Title:

Health Information Exchanges (HIEs) as Novel Sources for Population-Based Post Marketing Surveillance of Medical Products: A Pilot Study from the FDA Sentinel Innovation Center

Contributors:

Anjum Khurshid, MD, PhD[1]; David Kendrick, MD, MPH, FACP[2]; Haritha S. Pillai, MPH[3]; Jenny Sun, PhD[2], Eliel Oliveira, MBA, MSc, FAMIA[4]; Alexis Jaquish[1]; Sarah K. Dutcher, MS, PhD[5]; Patricia Bright, PhD[5]; Youjin Wang, PhD[5]; Jie Li, PhD[5], Tiffany Austin, MPH[5], Patrick J. Heagerty, MS, PhD[6]; Michael E. Matheny, MD, MS, MPH[7]; Sebastian Schneeweiss, MD, ScD[3]; Rishi J. Desai, MS, PhD[3]

Affiliations:

[1] Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA

[2] MyHealth Access Network, Tulsa, OK, USA

[3] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[4] Connxus, Austin, TX, USA

[5] Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, United States Food and Drug Administration, Silver Spring, MD, USA

[6] Department of Biostatistics, University of Washington, Seattle, WA, USA

[7] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

Abstract:

Introduction: Health information exchanges (HIEs) provide the capability to electronically move health care information among different health care information systems and store these data for downstream use cases. However, use of data from HIEs for postmarketing surveillance of medical products is not previously explored.

Objectives: To conduct a pilot descriptive study characterizing data from MyHealth Access Network - a statewide HIE for Oklahoma, to understand its utility for conducting pharmacoepidemiology studies.

Materials and methods: MyHealth Access Network connects 95% of all hospital activity, 100% of federally qualified health center activity, and most community behavioral health clinics, tribal health systems, and independent providers in the state. As a use case to understand data in MyHealth Access Network, we characterized patients with Type 2 Diabetes Mellitus (T2DM), aged 18 years or older and treated with one of two common antidiabetic drug classes: Sodium-Glucose Co-Transporter 2 inhibitors (SGLT-2i) and Dipeptidyl Peptidase 4 inhibitors (DPP-4i) in a new user cohort design. A cohort entry date was defined based on records of medication initiation for either of the two drug classes, using dispensing data from insurance claims (when available) or prescribing data from electronic health records (EHRs) when claims were not available. Patient characteristics including demographics, comorbidities, comedications, clinical and lifestyle related factors, and healthcare utilization factors were summarized for the two exposure groups.

Results: A total of 76,018 DPP-4i initiators and 101,599 SGLT-2i initiators met our inclusion criteria. The mean age was 59 years and nearly half were women in both the groups. While there was a large proportion of those with missing race information (32% DPP-4i, 24% SGLT-2i), a majority were White (47% DPP-4i, 57% SGLT-2i). The second most reported race category was American Indian or Alaskan

Natives (13% DPP-4i, 9.9% SGLT-2i), which is in line with their representation in the 2024 US Census for Oklahoma. Hemoglobin A1c (HbA1c) results were available for 63% and 66% of the DPP-4i and SGLT-2i groups with mean ± SD of 8.1± 1.9% and 8.2 ± 1.9%, respectively. Serum creatinine was recorded for 67% DPP-4i and 70% SGLT-2i initiators with mean ± SD of 1.30 ± 1.01 mg/dL and 1.14 ± 0.64 mg/dL, respectively.

Discussion: HIEs may offer a promising resource for population-based postmarketing surveillance of medical products owing to their comprehensive capture of information across various healthcare settings.

1. Introduction:

Fragmentation in the U.S. healthcare system poses critical challenges for both patient level care coordination and population level research. A 2007 analysis of Medicare data reported that patients visited, on average, more than 7 different healthcare providers annually, and that, on average, primary care providers serving a panel of patients found themselves coordinating care among 225 other providers in 117 other organizations.[1] As the adoption of electronic health records (EHR) by eligible providers and hospitals in the U.S. healthcare system surged following significant investments through the Health Information Technology for Economic and Clinical Health (HITECH) Act, the need to facilitate seamless exchange of clinical data securely across disparate competing EHR systems and provider organizations was clearly recognized.[2] As described in the Office of the National Coordinator for Health IT (ONC) Roadmap for Nationwide Interoperability (2015)[3], "every patient deserves to have their complete, longitudinal health record available where and when they need it for decisions about their care". As a result, approximately 3% of HITECH funding supported the dramatic expansion of state and regional health information exchanges (HIEs) throughout the country.[4] The primary purposes of HIEs is to promote interoperability and coordination of care. Implementing and operationalizing this mission has been done by creating trusted regional networks where the silos of data, existing due to the fragmented nature of healthcare delivery, are connected through a governed, secure, central exchange

platform. Since many of these exchanges also maintain the exchanged clinical documents and patient records in a centralized clinical data repository, over the years, these HIEs have become a comprehensive source of longitudinal, population level, linked EHR data in the country.[5]

Almost all states in the U.S. have a statewide HIE or multiple regional HIEs. A recent survey funded by the ONC has reported about 80 fully operational HIEs in the country with an average data repository comprising more than four million unique patients with data going back to around 15 years in some cases. [6,7] These HIEs are mostly non-profit entities with a multistakeholder community governance. They maintain a robust technical infrastructure to create bidirectional data interfaces with large and small hospitals, federally qualified health clinics, physician practices, Emergency Medical Services (EMS), behavioral health providers, pharmacies, laboratories, and public health clinics among other types of organizations that are sources of health-related data. These interfaces allow for exchange of detailed clinical data through secure connections using FHIR APIs (Fast Health Interoperability Resources Application Programming Interfaces), HL7 (Health Level 7) interfaces, CCD (Continuity of Care Documents), or even flat files (.csv).[8] Importantly, HIEs maintain governance and compliance required to build trust in the privacy and confidentiality of sensitive clinical data. The HIEs store and use this data as Business Associate Agreement (BAA) partners for all providers connected to them and fully comply with HIPAA (Health Insurance Portability and Accountability Act, 1996) and other federal, state and local laws.[9] There are also continuous data quality and data integrity assurance processes[10] implemented in all major HIEs. This broad technical infrastructure has matured over the last decade, making HIEs an attractive source of large-scale EHR data that is ready to be leveraged for multiple important use cases. However, historically HIEs have been underutilized for conducting public health surveillance, pharmacoepidemiologic studies, and population outcomes research.[11]

The Sentinel Initiative is a key component of the U.S. Food and Drug Administration's (FDA) postmarket surveillance for approved medical products. Results from Sentinel investigations are

routinely used to support the FDA in regulatory decision making.[12] The Sentinel System, a national medical product safety surveillance system within the Sentinel Initiative, has historically relied on large volumes of insurance claims data but more recently has developed an infrastructure linking EHRs with insurance claims to address questions related to medical product use and associated outcomes using more granular clinical information captured in EHRs.[13,14] The new infrastructure, the Real World Evidence Data Enterprise (RWE-DE), comprises linked claims-EHR data for 26 million individuals from six data sources involving four academic partner institutions (three tertiary care centers- Mass General Brigham, Vanderbilt University Medical Center, Duke University Health Systems, and an integrated care delivery network- Kaiser Permanente- Washington) collectively termed the Development Network and two commercial data networks (HealthVerity, TriNetX) collectively termed the Commercial Network.[14] While a major step forward, the RWE-DE may benefit from inclusion of additional data sources to increase the total population covered and availability of near-real time information. Additionally, all current sources face challenges in EHR data continuity due to the fragmentation of care described above. To explore the potential of HIEs for addressing some of these challenges, we initiated a pilot study with MyHealth Access Network- a statewide HIE from Oklahoma, to implement a descriptive analysis of initiators of two antidiabetic medications to reflect a typical Sentinel query. In this article, we describe the learnings from the pilot and discuss considerations for future use of HIEs as novel sources for population-based pharmacoepidemiology research and postmarketing surveillance of medical products.

2. Methods:

2.1 Data source

MyHealth Access Network is a non-profit organization and Oklahoma's State Designated Entity for Health Information Exchange that connects 95% of all hospital activity, 100% of federally qualified health center activity, and most community behavioral health clinics, tribal health systems, and independent providers

in the state.[15] It connects to more than 500 unique healthcare entities including independent pharmacies, emergency medical services (EMS), social services organizations, organ procurement organizations, Oklahoma Office of the Chief Medical Examiner, and health plans in the area. The HIE has a clinical data repository that goes back to encounters captured since 2011 and provides the most complete and near real-time access to population-level healthcare and social needs data for the over 4 million residents of Oklahoma. These data are standardized and linked across all providers into a unique unduplicated record for each patient. We used data from January 2016 to December 2023 available in MyHealth Access Network, which also included linked EHR data to claims data from three of the health plans.

2.2 Study design

We conducted a descriptive medical product utilization query (similar to a 'Type 5' query in the Sentinel Routine Querying tools)[16]- to identify and characterize patients with Type 2 Diabetes Mellitus (T2DM) treated with two common antidiabetic drug classes: Sodium-Glucose Co-Transporter 2 inhibitors (SGLT-2i) and Dipeptidyl Peptidase 4 inhibitors (DPP-4i) in a new user cohort design. Patients entered the cohort on a cohort entry date (CED) based on records of medication initiation for either of the two drug classes, using dispensing data from insurance claims (when available) or prescribing data from EHRs when claims were not available. Patients were required to be aged 18 years or older, have a diagnosis of type 2 diabetes mellitus, and have some EHR activity (operationalized as having any data available from EHRs any time prior to index date) to ensure adequate contact with the healthcare system and capture of clinical codes to measure baseline characteristics. Since the focus of this query was on patients with type 2 diabetes mellitus, we excluded patients with a diagnosis of type 1 diabetes mellitus.

2.3 Variables

We assessed demographic characteristics (age, sex, race, ethnicity and year of cohort entry); diabetes-related factors including duration of diabetes (operationalized as the time between the first diagnosis

codes found in the data and cohort entry) and use of other antihyperglycemic medications; a range of comedications and comorbid conditions; laboratory test results and other clinical parameters including serum creatinine, hemoglobin A1c (HbA1c), total cholesterol, high density lipoprotein (HDL), low density lipoprotein (LDL), and left ventricular ejection fraction (LVEF); vital signs (heart rate, blood pressure), body mass index (BMI), and lifestyle factors (any smoking status, alcohol use, drug misuse or abuse). All variables were assessed in a 365-day baseline period prior to the cohort entry date, unless otherwise specified. For laboratory tests, clinical parameters, and vital signs, values closest to the cohort entry were used for patients with multiple recordings in the baseline period. Standardized coding terminology including national drug codes (NDCs) and RxNorm Concept Unique Identifier (RxCUI) for medications with drug mechanism classifications determined by manual mapping, Logical Observational Identifiers Names and Codes (LOINC) for laboratory tests, international classification of diseases (ICD)-10-CM codes for diagnoses, and current procedural terminology (CPT) codes (with modifiers) to assess in-person healthcare utilization were used to identify relevant information in each domain.

2.4 Statistical analysis

We used basic descriptive statistics to summarize all characteristics for each of the two exposure groups using percentages for categorical variables and means (standard deviations) for quantitative measures. We also reported the proportion of individuals in each group who had at least one in-person EHR encounter in 30, 90, 180, and 365 days prior to the cohort entry to contextualize routine flow of information in the HIE.

2.5 Ethics review

This Sentinel project is a public health surveillance activity conducted under the authority of the Food and Drug Administration and, accordingly, is not subject to Institutional Review Board oversight.[17-19] Additionally, the HIE governance committee comprising local stakeholders and data contributors, reviewed and approved the secondary use of the data for this query.

3. Results:

Figure 1 shows the total number of DPP-4i and SGLT-2i initiators with type 2 diabetes mellitus along with source of the prescription information on their cohort entry date. A total of 76,018 DPP-4i initiators and 101,599 SGLT-2i initiators met our inclusion criteria, a majority of whom were included based on prescription orders found in EHRs. **Figure 2** outlines information sources for the two exposure groups during the study period. A large majority (63.0% and 60.1% for DPP-4i and SGLT-2i groups, respectively) were identified from EHR data alone and did not appear in any of the three claims datasets. A total of 15.7% and 20.1% for DPP-4i and SGLT-2i, respectively had information in both EHRs and claims from one or more of the three payers. Notably, a small subset of patients (0.003% and 0.01% for DPP-4i and SGLT-2i) existed in all three payer datasets and EHR at some point of time for both DPP-4i and SGLT-2i initiators, which highlights the issue of frequent changes in individuals' insurance coverage over time with claims data sources.

Patient demographics, medication use, and comorbid diagnoses

Table 1 summarizes the demographic characteristics of the DPP-4i and SGLT-2i initiators. The mean age was 59 years in each of the cohorts with the largest group of patients falling within the age category of 55-64 years (30%, 32% respectively). About half of the patients were female in both the cohorts (49%, 45% respectively). Less than 0.05% of patients had missing data for age and sex. While information on race was unknown or missing for 32% of the DPP-4i initiators and 24% of SGLT-2i initiators, a majority of patients were White in both the cohorts (47%, 57% respectively); the second most reported race category was American Indian or Alaskan Natives, representing 13% and 9.9%, respectively. A total of 8.5% of the DPP-4i initiators and 7.5% of the SGLT-2i initiators were of Hispanic origin.

Table 2 provides information on diabetes duration, comedications, and comorbid conditions. Most patients had a diagnosis of Type 2 DM for 1-5 years at the time of cohort entry across both DPP-4i initiators and SGLT-2i initiators (39%, 48%). Among other antidiabetic medications, metformin was most common, followed by sulfonylureas. GLP1 agonists were much more frequently used by SGLT-2i initiators (19% vs 6.3%). Angiotensin converting enzyme inhibitors and angiotensin receptor blockers were the most common co-medications in both the groups (42%, 52%) and hypertension was the most frequent comorbid diagnosis (32%, 45%).

Table 3 summarizes information on laboratory test results, lifestyle factors and vital signs. HbA1c results were available for 63% and 66% of the DPP-4i and SGLT-2i groups, respectively with mean $\pm$ SD of 8.1$\pm$ 1.9% and 8.2 $\pm$ 1.9%. Serum creatinine was recorded for 67% of DPP-4i and 70% of SGLT-2i initiators. The mean serum creatinine values were 1.30 $\pm$ 1.0 mg/dL and 1.14 $\pm$ 0.6 mg/dL for DPP-4i and SGLT-2i initiators respectively. Substance use including smoking status, alcohol use, and drug misuse/abuse was recorded for 5-23% of the patients across both DPP-4i and SGLT-2i initiators. BMI was recorded for 28.4% of the DPP-4i initiators (mean = 33.4 $\pm$ 8.0 kg/m^2) and 46% of the SGLT-2i initiators (mean = 34.1 $\pm$ 8.0 kg/m^2). Heart rate and blood pressure had similar mean values across both groups, with measurements recorded for 25-35% of DPP-4i initiators, and a higher proportion (45-50%) of SGLT-2i initiators.

Table 4 summarizes healthcare utilization-related information. We observed that 64% of the DPP-4i initiators and 76 % of the SGLT-2i initiators had at least 1 in-person encounter in the 365 days prior to the cohort entry. This proportion was consistent when restricting to past 180 days (62 %, 74 %), 90 days (60 %, 72 %), and 30 days (55 %, 67 %).

4. Discussion

In this study, we observed that it was feasible to create patient cohorts using near real-time EHR data to describe initiators of medications using data from an HIE. We noted that capture of important

clinical data elements such as HbA1c was comparable or more complete than diabetic cohorts from other EHR sources from the US.[20-22] Importantly, traditionally underrepresented racial groups such as American Indians were represented in line with their representation in the 2024 US Census for Oklahoma (10-13% in this study, 9.5% in the US Census)[23], which highlights the truly population-based nature of HIEs.

HIEs offer certain benefits that make them a promising resource for future use in national medical product surveillance systems like Sentinel for pharmacoepidemiology evaluations. First, since most HIEs are community-governed with a not-for-profit status and funding from public sources (state and federal health departments), they are quite efficient in terms of cost to acquire data for surveillance and research. As the source of almost real-time data on an entire population, HIEs may help to study emergent infections, adverse events for newly used treatments, and geographical coverage within the covered regions in a timely and comprehensive manner. Many HIEs also store unstructured data, like clinical notes, creating a valuable resource to develop and test natural language processing algorithms and large language models.[24] Increasingly, HIEs are expanding their role to collect nonmedical data such as social needs assessments, social services referrals, and criminal justice, environmental, and housing data to become health data utilities (HDUs) for their state or region, much like the energy or public transportation utilities that provide the service for multiple uses by consumers.[25] Some states have also passed legislation to mandate contribution of health and social services data to their HDUs to help state-level public health and Medicaid better coordinate, monitor, and evaluate care for its population. This rich and comprehensive data under a community-governed entity, which has high levels of technical expertise to curate representative population level data of clinical and non-clinical factors, is primed to help study health equity issues in much better detail than in any other dataset of this size and coverage.[26]

One of the most important requirements for conducting causal inference studies with longitudinal health records is the availability of EHR data without 'information leakage' that is common in traditional EHR sources from a single healthcare system because patients frequently seek care outside of the system.[27] A key service provided by HIEs is that of maintaining a master patient index that links all the various medical records across the fragmented health delivery system into one master record, thus saving a huge effort in deduplicating and linking clinical records for subsequent analysis. The ability to maintain a truly longitudinal patient record across healthcare systems makes HIEs an attractive data resource to conduct studies that require long term follow-up and consistent data availability over the course of a treatment or a disease. In the analysis of healthcare utilization, we observed that about 60-70% patients included in the current study had an in-person EHR visit in 365 days, 180 days, and 30 days prior to the medication initiation date, which suggests that HIEs have a stable pool of patients with consistent flow of EHR information.

Some governance related aspects of HIE data merit a discussion. While a single patient-level record can be shared across the network for treatment, payment, and operations for authorized users with an established relationship with that patient, the secondary use of these data for research and surveillance has established policies that vary across HIEs. These policies are largely focused on patient privacy and agreed upon governance policies with data contributors. Secondary use data sharing for public health, epidemiological studies, and population health research is determined on a case-by-case basis. Most HIEs will allow use of data for public health surveillance activities, but in many cases, use of data for research requires appropriate IRB supervision and may require patient consent for sharing any protected health information.

We also acknowledge some important limitations of our current work. First, we only explored a single HIE in Oklahoma which is a relatively mature and sophisticated HIE with a legislative mandate from the state. HIEs across the country differ in significant ways in terms of their geographical and

clinical data capture as well as their ability to link to other sources like insurance claims. However, many of the features described for MyHealth Access are likely to be representative of larger state level and regional HIEs, some with much larger populations such as in New York, California, Michigan, and Texas. Second, we only focused on a descriptive study of relatively commonly used medication classes, which precludes generalizable conclusions regarding completeness of information for more rare conditions or medications. However, it is reasonable to hypothesize that if there is a method of connecting a consortium of HIEs from across the country to act as a distributed network with a common data model and quality assurance, as established in the Sentinel Distributed Network, this limitation may be adequately addressed through a geographically and socioeconomically representative cohort from across the country. A preliminary feasibility study for such a consortium of HIEs to support pharmacoepidemiology and population health studies is already underway and showing encouraging results.[28]

In conclusion, the HIE network has grown over the last decade at the local, regional, and national level. Our pilot study conducted in partnership with MyHealth Access Network suggests that data from HIEs may offer a promising resource for conducting population-based postmarketing surveillance of medical products with EHR data.

Statements and Declarations:

Funding Acknowledgements:

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 U.S. FDA.

Competing Interests:

Dr. Khurshid and Eliel Oliveira are co-founders of Pulsar Health, a software IT company. 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. All other authors have no conflicts of interest to disclose relative to this work.

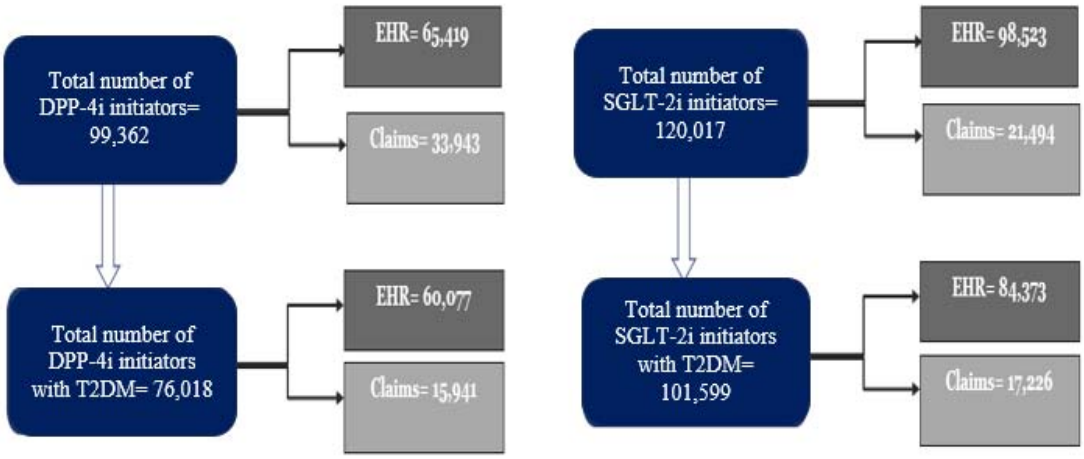
References

1. Pham HH, Schrag D, O'Malley AS, Wu B, Bach PB. Care Patterns in Medicare and Their Implications for Pay for Performance. *New England Journal of Medicine*. 2007-03-15;356(11)doi:10.1056/NEJMsa063979
2. What is the HITECH Act? The HIPAA Journal. https://www.hipaajournal.com/what-is-the-hitech-act/
3. Office of the National Coordinator for Health Information Technology. Connecting Health and Care for the Nation: A Shared Nationwide Interoperability Roadmap. https://www.healthit.gov/sites/default/files/ONC2015-Interop-Roadmap.pdf
4. State Health Information Exchange Cooperative Agreement Program. https://www.healthit.gov/topic/onc-hitech-programs/state-health-information-exchange
5. Holmgren AJ, Esdar M, Hüsers J, Coutinho-Almeida J. Health Information Exchange: Understanding the Policy Landscape and Future of Data Interoperability. *Yearbook of Medical Informatics*. 2023/08;32(01)doi:10.1055/s-0043-1768719
6. Civitas Networks for Health. Key Learnings from the National Health Information Organization Survey Results https://www.civitasforhealth.org/learnings-from-national-hio-survey-results/
7. Everson J, Chang W, Patel V, Adler-Milstein J. The state of health information organizations and plans to participate in the federal exchange framework. *Health Affairs Scholar*. 2024/08/02;2(8)doi:10.1093/haschl/qxae098
8. The Office of the National Coordinator for Health Information Technology. What Is HL7 FHIR? https://www.healthit.gov/
9. U.S. Department of Health and Human Services. HIPAA, Health Information Exchanges, and Disclosures of Protected Health Information for Public Health Purposes. https://www.hhs.gov/sites/default/files/hie-faqs.pdf
10. Garg N, Kuperman G, Onyile A, et al. Validating Health Information Exchange (HIE) Data For Quality Measurement Across Four Hospitals. *AMIA Annual Symposium*. 2014;2014:573–579
11. Yeager VA, Vest JR, Walker D, Diana ML, Menachemi N. Challenges to Conducting Health Information Exchange Research and Evaluation: Reflections and Recommendations for Examining the Value of HIE. *eGEMs*. 2017 Sep 4;5(1)doi:10.5334/egems.217
12. Maro JC, Nguyen MD, Kolonoski J, 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. *Clinical Pharmacology & Therapeutics*. 2023/10/01;114(4)doi:10.1002/cpt.2979
13. Brown JS, Mendelsohn AB, Nam YH, et al. The US Food and Drug Administration Sentinel System: a national resource for a learning health system. *Journal of the American Medical Informatics Association*. 2022/11/14;29(12)doi:10.1093/jamia/ocac153
14. Desai RJ, Marsolo K, Smith J, et al. The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE). *Pharmacoepidemiology and Drug Safety*. 2024/10/01;33(10)doi:10.1002/pds.70028
15. MyHealth Access Network. https://myhealthaccess.net/
16. Sentinel. Sentinel Routine Querying System Overview. https://dev.sentinelsystem.org/projects/SENTINEL/repos/sentinel-routine-querying-tool-documentation/browse/files/file160-type05-medprodutil.md
17. Basic HHS Policy for Protection of Human Research Subjects, 45 CFR §46.102(l)(2). Accessed April 14, 2025, 2025. https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A#46.102
18. U.S. Department of Health and Human Services. Federal Policy for the Protection of Human Subjects. 82 Fed Reg 7149. January 19, 2017;

19. Rosati K, Jorgensen NS, Melissa, Evans BJ. Sentinel Initiative Principles and Policies: HIPAA and Common Rule Compliance in the Sentinel Initiative. February 01, 2018;
20. Nanna MG, Kolkailah AA, Page C, Peterson ED, Navar AM. SGLT2i and GLP-1 RA in Patients With Diabetes and Cardiovascular Disease. *JAMA Cardiology*. 2023/01/01;8(1)doi:10.1001/jamacardio.2022.3839
21. Kukhareva PV, Facelli JC, O'Brien MJ, et al. Racial and Ethnic Disparities in Prescribing of GLP-1 Receptor Agonists in the United States: A Retrospective Cohort Analysis - PubMed. *medRxiv : the preprint server for health sciences*. 11/01/2024;doi:10.1101/2024.10.28.24316312
22. Li Y, Hu H, Zheng Y, et al. Impact of Contextual-Level Social Determinants of Health on Newer Antidiabetic Drug Adoption in Patients with Type 2 Diabetes. *International Journal of Environmental Research and Public Health 2023, Vol 20, Page 4036*. 2023-02-24;20(5)doi:10.3390/ijerph20054036
23. United States Census Bureau. https://www.census.gov/quickfacts/fact/table/OK
24. Alkasir A, Seidman G, Ritzo J, Bari L, Khurshid A. The Role of Health Data Utilities in Supporting Health AI. *NEJM AI*. 2024;1(10)doi:10.1056/AIpc2400401
25. Civitas Networks for Health. Health Data Utility Framework – A Guide to Implementation. https://www.civitasforhealth.org/hduframework/
26. Khurshid A, Sarkar IN. The health data utility and the resurgence of health information exchanges as a national resource - PubMed. *Journal of the American Medical Informatics Association : JAMIA*. 2025;doi:10.1093/jamia/ocaf032
27. Schneeweiss S, Desai RJ, Ball R. Invited commentary: A future of data-rich pharmacoepidemiology studies- transitioning to large-scale linked EHR+claims data. *Am J Epidemiol*. Jul 16 2024;doi:10.1093/aje/kwae226
28. American Medical Informatics Association-Enhancing Public Health Insights with HIE (Health Information Exchange) Data. https://amia.secure-platform.com/symposium/solicitations/102001/sessiongallery/schedule/items/94086

Figures and tables:

Figure 1 Final cohort of DPP-4i and SGLT-2i initiators with type 2 diabetes mellitus along with data source (EHR versus claims)



- SGLT-2i: Sodium-glucose co-transporter-2 inhibitors; DPP-4i: Dipeptidyl peptidase-4 inhibitors; T2DM: Type 2 Diabetes Mellitus; EHR: Electronic Health Record
- Patients were enrolled in the cohort on the cohort entry date (CED) based on medication initiation records from insurance claims or EHR prescribing data when claims were unavailable.

Figure 2 Claims versus EHR overlap of DPP-4i and SGLT-2i initiators with type 2 diabetes mellitus

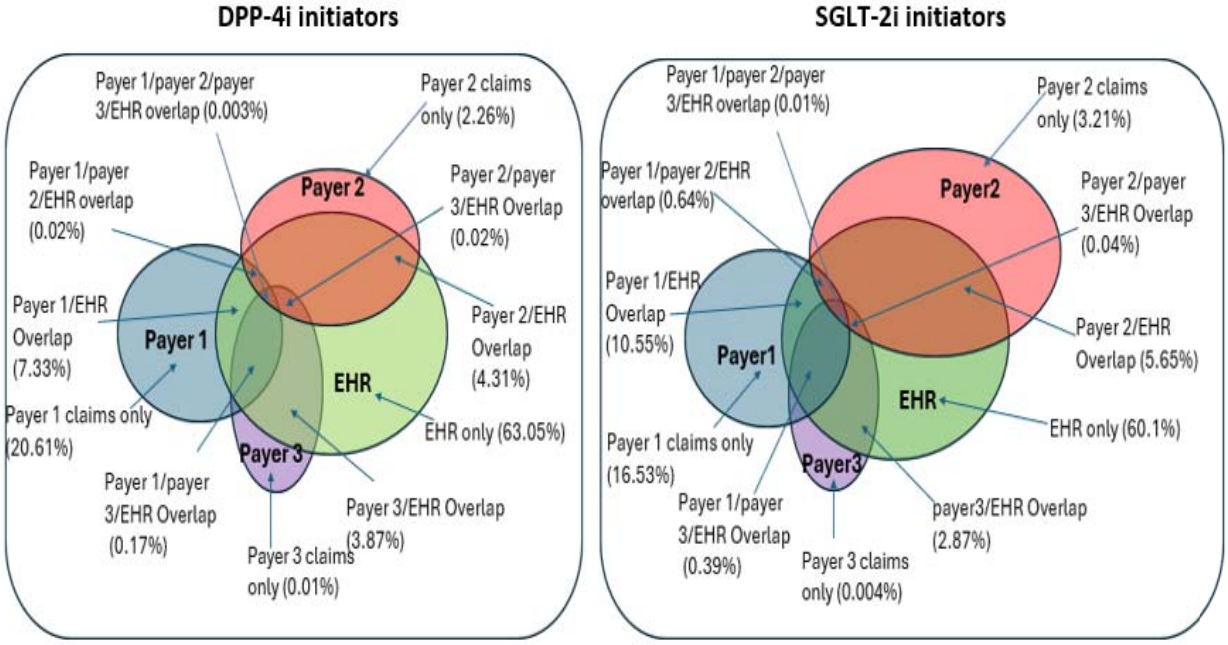


Table 1 Demographic characteristics of the DPP-4i and SGLT-2i initiators with type 2 diabetes mellitus

Patient Characteristics^	DPP-4i initiators with Type 2 DM		SGLT-2i initiators with Type 2 DM		Provenance* (EHR/Claims/ Both)
Total Count	76,018		101,599		
	Number/Mean	**Percent/ Standard Deviation**	**Number/Mean**	**Percent/ Standard Deviation**	
Age at Diagnosis (in years)	59.4	13.3	58.9	12.6	Both
Age Category					
18-24 years	469	0.62%	580	0.57%	Both
25-34 years	2,605	3.4%	3,250	3.2%	Both
35-44 years	8,378	11%	11,257	11%	Both
45-54 years	17,096	23%	23,434	23%	Both
55-64 years	22,942	30%	32,226	32%	Both
65-74 years	15,820	21%	21,414	21%	Both
75-84 years	7,226	9.5%	8,252	8.1%	Both
85+ years	1,459	1.9%	1,164	1.2%	Both
Not Reported[b]	23	0.03%	22	0.02%	Both
Sex					
Female	37,524	49%	45,627	45%	Both
Male	38,459	51%	55,930	55%	Both
Not Reported[b]	35	0.05%	42	0.04%	Both
Year of Cohort Entry[a]					
2016	3,265	4.3%	2,444	2.4%	Both
2017	3,489	4.6%	2,607	2.6%	Both
2018	5,911	7.8%	4,828	4.8%	Both
2019	8,862	12%	8,257	8.1%	Both
2020	15,478	20%	16,383	16%	Both
2021	11,724	15%	16,961	16.9%	Both
2022	8,780	12%	22,056	21.7%	Both
2023	18,509	24.3%	28,063	27.6%	Both
Race					
American Indian or Alaska Native	10,242	13%	10,005	9.9%	Both
Asian	1,109	1.5%	1,683	1.7%	Both
Black or African American	4,138	5.4%	6,832	6.7%	Both
Native Hawaiian/other Pacific Islander	134	0.18%	261	0.26%	Both
Unknown[b]	24,502	32%	24,700	24%	Both
White	35,893	47%	58,118	57%	Both
Ethnicity: Hispanic	6,442	8.5%	7,616	7.5%	Both

*This column specifies the source (claims versus EHR) of each variable; it is possible that this information comes from either, both or neither of the two sources.

[a]Cohort entry was determined based on medication start date depending on the earliest available record. Dispensed date was used if it was the earliest or occurred within 90 days of the earliest prescribed date; otherwise, the prescribed date was used. [b]Any discrepancies between EHR and claims after data cleaning were categorized as "not reported" or "unknown"

98 **Table 2** Diabetes duration, comedications, and comorbidities observed in the DPP-4i and SGLT-2i initiators with type 2
99 diabetes mellitus

Patient Characteristics	DPP-4 inhibitor initiators with Type 2 DM		SGLT-2 inhibitor initiators with Type 2 DM		Provenance* (EHR/Claims/Both)
	N/Mean	%/SD	N/Mean	%/SD	
Duration of diabetes (in years), N (%)[a]					
< 1	26,599	35%	32,330	32%	Both, although EHR had longer history
1-5	29,323	39%	48385	48%	Both, although EHR had longer history
6-10	4,549	6.0%	9920	9.8%	Both, although EHR had longer history
11-15	846	1.1%	1770	1.7%	Both, although EHR had longer history
>15	374	0.49%	398	0.39%	Both, although EHR had longer history
[b]History of Antihyperglycemic Agents					
Metformin	34,590	46%	53,066	52%	EHR: Prescribed; Claims: Dispensed
Thiazolidinediones	4,977	6.6%	6,493	6.4%	EHR: Prescribed; Claims: Dispensed
Sulfonylureas	18,679	25%	23,238	23%	EHR: Prescribed; Claims: Dispensed
Alpha Glucosidase Inhibitors	603	0.79%	452	0.44%	EHR: Prescribed; Claims: Dispensed
Meglitinides	201	0.26%	255	0.25%	EHR: Prescribed; Claims: Dispensed
Glucagon-like Peptide-1 Agonists	4,791	6.3%	19,189	19%	EHR: Prescribed; Claims: Dispensed
Short/Rapid Acting Insulins	7,023	9.2%	13,247	13%	EHR: Prescribed; Claims: Dispensed
Long/Intermediate Acting Insulins	4,100	5.4%	8,624	8.5%	EHR: Prescribed; Claims: Dispensed
Long and Short Acting Insulin Combinations	713	0.94%	1,355	1.3%	EHR: Prescribed; Claims: Dispensed
Count of total oral antihyperglycemic agents	0.87	0.96	1.23	1.06	EHR: Prescribed; Claims: Dispensed
[b]History of other medications					
Antiplatelets	20,123	26%	30,937	30%	EHR: Prescribed; Claims: Dispensed
Drugs for Obstructive Airway Diseases	19,063	25%	29,566	29%	EHR: Prescribed; Claims: Dispensed
Calcium Channel Blockers	13,794	18%	21,016	21%	EHR: Prescribed; Claims: Dispensed

Diuretics	20,425	27%	33,576	33%	EHR: Prescribed; Claims: Dispensed
Antidepressants	14,962	20%	24,054	24%	EHR: Prescribed; Claims: Dispensed
Beta Blockers	21,112	28%	35,158	35%	EHR: Prescribed; Claims: Dispensed
Agents Acting on the Renin-Angiotensin System	31,585	42%	52,466	52%	EHR: Prescribed; Claims: Dispensed
History of comorbidities					
Acute Kidney Failure/Chronic Kidney Disease	18,160	24%	36,092	36%	Both
Acute Myocardial Infarction	939	1.2%	2,321	2.3%	Both
Acute Respiratory Disease	7,569	10%	14,881	15%	Both
Anemia	4,099	5.4%	7,008	6.9%	Both
Atrial Fibrillation	2,373	3.1%	5,111	5.0%	Both
Cardiomyopathy	1,070	1.4%	3,835	3.8%	Both
Chronic Obstructive Pulmonary Disorder-Like Symptoms	3,372	4.4%	6,615	6.5%	Both
Coronary Revascularization	1,749	2.3%	4,053	4.00%	Both
Depressive Disorder	3,989	5.3%	6,952	6.8%	Both
Gastroesophageal Reflux Disease	5,456	7.2%	10,029	9.9%	Both
Heart Failure	3,094	4.1%	8,970	8.8%	Both
Hypercholesterolemia/Hyperlipidemia	18,123	24%	36,712	36%	Both
Hypertension/Hypertensive Disorder	24,154	32%	45,443	45%	Both
Malignant Neoplastic Disease	2,121	2.8%	3,590	3.5%	Both
Nicotine Dependency	3,348	4.4%	6,265	6.2%	EHR
Obesity	9,155	12%	20,377	20%	Both
Osteoarthritis	4,831	6.4%	9,486	9.3%	Both
Stroke (Ischemic/Hemorrhagic)	1,166	1.5%	1,742	1.7%	Both
Thromboembolism	659	0.87%	1,294	1.3%	Both
Transient Ischemic Attack	385	0.51%	658	0.65%	Both

*This column specifies the source (claims versus EHR) of each variable; it is possible that this information comes from either, both or neither of the two sources.

[a]Defined as difference between first recorded diagnosis for Type 2 DM and cohort entry date

[b]Medication history was determined based on the earliest available record. Dispensed date was used if it was the earliest or occurred within 90 days of the earliest prescribed date; otherwise, the prescribed date was used.

Table 3 Clinical characteristics, lifestyle factors, and vital signs of the DPP-4i and SGLT-2i initiators with type 2 diabetes mellitus

Patient Characteristics	**DPP-4 inhibitor initiators with Type 2 DM**		**SGLT-2 inhibitor initiators with Type 2 DM**		**Provenance* (EHR/Claims/Both)**
	N/Mean	**%/SD**	**N/Mean**	**%/SD**	
Clinical characteristics					
Ejection fraction recorded, N(%)	3,357	4.4%	5,312	5.2%	EHR
LVEF with valid value (any units), N (%)	3,296	4.3%	5,201	5.1%	EHR
LVEF (%), N (%)	3,290	4.3%	5,190	5.1%	EHR
LVEF (%), mean (SD)	54.14	14.2	51.9	15.7	EHR
LVEF (missing units), N (%)	6	0.01%	11	0.01%	EHR
LVEF (missing units), mean (SD)	55.83	8.9	49.63	9.0	EHR
HbA1c test recorded, N(%)	48,054	63%	67,350	66%	
HbA1c with valid lab value (any units), N (%)	47,711	63%	66,700	66%	EHR
HbA1c (%), N (%)	37,982	50%	58,085	57%	EHR
HbA1c (%), mean (SD)	7.9	1.7	8.2	1.9	EHR
HbA1c (missing units), N (%)	9,729	13%	8,615	8.5%	EHR
HbA1c (missing units), mean (SD)	8.1	1.9	8.2	1.9	EHR
Serum creatinine test recorded, N (%)	51,268	67%	71,328	70%	
Serum creatinine with valid lab value (any units), N (%)	51,254	67%	71,321	70%	EHR
Serum creatinine (mg/dl), N (%)	47,050	62%	68,265	67%	EHR
Serum creatinine (mg/dl), mean (SD)	1.30	1.0	1.14	0.6	EHR
Serum creatinine (missing units), N (%)	4,204	5.5%	3,056	3.0%	EHR
Serum creatinine (missing units), N (%)	1.016	0.7	0.964	0.5	EHR
Total cholesterol test recorded, N (%)	38,519	51%	55,597	55%	
Total cholesterol with valid lab value (any units), N (%)	38,492	51%	55,557	55%	EHR
Total cholesterol (mg/dl), N (%)	33,782	44%	52,236	51%	EHR
Total cholesterol (mg/dl), mean (SD)	168.563	46	169.432	48	EHR
Total cholesterol (missing units), N (%)	4,710	6.2%	3,321	3.3%	EHR
Total cholesterol (missing units), mean (SD)	163.136	46	164.898	47	EHR
LDL cholesterol test recorded, N (%)	38,482	51%	56,158	55%	

LDL cholesterol with valid lab value (any units), N (%)	37,867	49.8%	55,052	54%	EHR
LDL cholesterol (mg/dl), N (%)	32,189	42%	50,077	49%	EHR
LDL cholesterol (mg/dl), mean (SD)	89.5	35.9	91.1	38.8	EHR
LDL cholesterol (missing units), N (%)	5,678	7.5%	4,975	4.9%	EHR
LDL cholesterol (missing units), mean (SD)	85.7	41.0	87.3	44.7	EHR
HDL cholesterol test recorded, N (%)	37,965	50%	55,337	55%	
HDL cholesterol with valid lab value (any units), N (%)	37,889	50%	55,154	54%	EHR
HDL cholesterol (mg/dl), N (%)	32,550	43%	51,178	50%	EHR
HDL cholesterol (mg/dl), mean (SD)	44.0	12.8	431	12.3	EHR
HDL cholesterol (missing units), N (%)	5,339	7%	3,976	3.9%	EHR
HDL cholesterol (missing units), mean (SD)	44.2	18.3	43.1	16.9	EHR
Triglycerides test recorded, N (%)	39,844	52.4%	56,940	56%	
Triglycerides with valid lab value (any units), N (%)	39,330	51.7%	56,108	55.2%	EHR
Triglycerides (mg/dl), N (%)	34,246	45.1%	52,228	51.4%	EHR
Triglycerides (mg/dl), mean (SD)	182.4	93.2	184.5	95.2	EHR
Triglycerides (missing units), N (%)	5,084	6.7%	3,880	3.8%	EHR
Triglycerides (missing units), mean (SD)	177.8	87.2	182.8	91.5	EHR
Lifestyle factors					
Smoking Recorded, N (%)	17,533	23%	18,758	19%	EHR
Alcohol Use Recorded, N (%)	5,049	6.6%	6,355	6.3%	EHR
Drug Misuse and Abuse Recorded, N(%)	4,211	5.5%	6,043	6.0%	EHR
Vitals					
BMI Recorded, N(%)	21,622	28.4%	46,598	46%	EHR
BMI, mean (SD)	33.4	8.0	34.1	8.0	EHR
BMI low, < 18.5 kg/m^2	124	0.16%	168	0.17%	EHR

BMI medium, 18.5 to 24.9 kg/m^2	2,230	2.9%	4,028	4.0%	EHR
BMI high, >24.9 kg/m^2	19,268	25%	42,402	42%	EHR
Heartrate Recorded, N(%)	20,675	27%	47,054	46%	EHR
Heartrate, BPM, mean (SD)	80.325	14	80.00	14.215	EHR
Diastolic Blood Pressure Recorded, N (%)	24,587	32.3%	51,422	50.6%	EHR
Diastolic Blood Pressure, mean (SD)	78	12	77	11.4	EHR
Systolic Blood Pressure Recorded, N (%)	24,591	32%	51,423	51%	EHR
Systolic Blood Pressure, mean (SD)	134.1	19.5	132.7	18.8	EHR

*This column specifies the source (claims versus EHR) of each variable; it is possible that this information comes from either, both or neither of the two sources

LVEF- Left Ventricular Ejection Fraction; HbA1c- Hemoglobin A1c; LDL- Low Density Lipoprotein; HDL- High Density Lipoprotein; BMI- Body Mass Index; BPM- Beats Per Minute

Table 4 Healthcare utilization features of the DPP-4i and SGLT-2i initiators with type 2 diabetes mellitus

[a]**Patient Characteristics**	**DPP-4 inhibitor initiators with Type 2 DM**		**SGLT-2 inhibitor initiators with Type 2 DM**		**Provenance* (EHR/Claims/Both)**
Patients with at least 1 in-person EHR encounter	**N**	**%**	**N**	**%**	
In the past 365 days, n (%)	48,792	64%	77,441	76%	Both
In the past 180 days, n (%)	47,340	62%	75,619	74%	Both
In the past 90 days, n (%)	45,484	60%	73,067	72%	Both
In the past 30 days, n (%)	41,560	55%	67,923	67%	Both

*This column specifies the source (claims versus EHR) of each variable; it is possible that this information comes from either, both or neither of the two sources

[a] Operationalized based on patient care setting at the time of encounter from EHR and Evaluation and Management CPT codes with modifiers from claims data