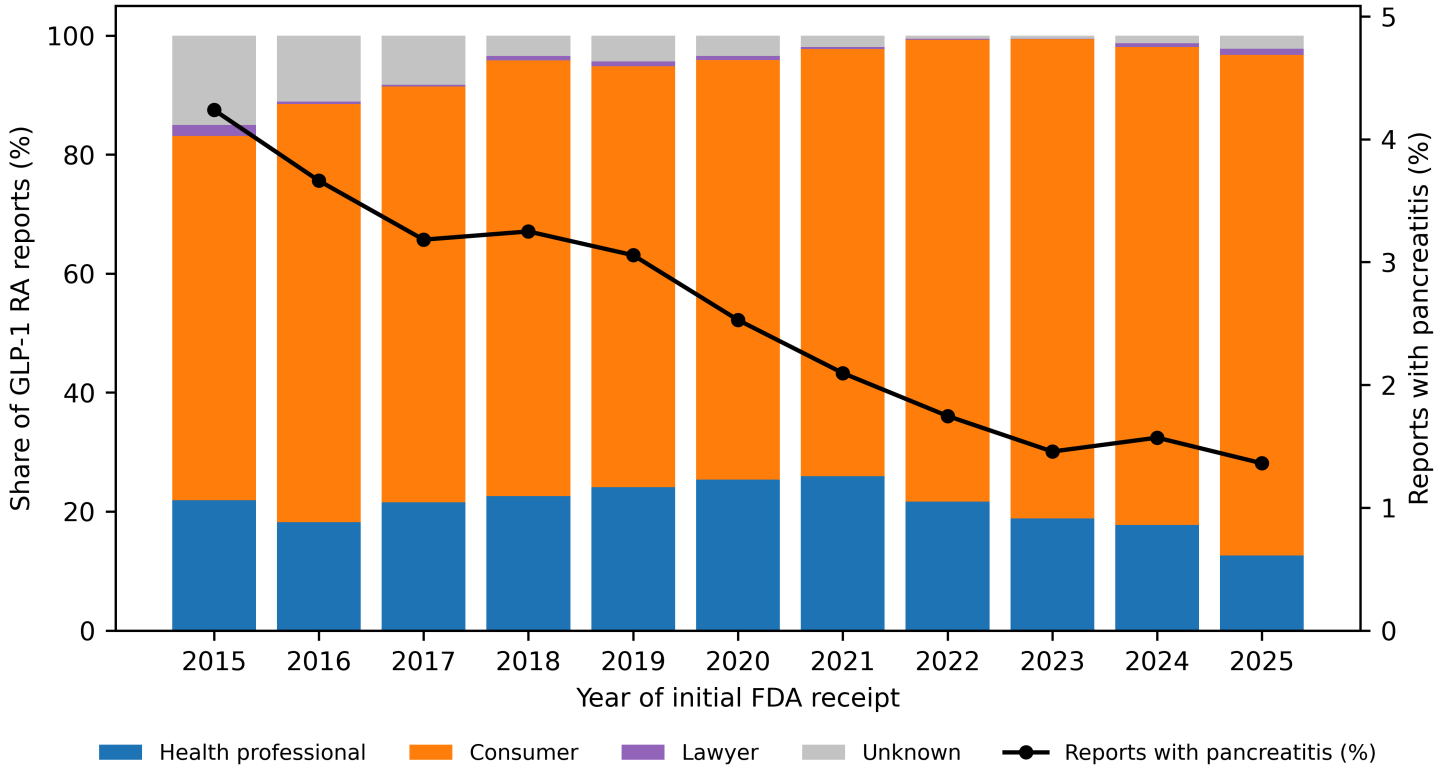


Figure 3

Reporter type of GLP-1 receptor agonist reports and proportion including pancreatitis, 2015–2025

Bars show the share of GLP-1 RA reports by reporter type in each year of initial FDA receipt; the line shows the percentage of GLP-1 RA reports that included a pancreatitis-related preferred term.

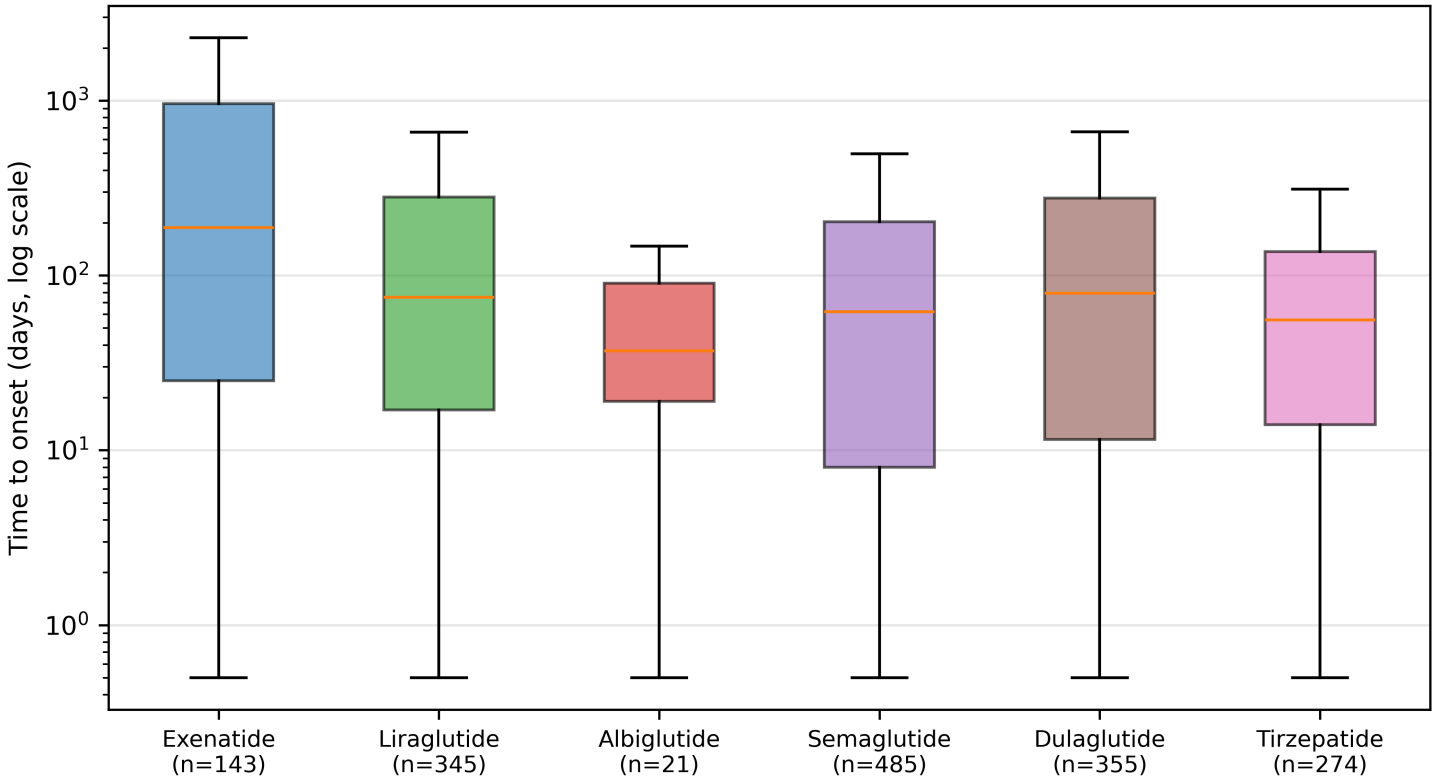


Figure 4

Time to onset of pancreatitis after GLP-1 receptor agonist initiation

Box plots of time from the earliest therapy start date to the event date (days, logarithmic scale; same-day onsets plotted at 0.5 days) for agents with at least 10 reports with complete dates. Boxes show the IQR and median; whiskers extend to 1.5 × IQR, and outlying values are not shown. Albiglutide (n = 21) is shown although it is not included in Table 3. IQR interquartile range.

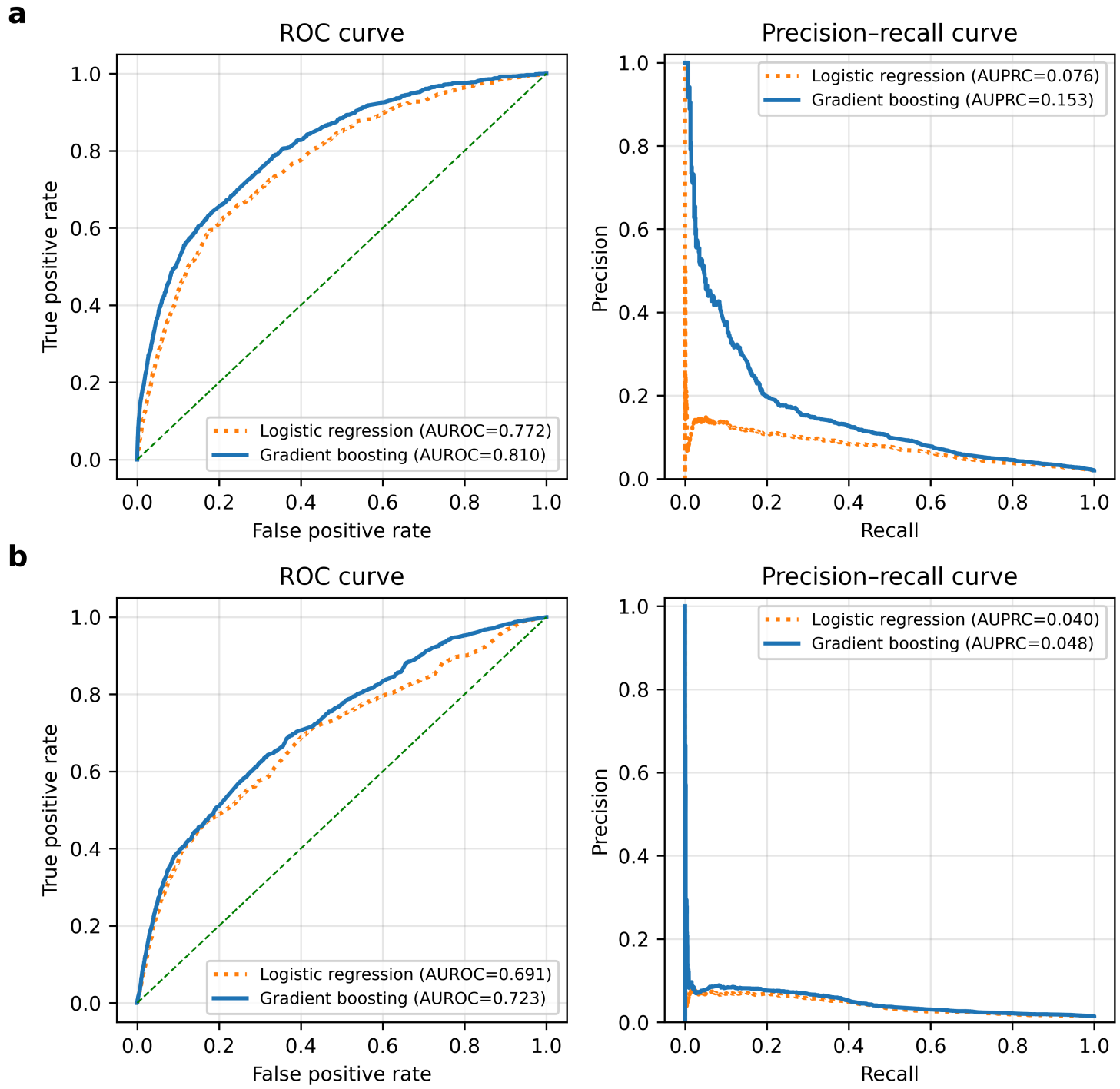


Figure 5

Discrimination of the primary pancreatitis report prioritization models

(a) ROC and precision–recall curves in the stratified random 80/20 train–test split. (b) Temporal-validation ROC and precision–recall curves for models trained on 2015–2022 reports and tested on 2023–2025 reports. Models used intake-type variables extracted from the latest case version. Dashed diagonal lines show chance performance (random ranking). AUPRC area under the precision–recall curve, AUROC area under the receiver operating characteristic curve, ROC receiver operating characteristic.

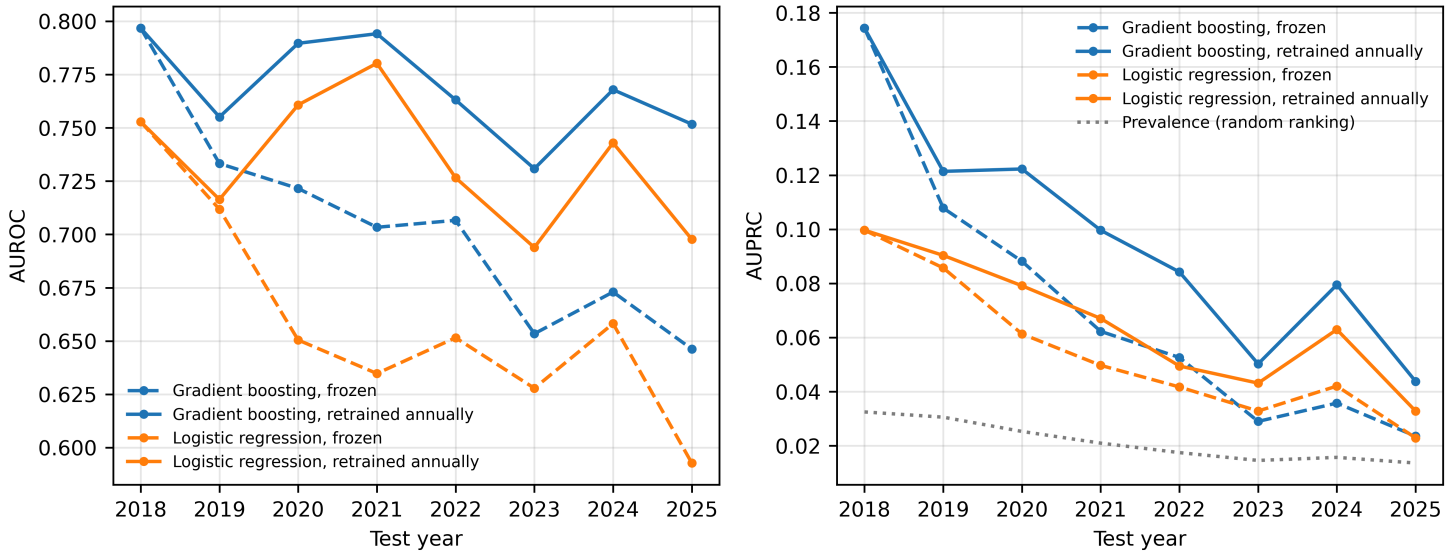


Figure 6

Annual retraining versus a frozen model for pancreatitis report prioritization

AUROC (left) and AUPRC (right) by test year for models retrained each year on all earlier reports (solid lines) and for models trained once on 2015–2017 reports (dashed lines); results for retraining on the three preceding years are given in Additional file 1. The dotted line shows the prevalence of pancreatitis, which equals the AUPRC of random ranking. AUPRC area under the precision–recall curve, AUROC area under the receiver operating characteristic curve.

Supplementary Files

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

- Additionalfile1SupplementaryMaterial.docx