Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02234-5

Strengthening inferential studies in the FDA Sentinel initiative: results from a methodological demonstration project

Check for updates

Rishi J. Desai[1] ✉, Janick Weberpals[1], Haritha Pillai[1], Adebola Ajao[2], Mukund Desibhatla[3], Rebecca Hawrusik[3], José J. Hernández-Muñoz[2], Chanelle Jones[2], Jamal T. Jones[2], Joyce Lii[1], Jie Li[2], Jennifer G. Lyons[4], Ryan Schoeplein[3], Fatma M. Shebl[2], Sengwee Toh[4], Elisabetta Patorno[1] & Sebastian Schneeweiss[1]

The FDA Sentinel Real-World Evidence Data Enterprise (RWE-DE) contains linked electronic health records with claims data for over 25 million patients. To demonstrate the applicability of the RWE-DE to a study previously considered infeasible in claims-based Sentinel network, we emulated a target trial using a case study comparing acute pancreatitis among new users of sodium-glucose cotransporter-2 inhibitors (SGLT-2i) with new users of dipeptidyl peptidase-4 inhibitors (DPP-4i) among 97,119 patients with type 2 diabetes mellitus from HealthVerity [2018–2020] and TriNetX [2013–2024] databases. We applied a previously validated computable phenotyping algorithm using EHRs to identify acute pancreatitis as the primary outcome. After confounding adjustment for >135 variables using propensity score fine stratification weighting, the hazard ratio (95% confidence interval) for acute pancreatitis following SGLT-2i initiation compared to DPP-4i initiation was 0.85 (0.67–1.07) for intent-to-treat and 0.84 (0.58–1.22) for per-protocol analysis. This study serves as a proof-of-concept for future safety assessments in Sentinel.

The U.S. FDA's Sentinel System forms a critical component of the national active post-marketing surveillance of medical products[1]. Historically, Sentinel's reliance on insurance claims data has led to insufficiency in addressing some emerging safety questions requiring more granular clinical information[2]. The FDA Sentinel Real-World Evidence Data Enterprise (RWE-DE), a data infrastructure linking large volumes of electronic health records (EHRs) with claims data, was created to help the FDA address emerging safety questions for which claims data may be insufficient[3,4].

A common scenario where EHR linkage can be particularly helpful is when certain outcomes of interest may not be captured in administrative claims. For instance, when assessing the suitability of evaluating the potential risk of acute pancreatitis in Sentinel, the FDA considered that claims-based diagnosis codes for acute pancreatis may have poor positive predicted value (PPV)[5], which was confirmed in a later validation study to be in the range of 55-66%[6]. Due to the potential for outcome misclassification that may have led to an underestimation of the effect, outcome ascertainment in linked EHR-claims data was proposed as an alternative. In this report, we describe results from a demonstration project that aimed to assess the applicability of the RWE-DE in a use case of the risk of acute pancreatitis following initiation of SGLT-2 inhibitors compared to dipeptidyl peptidase-4 inhibitors (DPP-4i) in patients with type 2 diabetes mellitus (T2DM). Specifically, we address the limitation of low PPV for pancreatitis in claims-based diagnosis codes by deploying a previously developed computable phenotyping algorithm using elements from EHRs in addition to claims diagnosis codes, which is reported to have a PPV of >90%[7]. Additionally, we also outline typical workflow of inferential studies utilizing EHR-claims linked data in Sentinel and provide readily usable analytic codes as a turnkey solution for conducting rigorous analyses in a timely way for future needs of the program.

Results

Study cohorts

After applying all eligibility criteria, the final cohorts included 72,429 patients from HealthVerity (30,174 SGLT-2i initiators; 42,155 DPP-4i

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [2]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA. [3]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA, USA. [4]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA. ✉e-mail: rdesai@bwh.harvard.edu

initiators) and 24,690 patients from TriNetX (11,943 SGLT-2i initiators; 12,747 DPP-4i initiators).

Table 1 summarizes the key patient characteristics of the included participants. For both data sources, the average age was lower in SGLT-2i group than DPP-4i group (57 years versus 60 years in HealthVerity; 55 years versus 56 years in TriNetX). Overall indicators of health including CFI and CCI were comparable between treatment groups in both data sources. Metformin was the most common comedication in both treatment groups across both data sources. The mean count of antidiabetic drug classes was generally comparable (HealthVerity: 1.4 ± 0.8 in the SGLT-2i group and 1.3 ± 0.8 in the DPP-4i; TriNetX: 1.2 ± 0.8 for the SGLT-2i group and 1.1 ± 0.8 for the DPP-4i group). For HealthVerity, the proportion of patients with myocardial infarction (1.7%, 1.1%), stable angina (4.3%, 3.7%), and heart failure (6%, 5.9%) was similarly distributed between the two treatment groups; while for TriNetX, the proportion was generally higher in the SGLT-2i groups for myocardial infarction (3.7%, 1.2%), stable angina (4.8%, 2.3%), and heart failure (13.3%, 5.7%), likely reflecting increasing use of SGLT-2i in more recent time periods after knowledge of their cardiovascular benefits accumulated.

Structural missing data investigation

Missingness in EHR-based variables was commonly observed as expected (Table 1). In HealthVerity, HbA1c results were available for 29.4% SGLT-2i initiators and 27.3% DPP-4i initiators (mean HbA1c: 8.7 ± 1.9, 8.6 ± 1.9). Serum creatinine and triglyceride levels were recorded for around 20–25% of the patients (mean serum creatinine: 0.9 ± 0.5, 0.9 ± 0.5; mean triglycerides: 176 ± 90.6; 171.4 ± 87.4). BMI was recorded for >60% of the patients in both groups (mean BMI: 32.4 ± 5.5, 31.6 ± 5.7). Blood pressure was recorded for >80% of the patients in both groups (mean systolic: 131.3 ± 16.5, 131.3 ± 16.8; mean diastolic: 79 ± 10.2; 78.3 ± 10.3). Total number of EHR encounters were comparable across both groups (mean EHR encounters: 3.4 ± 2.8, 3.5 ± 2.9). In TriNetX, HbA1c results were available for around 50% of the patients in both groups (mean HbA1c: 8.6 ± 2.0, 8.5 ± 1.9). Serum creatinine and triglyceride levels were recorded for 35–60% of the patients (mean serum creatinine: 0.9 ± 0.3, 0.9 ± 0.4; mean triglycerides: 174.2 ± 93.6; 174.3 ± 92.7). BMI was recorded for approximately half of the patients in both groups (mean BMI: 34.8 ± 8.0, 34.5 ± 7.9). Blood pressure was recorded for >60% of the patients in both groups (mean systolic: 134.6 ± 19.6, 134.2 ± 19.0; mean diastolic: 79.8 ± 12.2; 79.5 ± 11.6).

We observed a monotonic missingness pattern in EHR-based variables for both data sources, as patients with missing data for these key variables are likely to exhibit consistent gaps in other EHR-based measurements (Supplementary Figs. 1 and 2). For the missingness diagnostics (Supplementary Table 3), we observed differences in measured variables between those with and without missing data for EHR-based variables as seen by absolute standardized mean distribution, with medians of around 0.02–0.05. Next, for models predicting missingness, we observed relatively high area under the curve (AUCs) for each of these variables, especially in TriNetX. High AUCs suggest that missing at random (MAR) conditional on measured information may be a likely missingness mechanism. Finally, we evaluated associations between missingness indicator in each of these EHR variables and the outcome (acute pancreatitis outcome of interest). These results indicated that when adjusting for other measured variables, no significant association was observed between missingness indicator and the outcome. This observation provides some reassurance against missing not at random (MNAR) mechanism. Overall, we concluded that MAR may be a reasonable assumption regarding a missingness mechanism for these variables and therefore, multiple imputations are likely to provide best bias-variance trade-off[8].

Acute pancreatitis risk

Table 2 shows a comparison of incidence rates (IRs) of acute pancreatitis in new users of SGLT-2i and DPP-4i in HealthVerity and TriNetX. The total event count was 236 and 138 in HealthVerity and was 107 and 41 in TriNetX, for ITT and PP causal contrasts, respectively. Overall, we observed event rates in the range of two to three per 1000 person-years across both data sources in the two follow-up schemes. Figure 1 compares the unadjusted cumulative incidence of acute pancreatitis in new users of SGLT-2i versus DPP-4i with T2DM, including both ITT and PP analyses in both data sources. Overall, the plots suggested that the cumulative incidence of acute pancreatitis was comparable between the two treatment groups for both ITT and PP approach in both data sources.

Supplementary Figs. 3 and 4 show high covariate balance across all covariates after the PS weighting procedure, reported as mean difference range between two treatment groups across 20 imputations. Figure 2 provides the hazard ratios (HRs) for acute pancreatitis in SGLT-2i initiators compared to DPP-4i initiators with T2DM in HealthVerity and TriNetX. In the claims-only analysis, the pooled adjusted HRs were 0.99 (0.84–1.16) for ITT and 0.94 (0.73–1.20); which notably moved downwards in the claims + EHR augmented analysis [pooled HR 0.85 (95% CI: 0.67–1.07) for ITT and 0.84 (0.58–1.22) for PP].

Subgroup analyses and robustness evaluation

For all subgroups considered (age <65 and ≥65, males, females, and history of acute pancreatitis risk factors), we found results consistent with the overall population (Fig. 3). In the two sensitivity analyses where we attempted to reduce missingness proportions for EHR-based covariates by increasing the lookback period and by restricting to those with ≥3 EHR encounters, we noted that the capture increased for all EHR-based covariates (Supplementary Fig. 5) and results did not change meaningfully from the primary analysis. The analysis of the stroke endpoint as a negative control outcome confirmed no difference in risk between SGLT-2i and DPP-4i (pooled HR 0.97, 95% CI 0.82–1.14).

Discussion

The primary contribution of this study is that it provides an insight into workflows and salient challenges for complex inferential analyses in the Sentinel System moving forward where diverse data types including insurance claims and EHRs from numerous data sources are expected to be routinely used. First and foremost, as soon as a safety concern arises regarding a medical product, it will be crucial to identify data elements that are essential for a given study to ensure fitness-for-purpose of the data sources. For instance, in this study, the need for additional clinical data including amylase and lipase test results to reliably identify acute pancreatitis drove the choice of utilizing the claims-EHR linked RWE-DE, which is relatively smaller with approximately 25 million total covered lives between the two sources used in this study, over the insurance claims-based Sentinel Distributed Database, which is much larger with 128.7 million members currently accruing new data. Next, availability of validated algorithms for complex outcomes that are insufficiently defined by diagnosis codes alone but are possible to identify with computable phenotyping will be critical to deploy in a scalable and timely way. Sentinel has made steady progress in this area with development and validation of numerous computable phenotypes that are likely to be of high interest as safety outcomes in the future including anaphylaxis, suicidal ideation, and sleep-related behaviors[9,10]. Finally, use of EHR data opens the possibility of more robust confounding adjustment for elements that are traditionally not captured by claims data such as BMI, vital signs, and laboratory test results. However, pervasive missingness in these variables in real-world data sources remain a challenge and appropriate methods to diagnose likely mechanisms and analytically correct missingness will continue to be of vital importance. Sentinel has made substantial advances in this area as well with development of reusable analytic tools and methods designed specifically to handle data missingness in EHR-based variables[8,11,12]. In this study, we were able to demonstrate use of these tools in routine analytic workflow in an efficient way.

In this use case of a large study involving more than 97,000 patients from the FDA Sentinel RWE-DE commercial network, the incidence of acute pancreatitis was low, and we did not observe evidence for a statistically significant difference in acute pancreatitis risk following initiation of SGLT-

Table 1 | Select patient characteristics before propensity score adjustment in the study cohort of patients with Type 2 diabetes mellitus initiating SGLT-2 inhibitors or DPP-4 inhibitors

	HealthVerity				TriNetX			
	(January 2018–December 2020)				(January 2013–February 2024)			
	SGLT-2i initiators		DPP-4i initiators		SGLT-2i initiators		DPP-4i initiators	
Patient characteristics	**N/mean**	**%/Std deviation**	**Number/ Mean**	**%/Std deviation**	**Number/ Mean**	**%/Std deviation**	**Number/ Mean**	**%/Std deviation**
Unique Patients	30,174	N/A	42,255	N/A	11943	N/A	12,747	N/A
Demographic characteristics								
Age (years)	56.9	11.1	59.6	12.9	55.4	11.4	55.6	11.5
Sex								
Female	14,634	48.5	23,106	54.7	5743	48.1	6521	51.2
Male	15,540	51.5	19,149	45.3	6200	51.9	6226	48.8
Health characteristics								
Claims-Based Frailty Index	0.1	0	0.2	0	0.2	0	0.2	0
Combined comorbidity Index	1.2	1.8	1.4	2	1.5	2.1	1.2	2
Prior Metformin users	22,764	75.4	29,922	70.8	7894	66.1	7792	61.1
Current Metformin users	19,907	66	30,588	72.4	6941	58.1	8469	66.4
Prior Sulfonylureas users	9770	32.4	15,940	37.7	2885	24.2	3562	27.9
Current Sulfonylureas users	8203	27.2	14,532	34.4	2427	20.3	3247	25.5
Prior Insulin users	7168	23.8	7271	17.2	2607	21.8	1898	14.9
Current Insulin users	6249	20.7	6457	15.3	2278	19.1	1737	13.6
ACE inhibitors/ARBs	20,899	69.3	29,163	69	7716	64.6	7484	58.7
Beta Blockers	10,570	35	14,594	34.5	4305	36	3702	29
Calcium Channel Blockers	7054	23.4	10,836	25.6	3097	25.9	2866	22.5
Hypertension	22,724	75.3	31,973	75.7	9309	77.9	9606	75.4
Hyperlipidemia	21,737	72	29,063	68.8	8533	71.4	8749	68.6
Myocardial Infarction	510	1.7	470	1.1	442	3.7	155	1.2
Heart Failure	1813	6	2498	5.9	1589	13.3	725	5.7
Diabetic Nephropathy	2805	9.3	5118	12.1	1275	10.7	965	7.6
Diabetic Neuropathy	5844	19.4	8748	20.7	1949	16.3	1714	13.4
Diabetic Retinopathy	3072	10.2	4571	10.8	855	7.2	658	5.2
Type 2 Diabetes Mellitus without Mention of Complications	23,235	77	32,988	78.1	9071	76	8275	64.9
Type 2 Diabetes mellitus with Unspecified Complications	2401	8	3509	8.3	755	6.3	726	5.7
History of Autoimmune Diseases	1557	5.2	2088	4.9	569	4.8	567	4.4
Gallstones	390	1.3	586	1.4	186	1.6	195	1.5
Hypertriglyceridemia	1248	4.1	1420	3.4	380	3.2	377	3
Health service utilization intensity metrics								
Mean number of ambulatory encounters	8.9	9.2	8.5	9.8	8.9	9.5	8.3	8.7
Mean number of emergency room encounters	0.3	0.9	0.4	1	0.7	1.6	0.7	1.6
Mean number of inpatient hospital encounters	0.1	0.9	0.2	0.9	0.9	3.5	1	4.2
Mean number of non-acute institutional encounters	0.1	0.6	0.1	1	0.1	2	0.2	2.4
Mean number of other ambulatory encounters	4.4	19.7	6.5	24.8	6.1	18.1	5.7	17.4
Mean number of filled prescriptions	26	21.2	26.5	22.1	23.8	20.3	22.5	20.9
Mean number of generics dispensed	10.3	5.7	10.6	5.9	10.2	5.8	9.5	5.8
Count of antidiabetic medications	1.4	0.8	1.3	0.8	1.2	0.8	1.1	0.8
Laboratory test results								
Hemoglobin A1c								
Test record in PERCENT	8874	29.4	11518	27.3	5802	48.6	6560	51.5

Table 1 (continued) | Select patient characteristics before propensity score adjustment in the study cohort of patients with Type 2 diabetes mellitus initiating SGLT-2 inhibitors or DPP-4 inhibitors

	HealthVerity				TriNetX			
	(January 2018–December 2020)				(January 2013–February 2024)			
	SGLT-2i initiators		DPP-4i initiators		SGLT-2i initiators		DPP-4i initiators	
Patient characteristics	N/mean	%/Std deviation	Number/Mean	%/Std deviation	Number/Mean	%/Std deviation	Number/Mean	%/Std deviation
Mean, standard deviation	8.7	1.9	8.6	1.9	8.6	2	8.5	1.9
No test record	21,300	70.6	30,737	72.7	6141	51.4	6187	48.5
Serum Creatinine								
Test record in MG/DL	7298	24.2	10,020	23.7	6364	53.3	7377	57.9
Mean, standard deviation	0.9	0.5	0.9	0.5	0.9	0.3	0.9	0.4
No test record	22,876	75.8	32,235	76.3	5579	46.7	5370	42.1
Triglycerides								
Test record in MG/DL	5955	19.7	8109	19.2	4152	34.8	4877	38.3
Mean, standard deviation	176	90.6	171.4	87.4	174.2	93.6	174.3	92.7
No test record	24,219	80.3	34,146	80.8	7791	65.2	7870	61.7
Vitals/Lifestyle factors								
Body Mass Index (BMI)								
Recorded in kg/m^2	18,326	60.7	26,239	62.1	5950	49.8	6055	47.5
Mean, standard deviation	32.4	5.5	31.6	5.7	34.8	8	34.5	7.9
Not recorded	11,848	39.3	16,016	37.9	5993	50.2	6692	52.5
Diastolic Blood Pressure (DBP)								
DBP recorded in mmHg	24,896	82.5	34,932	82.7	7884	66	7830	61.4
Mean, standard deviation	79	10.2	78.3	10.3	79.8	12.2	79.5	11.6
No test record	5278	17.5	7323	17.3	4059	34	4917	38.6
Systolic Blood Pressure (SBP)								
SBP recorded in mmHg	24,896	82.5	34,933	82.7	7834	65.6	7752	60.8
Mean, standard deviation	131.3	16.5	131.3	16.8	134.6	19.6	134.2	19
No test record	5278	17.5	7323	17.3	4109	34.4	4995	39.2
Tobacco Use								
Recorded as Yes	4384	14.5	5663	13.4	1535	12.9	1608	12.6
Recorded as No	7588	25.1	11,807	27.9	0	0	0	0
Not recorded	18,202	60.3	24,785	58.7	10,408	87.1	11,139	87.4
EHR Encounters								
Total number of encounters	3.4	2.8	3.5	2.9	3.9	4.7	3.9	4.8

2 inhibitors compared to DPP-4 inhibitors in patients with T2DM. We noted that when using claims-only diagnosis codes for acute pancreatitis that are known to have poor PPV (0.55–0.65), the estimates were closer to the null versus when using Sentinel's phenotyping algorithm for pancreatitis with a PPV of 0.90. These observations are consistent with the general expectations that non-differential misclassification of the outcome, as likely observed here with the claims-based definition, results in a bias towards the null and could result in masking of important differences in the outcome risk between exposures[13]. Future investigations focused on the outcome of acute pancreatitis should be wary of this potential bias when using claims-based definition. Our findings were consistent across two data sources and across subgroups of age, sex, and acute pancreatitis risk with and without additional adjustment for EHR-derived covariates. This study was conducted in collaboration with the FDA as a methodological demonstration and the findings of this study should be assessed considering the totality of the evidence and not this individual result. Further, direct interpretation of this comparison is inherently challenging as the comparator (DPP4-i) carries a label for acute pancreatitis.

As with any observational investigation, our study has important limitations. First, despite adjusting for many covariates, residual confounding may still be present as treatment decisions are made by treating physician non-randomly and the factors that influence treatment decisions cannot be readily assessed even using RWE-DE. While EHR data allowed us to capture laboratory test results and offered enhanced confounding adjustment, we also observed large missingness in capture of some of these results such as triglycerides which is an important risk factor for acute pancreatitis. As a result, residual confounding by factors that either not recorded or incompletely recorded is possible. Second, while the RWE-DE is one of the largest claims-EHR network constructed to be used for post-marketing safety surveillance purposes in the U.S, richer data from the claims-EHR linkages are obtained at the expense of a substantially smaller total available population size than claims-only networks. Therefore, when investigating rare adverse events such as acute pancreatitis, lack of precision may present challenges in drawing conclusions. As such, it should be noted that the conclusion of no difference in acute pancreatitis risk between SGLT2i and DPP4i in the primary and subgroup analyses in this study may reflect insufficient power for detecting differences small effect size power rather than definitive equivalence. Future work integrating more data sources may be helpful in increasing the population size for the RWE-DE. Finally, validated computable phenotyping algorithms may show

Table 2 | Crude incidence rates of acute pancreatitis among patients with Type 2 diabetes mellitus initiating SGLT-2 inhibitors or DPP-4 inhibitors

Data source	Treatment group	Measure	Intent to treat follow-up (as-started)	Per protocol follow-up (on-treatment)
HealthVerity (Jan 2018–Dec 2020)	SGLT-2i initiators (*n* = 30,174)	Number of events/person years	88/33,889	40/16,374
		Incidence rate/1000 person years	2.6 (2.1–3.2)	2.4 (1.7–3.3)
	DPP-4i initiators (*n* = 42,255)	Number of events/person years	148/51,561	67/24,608
		Incidence rate/1000 person years	2.9 (2.4–3.4)	2.7 (2.1–3.5)
TriNetX (Jan 2013–Feb 2024)	SGLT-2i initiators (*n* = 11,943)	Number of events/person years	44/22,756	15/7,891
		Incidence rate/1000 person years	1.9 (1.4–2.6)	1.9 (1.1–3.1)
	DPP-4i initiators (*n* = 12,747)	Number of events/person years	94/36,783	26/10,499
		Incidence rate/1000 person years	2.6 (2.1–3.1)	2.5 (1.6–3.6)

performance degradation at different sites. However, it is often infeasible to develop and validate algorithms separately in each data source in near real time when investigating safety of medications in a large scale national post-marketing surveillance program like Sentinel. When resources permit, a smaller scale validation study using manual review of data from newer sites where the algorithm was applied may be considered.

In conclusion, the successful completion of this study in FDA Sentinel's RWE-DE commercial network serves as a proof-of-concept for future protocol-based assessments in Sentinel requiring EHR data. Analytic pipelines and packages developed by the FDA Sentinel System provide key building blocks to achieve scalable and timely execution of complex analyses using claims-EHR linked data assets.

Methods

Data sources

We used data from the FDA Sentinel RWE-DE commercial network comprising two data partners—HealthVerity and TriNetX. HealthVerity included ambulatory care EHRs from three sources linked to closed medical claims and pharmacy data from 2018 through 2020 while TriNetX included inpatient and ambulatory care EHRs from 20 unique health care organizations (HCOS) linked to closed claims data for the period of 2013–2024[3].

Specification and emulation of the target trial

We leveraged the "PRocess guide for INferential studies using healthcare data from routine ClinIcal Practice to evaLuate causal Effects of Drugs (PRINCIPLED)" framework[14], established specifically to guide conduct of inferential studies in Sentinel, for the proposed study. First, we developed the causal question by specifying a target trial protocol comparing the risk of acute pancreatitis in SGLT-2i initiators to DPP-4i (Table 3). We identified DPP-4i as a comparator group as they are also frequently used as second-line treatment for T2DM and may serve as realistic clinical comparators to SGLT-2i. Of note, the DPP-4i prescribing information includes a Warnings and Precaution for acute pancreatitis based upon post-marketing data and imbalances in clinical trials[15]. Next, the emulation of each component of the target trial protocol was described using fit-for-purpose linked claims-EHR data with exposure information coming from pharmacy claims and outcome and confounder information from both claims and EHR data as described below. The study protocol was publicly posted on the Sentinel website before the analysis began[16]. Key components of the target trial protocol are described below.

Eligibility criteria. Cohort entry was the day of first dispensing of either a SGLT-2i or DPP-4i. The eligibility criteria for the target trial included presence of T2DM diagnosis, no use of study medications, no history of end stage renal disease (ESRD), HIV, or acute pancreatitis, and no use of glucagon-like peptide-1 receptor agonists (GLP-1 RAs) within six months before study medication start. We excluded users of GLP-1 RAs because they share a similar mechanism with DPP-4is and there remains uncertainty regarding the risk of pancreatitis after their use, with some studies suggesting increased risk[17,18]. Patients with a history of pancreatitis, ESRD, or HIV were excluded as these patients may have elevated risk of future acute pancreatitis events which may not be attributable to the treatment[19,20]. We further required 6 months of medical and prescription coverage before cohort entry with an allowable enrollment gap of up to 30 days as well as at least one EHR encounter to ensure that patients had observable time in the data. This requirement ensured that patients had contact with the healthcare system to allow for adequate capture of clinical codes to measure eligibility criteria and baseline covariates.

Treatment strategies and follow-up. The two treatment strategies comprised initiation of SGLT-2i (canagliflozin, dapagliflozin, empagliflozin, ertugliflozin, bexagliflozin) or DPP-4i (alogliptin, linagliptin, saxagliptin, sitagliptin) assessed based on pharmacy dispensing data. We estimated observational analogues of the intent-to-treat (ITT or as-started) and per-protocol (PP or on-treatment) effect. Accordingly,

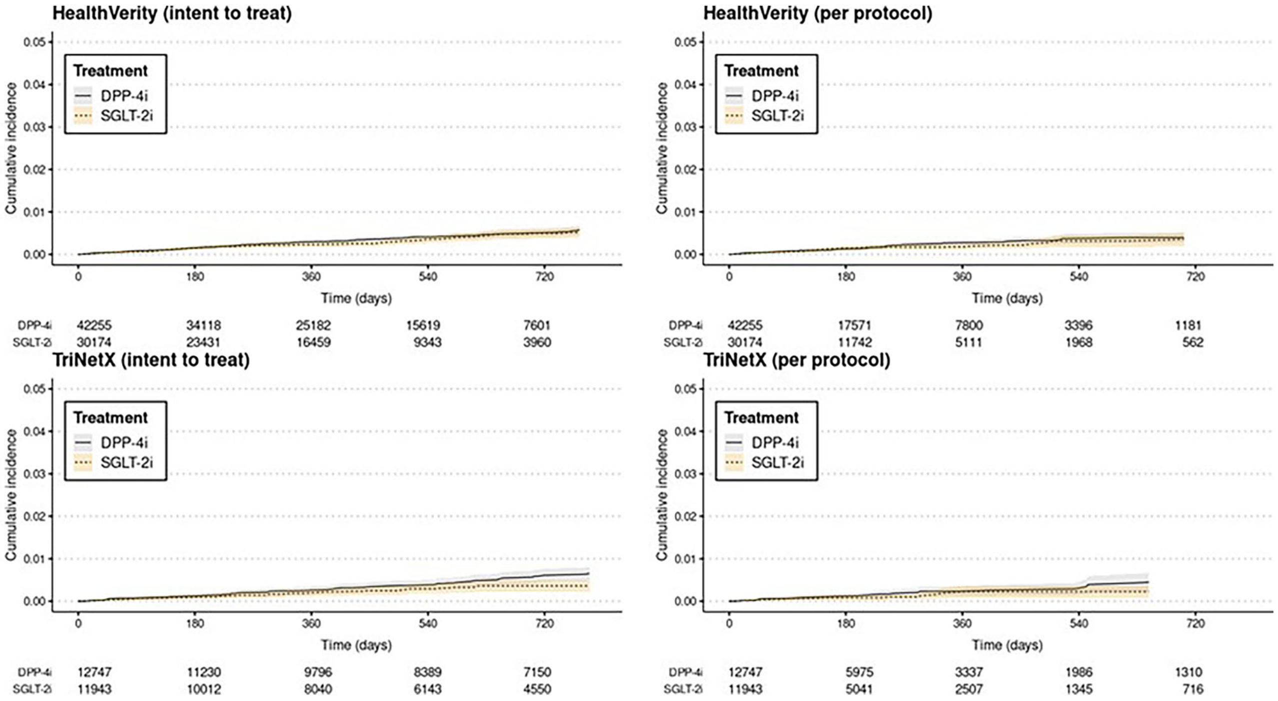
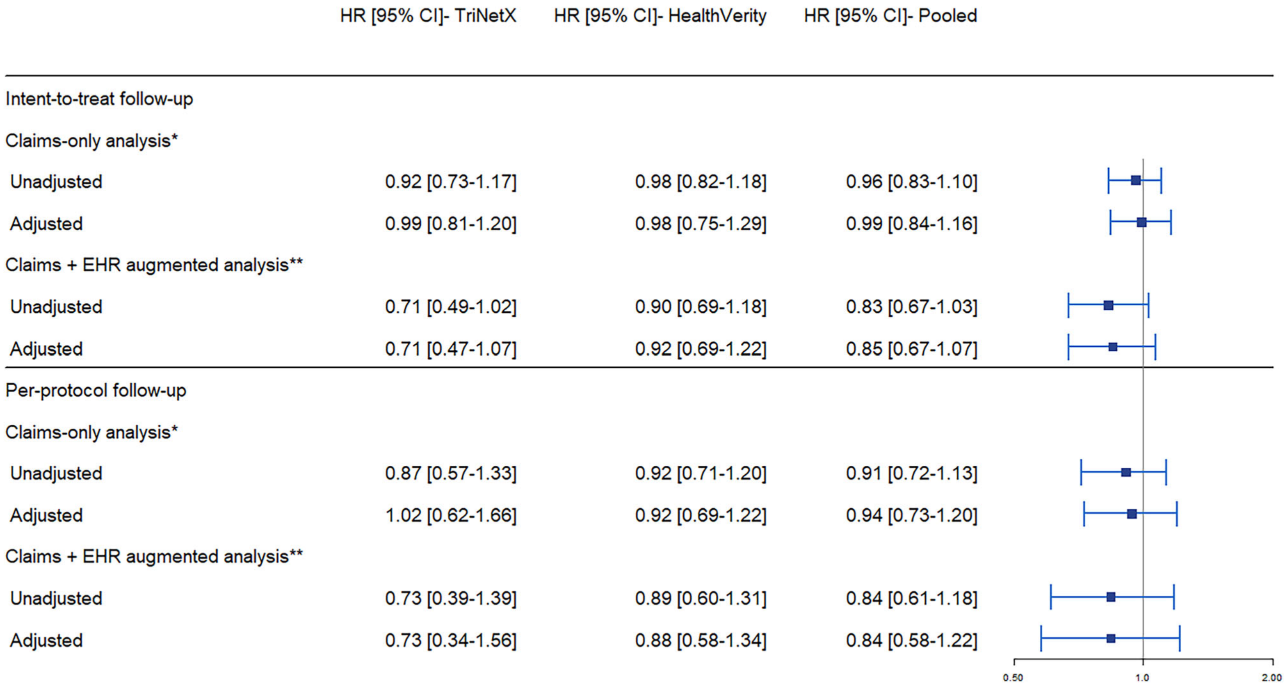


Fig. 1 | Cumulative incidence of acute pancreatitis before propensity score adjustment in the study cohort of patients with Type 2 diabetes mellitus initiating SGLT-2 inhibitors or DPP-4 inhibitors.

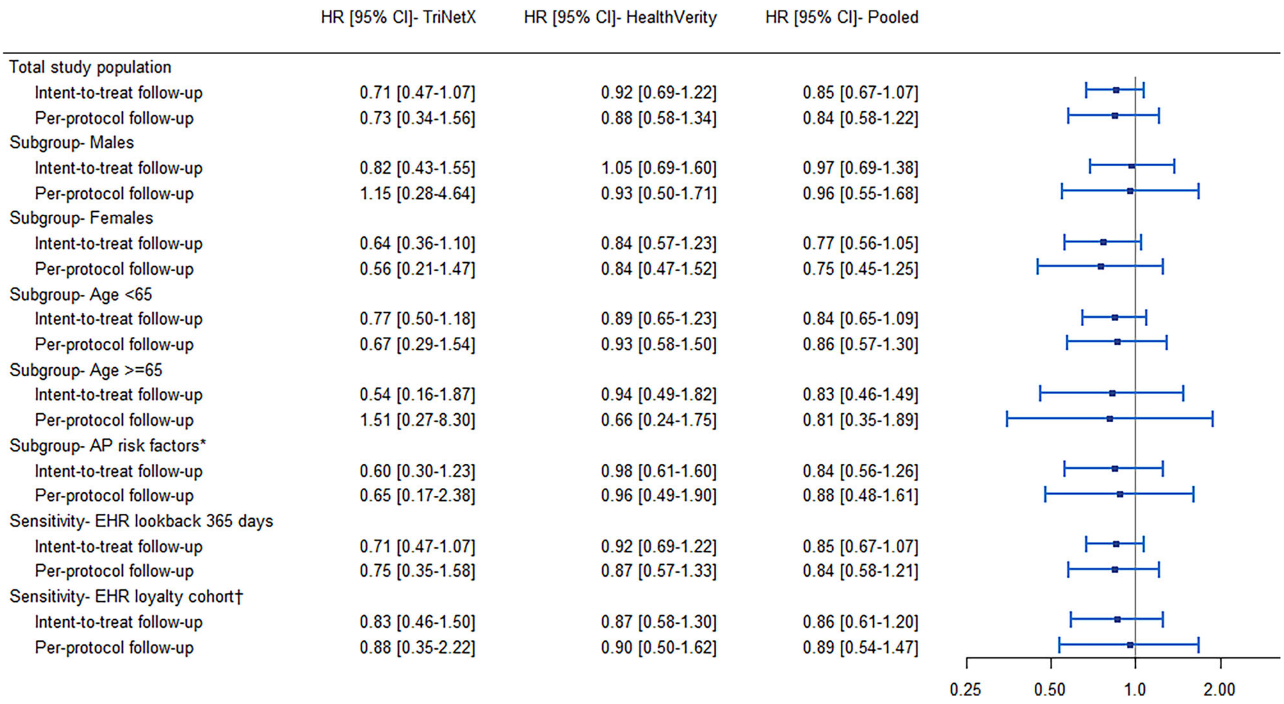


	HR [95% CI]- TriNetX	HR [95% CI]- HealthVerity	HR [95% CI]- Pooled
Intent-to-treat follow-up			
Claims-only analysis*			
Unadjusted	0.92 [0.73-1.17]	0.98 [0.82-1.18]	0.96 [0.83-1.10]
Adjusted	0.99 [0.81-1.20]	0.98 [0.75-1.29]	0.99 [0.84-1.16]
Claims + EHR augmented analysis**			
Unadjusted	0.71 [0.49-1.02]	0.90 [0.69-1.18]	0.83 [0.67-1.03]
Adjusted	0.71 [0.47-1.07]	0.92 [0.69-1.22]	0.85 [0.67-1.07]
Per-protocol follow-up			
Claims-only analysis*			
Unadjusted	0.87 [0.57-1.33]	0.92 [0.71-1.20]	0.91 [0.72-1.13]
Adjusted	1.02 [0.62-1.66]	0.92 [0.69-1.22]	0.94 [0.73-1.20]
Claims + EHR augmented analysis**			
Unadjusted	0.73 [0.39-1.39]	0.89 [0.60-1.31]	0.84 [0.61-1.18]
Adjusted	0.73 [0.34-1.56]	0.88 [0.58-1.34]	0.84 [0.58-1.22]

Fig. 2 | Relative risk of acute pancreatitis before and after propensity score adjustment in the study cohort of patients with Type 2 diabetes mellitus initiating SGLT-2 inhibitors compared to DPP-4 inhibitors. *Claims only analysis defined acute pancreatitis based on ICD codes alone and adjusted for >130 claims-based variables **Claims + EHR augmented analysis defined acute pancreatitis based on a phenotyping algorithm using EHR data and added 6 additional variables for confounding adjustment with missing data imputed using multiple imputation.

follow-up began on the day after exposure initiation and continued until the first occurrence of any of the following: (1) outcome occurrence (acute pancreatitis); (2) health plan disenrollment; (3) recorded death; (4) end of available data; (5) discontinuation/switching from initiated treatment (only for PP analysis).

Outcome. The primary outcome measure was acute pancreatitis assessed using a validated phenotyping algorithm developed for use in Sentinel studies[7]. Briefly, the outcome was defined probabilistically conditional on information recorded in diagnosis codes and laboratory findings (amylase, lipase, triglycerides). Additionally, for TriNetX, features extracted



	HR [95% CI]- TriNetX	HR [95% CI]- HealthVerity	HR [95% CI]- Pooled
Total study population			
Intent-to-treat follow-up	0.71 [0.47-1.07]	0.92 [0.69-1.22]	0.85 [0.67-1.07]
Per-protocol follow-up	0.73 [0.34-1.56]	0.88 [0.58-1.34]	0.84 [0.58-1.22]
Subgroup- Males			
Intent-to-treat follow-up	0.82 [0.43-1.55]	1.05 [0.69-1.60]	0.97 [0.69-1.38]
Per-protocol follow-up	1.15 [0.28-4.64]	0.93 [0.50-1.71]	0.96 [0.55-1.68]
Subgroup- Females			
Intent-to-treat follow-up	0.64 [0.36-1.10]	0.84 [0.57-1.23]	0.77 [0.56-1.05]
Per-protocol follow-up	0.56 [0.21-1.47]	0.84 [0.47-1.52]	0.75 [0.45-1.25]
Subgroup- Age <65			
Intent-to-treat follow-up	0.77 [0.50-1.18]	0.89 [0.65-1.23]	0.84 [0.65-1.09]
Per-protocol follow-up	0.67 [0.29-1.54]	0.93 [0.58-1.50]	0.86 [0.57-1.30]
Subgroup- Age >=65			
Intent-to-treat follow-up	0.54 [0.16-1.87]	0.94 [0.49-1.82]	0.83 [0.46-1.49]
Per-protocol follow-up	1.51 [0.27-8.30]	0.66 [0.24-1.75]	0.81 [0.35-1.89]
Subgroup- AP risk factors*			
Intent-to-treat follow-up	0.60 [0.30-1.23]	0.98 [0.61-1.60]	0.84 [0.56-1.26]
Per-protocol follow-up	0.65 [0.17-2.38]	0.96 [0.49-1.90]	0.88 [0.48-1.61]
Sensitivity- EHR lookback 365 days			
Intent-to-treat follow-up	0.71 [0.47-1.07]	0.92 [0.69-1.22]	0.85 [0.67-1.07]
Per-protocol follow-up	0.75 [0.35-1.58]	0.87 [0.57-1.33]	0.84 [0.58-1.21]
Sensitivity- EHR loyalty cohort†			
Intent-to-treat follow-up	0.83 [0.46-1.50]	0.87 [0.58-1.30]	0.86 [0.61-1.20]
Per-protocol follow-up	0.88 [0.35-2.22]	0.90 [0.50-1.62]	0.89 [0.54-1.47]

Fig. 3 | Results from subgroup and sensitivity analysis (propensity score adjusted estimates from claims + EHR augmented analyses). *Acute pancreatitis risk factors included gallstones, tobacco use, or alcohol abuse † EHR loyalty cohort analysis was restricted to subjects with ≥3 EHR encounters in the 6-months before cohort entry.

Table 3 | Target trial specification and emulation

Element	Specification of the hypothetical target trial	Emulation using real-world data sources
Eligibility Criteria	Patients with type 2 diabetes mellitus, no use of study medications before randomization, no history of end stage renal disease (ESRD), no history of HIV, no history of acute pancreatitis, no history of GLP-1 receptor agonist use	Same as target trial
	Continuous health plan enrollment and at least one recorded encounter in EHRs in 6 months prior to treatment initiation	
Treatment Strategies	1. Initiation SGLT-2 inhibitors (canagliflozin, dapagliflozin, empagliflozin, ertugliflozin, bexagliflozin)	Same as target trial
	2. Initiation of DPP-4 inhibitors (alogliptin, linagliptin, saxagliptin, sitagliptin)	
Treatment Assignment	Randomized non-blinded	Non-blinded and assumed to be randomized within levels of measured confounders
Follow-Up Start (Time 0)	At assignment	Same as target trial
Follow-Up End	First of administrative end of follow-up (most recent data), loss to follow-up, death, or outcome occurrence	Same as target trial
Primary Outcome	Acute pancreatitis	Same as target trial
Causal Contrast	Intent to treat effect (effect of being assigned to the treatment)	Observational analogue of intent to treat effect
	Per protocol effect (effect of staying on the treatment)	Observational analogue of per protocol effect

from clinical notes using natural language processing (NLP), were used in the phenotyping model. In a validation study, it was observed that fixing the PPV at 0.90, the phenotyping model achieves sensitivity of 0.88 with structured features only and 0.92 when adding NLP features. A detailed list of structured diagnosis codes, lab tests, and NLP features used in the phenotyping model can be found in the Supplementary Table 1.

Covariates. Patient characteristics were assessed during 180 days before and including the cohort entry date. These included several claims-based characteristics such as demographics, medications, comorbidities, health service utilization metrics, and indices for general health including a Claims-based Frailty Index (CFI)[21] and Combined Comorbidity Index (CCI)[22]. EHR-based characteristics such as laboratory test results (HbA1c, serum creatinine, triglycerides), vitals and lifestyle factors (body mass index, blood pressure, tobacco use) were also assessed. Supplementary Table 2 contains a detailed description of all covariates.

Statistical analysis

To investigate the added value of EHR data in this study, we first conducted a claims-only analysis that defined acute pancreatitis based on ICD codes alone and adjusted for >130 claims-based variables and then conducted a claims + EHR augmented analysis that defined acute pancreatitis based on the phenotyping algorithm and added 6 additional variables from EHRs for confounding adjustment, both described above.

In all analyses, we used a propensity score (PS) based fine-stratification weighting method with 50 strata for confounding adjustment by measured factors[23]. We estimated PS as the predicted probability of initiating SGLT-2i compared to DPP-4i given the baseline patient characteristics from fitting a multivariable logistic regression model separately by database. Fifty strata were created based on the distribution of PS in SGLT-2i initiators, and DPP-4i initiators were assigned into these strata based on their PS resulting in 50 unequally sized strata. In the weighting step, DPP-4i initiators in each stratum were weighted proportional to the number of SGLT-2i initiators to account for stratum membership and achieve balance. The PS fine-stratification weighting approach, as implemented in this study, targets the average treatment effect in the treated (ATT), which is considered to be a highly relevant estimand for drug safety investigations[24,25]. Notably, other PS-based approaches that target different estimands such as average treatment effect in the whole population or average treatment effect in an overlapping population are available and can also be considered depending on the research question of interest.

As diagnostics for PS models, we evaluated distributional overlap, weight distribution, and individual covariate balance using standardized differences post-weighting. In the weighted population, we estimated the hazard ratio for acute pancreatitis among initiators of SGLT-2i versus DPP-4i using Cox proportional hazards model. Cumulative incidence was calculated using cumulative incidence functions and reported stratified by treatment groups[26].

Anticipating missing data in EHR-derived variables, we identified and described possible missingness patterns and mechanisms among partially observed covariates based on the observed data using the smdi R package[11]. After diagnosing the likely missingness mechanisms, we applied the corresponding multiple imputation methods to analytically address missingness in all EHR-based covariates. We created 20 imputed datasets where missing covariates were imputed based on random forest algorithms. In each of the imputed datasets, we fit the PS models and conducted PS fine stratification to calculate adjusted treatment effect estimates using the MatchThem R package[27]. The results were reported after pooling results using Rubin's rule to account for variance both within and across the imputed datasets[28].

Subgroup analyses and robustness evaluations

All subgroup and robustness evaluations were conducted with claims + EHR augmented analysis. We performed subgroup analyses in the following prespecified strata: age (<65 versus ≥ 65), sex (male versus female), and history of risk factors for acute pancreatitis (gallstones, tobacco use, alcohol abuse). We conducted two sensitivity analyses aimed at reducing missingness in EHR-based confounders and included: (1) increasing baseline window to 12 months before cohort entry, and (2) restricting the analysis to subjects with ≥ 3 EHR encounters in the 6 months before cohort entry. Finally, we evaluated ischemic stroke as a negative control outcome to detect net bias in the primary analysis[29].

Data availability

The data used in this study are derived from the RWE-DE commercial network and are not publicly available. However, access to these data can be obtained through licensing agreements with the respective commercial data providers.

Code availability

Reusable R codes to conduct study specific diagnostic and inferential analyses are made available publicly at https://gitlab-scm.partners.org/rjd48/sentinel_ic_uc1.

Received: 2 May 2025; Accepted: 30 November 2025;
Published online: 15 December 2025

References

1. Ball, R., Robb, M., Anderson, S. A. & Dal Pan, G. The FDA's sentinel initiative—a comprehensive approach to medical product surveillance. *Clin. Pharm. Ther.* **99**, 265–268 (2016).
2. Maro, J. C. et al. Six years of the US Food and Drug Administration's postmarket active risk identification and analysis system in the Sentinel initiative: implications for real world evidence generation. *Clin. Pharmacol. Ther.* https://doi.org/10.1002/cpt.2979 (2023).
3. Desai, R. J. et al. The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE). *Pharmacoepidemiol. Drug Saf.* https://doi.org/10.1002/pds.70028 (2024).
4. Schneeweiss, S., Desai, R. J. & Ball, R. Invited commentary: a future of data-rich pharmacoepidemiology studies- transitioning to large-scale linked EHR+claims data. *Am. J. Epidemiol*. https://doi.org/10.1093/aje/kwae226 (2024).
5. Moores, K., Gilchrist, B., Abrams, T. & Carnahan, R. *Mini-Sentinel Systematic Evaluation of Health Outcome of Interest Definitions for Studies Using Administrative Data—Pancreatitis Report* (2011).
6. Floyd, J. S. et al. Validation of acute pancreatitis among adults in an integrated healthcare system. *Epidemiology*. https://doi.org/10.1097/EDE.0000000000001541 (2023).
7. Sentinel Initiative. Advancing Scalable Natural Language Processing Approaches for Unstructured Electronic Health Record Data. Available at https://www.sentinelinitiative.org/methods-data-tools/methods/advancing-scalable-natural-language-processing-approaches-unstructured (2021).
8. Weberpals, J. et al. A principled approach to characterize and analyze partially observed confounder data from electronic health records. *Clin. Epidemiol*. **16**, 329–343 (2024).
9. Walsh, C. G. et al. Scalable incident detection via natural language processing and probabilistic language models. *Sci. Rep*. https://doi.org/10.1038/s41598-024-72756-7 (2024).
10. Carrell, D. S. et al. Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. *Am. J. Epidemiol.* **192**, 283–295 (2023).
11. Weberpals, J. et al. smdi: an R package to perform structural missing data investigations on partially observed confounders in real-world evidence studies. *JAMIA Open*. https://doi.org/10.1093/jamiaopen/ooae008 (2024).
12. Williamson, B. D. et al. Assessing treatment effects in observational data with missing confounders: a comparative study of practical doubly-robust and traditional missing data methods. Preprint at https://doi.org/10.48550/arXiv.2412.15012 (2024).
13. Stayner, L. et al. *Statistical Methods in Cancer Research Volume V: Bias Assessment in Case–Control and Cohort Studies for Hazard Identification* (International Agency for Research on Cancer, 2024).
14. Desai, R. J. et al. Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center. *BMJ*. https://doi.org/10.1136/bmj-2023-076460 (2024).
15. Elashoff, M., Matveyenko, A. V., Gier, B., Elashoff, R. & Butler, P. C. Pancreatitis, pancreatic, and thyroid cancer with glucagon-like peptide-1-based therapies. *Gastroenterology*. https://doi.org/10.1053/j.gastro.2011.02.018 (2011).
16. Sentinel Initiative. Risk of acute pancreatitis following SGLT2 inhibitor use in patients with type 2 diabetes mellitus. https://www.sentinelinitiative.org/sites/default/files/documents/Risk-of-Acute-Pancreatitis-Following-SGLT2-Inhibitor-Use-in-Patients-with-Type-2-Diabetes-Mellitus_Study-Protocol_19September2024.pdf (2024).
17. Sodhi, M., Rezaeianzadeh, R., Kezouh, A. & Etminan, M. GLP-1 agonists and gastrointestinal adverse events. *JAMA*. https://doi.org/10.1001/jama.2023.19574 (2023).
18. Cao, C. et al. GLP-1 receptor agonists and pancreatic safety concerns in type 2 diabetic patients: data from cardiovascular outcome trials. *Endocrine* https://doi.org/10.1007/s12020-020-02223-6 (2020).
19. Wang, H. et al. Hemodialysis and risk of acute pancreatitis: a systematic review and meta-analysis. *Pancreatology*. https://doi.org/10.1016/j.pan.2020.11.004 (2021).
20. Dassopoulos, T. & Ehrenpreis, E. D. Acute pancreatitis in human immunodeficiency virus-infected patients: a review. *Am. J. Med*. https://doi.org/10.1016/S0002-9343(99)00169-2 (1999).

21. Kim, D. H. et al. Measuring frailty in medicare data: development and validation of a claims-based frailty index. *J. Gerontology: Ser. A* **73**, 980–987 (2018).
22. Gagne, J. J., Glynn, R. J., Avorn, J., Levin, R. & Schneeweiss, S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J. Clin. Epidemiol.* **64**, 749–759 (2011).
23. Desai, R. J., Rothman, K. J., Bateman, B. T., Hernandez-Diaz, S. & Huybrechts, K. F. A propensity-score-based fine stratification approach for confounding adjustment when exposure is infrequent. *Epidemiology*. https://doi.org/10.1097/EDE.0000000000000595 (2017).
24. Cafri, G. Why we should be prioritizing the average treatment effect on the treated over other Estimands when evaluating drug and device safety. *Am. J. Epidemiol*. **194**, 3602–3608 (2025).
25. Desai R. J., Franklin J. M. Alternative approaches for confounding adjustment in observational studies using weighting based on thepropensity score: a primer for practitioners. *BMJ* **367**, l5657 (2019).
26. Austin, P. C., Lee, D. S. & Fine, J. P. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*. https://doi.org/10.1161/CIRCULATIONAHA.115.017719 (2016).
27. Pishgar, F., Greifer, N., Leyrat, C. & Stuart, E. MatchThem: Matching and Weighting after Multiple Imputation. *The R. Journal.* https://doi.org/10.32614/RJ-2021-073, https://journal.r-project.org/archive/2021/RJ-2021-073/ (2021).
28. Rubin, D. B. Multiple imputation for nonresponse in surveys. *Wiley Series in Probability and Statistics* (1987).
29. Shi, X., Miao, W. & Tchetgen Tchetgen, E. A selective review of negative control methods in epidemiology. *Curr. Epidemiol. Rep.* https://doi.org/10.1007/s40471-020-00243-4 (2020).

Acknowledgements

Many thanks are due to TriNetX (Cambridge, MA) and HealthVerity (Philadelphia, PA) for providing data for this analysis, with data from HealthVerity containing in part certain data from Veradigm. This project was supported by Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA). FDA coauthors reviewed the study protocol, statistical analysis plan, and the manuscript for scientific accuracy and clarity of presentation. Representatives of the FDA reviewed a draft of the manuscript for presence of confidential information and accuracy regarding statement of any FDA policy. The views expressed are those of the authors and not necessarily those of the US FDA.

Author contributions

R.J.D., J.W., S.S., S.T., E.P., J.J.H., F.M.S., J.L. (Jie Li), J.T.J., and A.A. conceived and designed the study. R.J.D., J.W., S.S., S.T., E.P., H.P., R.S., and M.D. worked on protocol development. R.J.D., H.P., J.L. (Joyce Lii), R.H., and J.G.L. conducted data analysis. R.J.D. wrote the manuscript draft. R.S., M.D., and C.J. provided project management and administrative support. All authors reviewed and critically revised the manuscript draft.

Competing interests

Dr. Desai reports serving as Principal Investigator on investigator-initiated grants to the Brigham and Women's Hospital from Novartis, Vertex, and Bayer on unrelated projects. Dr. Schneeweiss is co-principal investigator of an investigator-initiated grant to the Brigham and Women's Hospital from Boehringer Ingelheim unrelated to the topic of this study. He is a consultant to Aetion Inc., a software manufacturer of which he owns equity. His interests were declared, reviewed, and approved by the Brigham and Women's Hospital and MGB HealthCare System in accordance with their institutional compliance policies. Dr. Weberpals is now an employee of AstraZeneca and owns stocks in AstraZeneca. Other authors declare no financial or non-financial competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-02234-5.

Correspondence and requests for materials should be addressed to Rishi J. Desai.