Empirical Evaluation of EHR-Enhanced Signal Detection Using Tree Based Scan Statistic Methods

Sentinel Protocol

Primary Investigators
MGB: Shirley V Wang
VUMC: Joshua Smith

Contributors
MGB: Massimiliano Russo, Sushama Kattinakere Sreedhara, Joyce Li, Thomas Deramus, Rishi Desai
VUMC: Sharon Davis, Ruth Reeves, Daniel Park, Robert Winter, Jill Whittaker
SOC: Judy Maro, Audrey Wolfe Evans, Sam McGown
FDA: Jose Hernandez, Yong Ma, Youjin Wang, Jamal Jones

Version 4.0
May 17, 2024

1. Title Page

Title	Empirical Evaluation of EHR-Enhanced Signal Detection Using Tree Based Scan Statistic Methods
Research Question & Objectives	Our objective is to examine the performance of propensity score matched tree-based scan statistics (TBSS) with laboratory outcomes as well as NLP derived outcomes in the outcome tree for one empirical example where there is an expected effect that could be detected in billing codes; however, laboratory values and clinical notes may add greater sensitivity for detecting sub-clinical effects. The example we will focus on is second generation sulfonylureas vs dipeptidyl peptidase-4 inhibitors, where the expected adverse effect is hypoglycemia.
Protocol Version	4
Last Update Date	May 17, 2024
Contributors	**Primary investigator contact information:** MGB: Shirley Wang, swang1@bwh.harvard.edu VUMC: Joshua Smith, joshua.smith@vumc.org **Contributor names:** MGB: Massimiliano Russo, Sushama Kattinakere Sreedhara, Joyce Li VUMC: Sharon Davis, Ruth Reeves, Daniel Park, Robert Winter, Jill Whittaker SOC: Judy Maro, Audrey Wolfe, Sam McGown FDA: Jose Hernandez, Yong Ma, Youjin Wang, Jamal Jones
Study Registration	**Site:** N/A **Identifier:** N/A
Sponsor	**Organization:** United States Food and Drug Administration, Sentinel Initiative, Contract number: 75F40119D10037 **Contact:** Meighan Rogers Driscoll, meighan.rogers.driscoll@point32health.org
Conflict of Interest	Dr. Wang has consulted for Veracity Healthcare Analytics, Exponent Inc, and MITRE, an FFRDC for the Centers for Medicare and Medicaid Services for unrelated work.

Table of Contents

1. Title Page **2**
2. Abstract **4**
3. Amendments and Updates **5**
4. Milestones **6**
Table 1 Milestones 6
5. Rationale and Background **6**
6. Research Question and Objectives **7**
Table 2 Methods research questions and objective (exploratory) 7
7. Research Methods **8**
7.1. Study design 8
7.2. Study design diagram 9
7.3. Setting 10
7.3.1 Context and rationale for definition of time 0 (and other primary time anchors) for entry to the study population 10
Table 3 Operational Definition of Time 0 (index date) and other primary time anchors 10
7.3.2 Context and rationale for study inclusion criteria: 10
Table 4. Operational Definitions of Inclusion Criteria 10
7.3.3 Context and rationale for study exclusion criteria 11
Table 5. Operational Definitions of Exclusion Criteria 11
7.4. Variables 13
7.4.1 Context and rationale for exposure(s) of interest 13
Table 6. Operational Definitions of Exposure 15
7.4.2 Context and rationale for outcome(s) of interest 16
Table 7. Operational Definitions of Outcome 17
7.4.3 Context and rationale for follow up 17
Table 8. Operational Definitions of Follow Up 18
7.4.4 Context and rationale for covariates (confounding variables and effect modifiers, e.g. risk factors, comorbidities, comedications) 18
Table 9. Operational Definitions of Covariates 19
7.5. Data analysis 31
7.5.1 Context and rationale for analysis plan 31
Table 10. Analysis specification 32
7.6. Data sources 33
7.6.1 Context and rationale for data sources 33
Table 12. Metadata about data sources and software 35
7.7. Data management 36
7.8. Quality control 40
7.9. Study size and feasibility 40
8. Limitations of the Methods **40**
9. Protection of Human Subjects **40**
10. Reporting of Adverse Events **40**
11. References **41**
12. Appendices **42**

2. Abstract

Tree-based scan statistics (TBSS) are a data mining method that use a hierarchical tree to group and map relationships between thousands of correlated outcomes and appropriately adjust for multiple hypothesis testing when screening to prioritize alerts for further investigation.[1,2] Prior work has included evaluation of the method with different study designs using both simulated and empirical data. [3-6] Although TBSS can be used to scan exposures as well as outcomes, it is most commonly used to simultaneously evaluate thousands of potential adverse events (and groups of related adverse events) for associations with a drug of interest. After identifying a drug of interest and an appropriate comparator, propensity score matching can be used to adjust for confounders.[4]

From a methodological perspective, there are a few major differences when conducting drug safety surveillance with data from electronic health records (EHRs) as opposed to insurance claims data. In insurance claims data, we define incident outcomes such as a stroke or a seizure, using date stamped diagnosis or other billing codes. Such billing code data is also available in EHRs, and we could apply the same drug safety surveillance methods as used for insurance claims data to detect safety signals. Historically, TBSS has been applied to screen for potential adverse events captured with ICD9 and ICD10 codes. However, an untapped advantage of EHRs in post-market safety surveillance activities is the rich clinical information, including laboratory test results, vital signs, and clinical notes.

This project will develop approaches for abstracting and combining structured and unstructured EHR data as well as expanding TBSS methods to also identify signals for outcomes more suitable to be identified through EHR data (e.g., natural language processing, laboratory values).

For the study described in this protocol, we will examine the performance of propensity score matched TBSS with laboratory value outcomes as well as natural language processing (NLP) derived outcomes in the outcome tree for an applied example where there is an expected adverse effect that could be detected in billing codes; however, laboratory values and clinical notes may add greater sensitivity for detecting sub-clinical effects. The example that we will focus on is second generation sulfonylureas versus dipeptidyl peptidase-4 inhibitors, where the expected adverse effect is hypoglycemia. This example was selected collaboratively with the FDA.

We will compare the results from applying TBSS with a propensity score matched cohort where 1) the outcome tree is based only on billing coded data, 2) after adding NLP derived outcomes to the outcome tree, and 3) after adding laboratory results to the outcome tree.

We will use existing ARIA tools to extract cohorts for the empirical example(s). Any code developed during this project will be made publicly available on Sentinel's Git site.

3. Amendments and Updates

Version date	Version number	Section of protocol	Amendment or update	Reason
Jan 27 2024	1	Section 6 & Section 7	DPP-4 inhibitors group as comparator rather than Sitagliptin, Instead of concomitant Metformin use, Metformin used as covariate	Not enough sample size
Feb 7 2024	2	Section 6 & 7	Instead of excluding other anti-diabetic medications, use them as covariate. Use 180 days baseline period, instead of 365 days. Types 2 Diabetes at least 1 diagnosis code in any care setting (rather than 1 IP or 2 OP diagnosis)	Not enough sample size

Apr 27 2024	3	Section 6 & 7	• Drop exclusion for prior hypoglycemia [-30, 0] and change hypoglycemia covariate assessment window to [-180, 0].
			• Revise covariates as follows:
			  • Add low-income subsidy indicator
			  • Add NDC codes to diagnosis codes for defining selected covariates
			  • Measure most recent hbA1c value as continuous variable with missing indicator
			• Change timeline for milestones to reflect no-cost extension
				Feedback from workgroup and FDA

May 17 2024	4	Appendices, design diagram	• Clarifications and corrections as requested by FDA	Clarifications and corrections as requested by FDA

4. Milestones

Table 1 Milestones

Milestone	Date
Workgroup kick-off meeting	Feb 2023
Complete protocol prepared	April 2024
Submission ready manuscript prepared (NLP portable pipeline)	Oct 2024
Submission ready manuscript prepared (empirical example)	Oct 2024

5. Rationale and Background

What is known about the condition: Hypoglycemia is a common adverse event related to glucose lowering medications for patients with type II diabetes. It is associated with occurrence of other severe adverse health outcomes and mortality and contributes to excess hospital admissions and emergency department visits in older adults [1,2].

What is known about the exposure of interest: Second generation sulfonylureas (SUs) have a higher risk of hypoglycemia than DPP-4 inhibitors.[3].

What is known about the method: Tree-based scan statistics (TBSS) are a data mining method that use a hierarchical tree to group and map relationships between thousands of correlated outcomes and appropriately adjust for multiple hypothesis testing when screening to prioritize alerts for further investigation[3,4]. Prior work has included evaluation of the method with different study designs using both simulated and empirical data. [3-6] Although TBSS can be used to scan exposures as well as outcomes, it is most commonly used to simultaneously evaluate thousands of potential adverse events (and groups of related adverse events) for associations with a drug of interest. After identifying a drug of interest and an appropriate comparator, propensity score matching can be used to adjust for confounders.[7]

Gaps in knowledge: From a methodological perspective, there are a few major differences when conducting drug safety surveillance with data from electronic health records (EHRs) as opposed to insurance claims data. In insurance claims data, we define incident outcomes such as a stroke or a seizure, using date stamped diagnosis or other billing codes. Such billing code data is also available in EHRs, and we could apply the same drug safety surveillance methods as used for insurance claims data to detect safety signals. Historically, TBSS has been applied to screen for potential adverse events captured with ICD9 and ICD10 codes. However, an untapped advantage of EHRs in post-market safety surveillance activities is the rich clinical information, including laboratory test results, vital signs and clinical notes.

What is the expected contribution of this study? This study will examine the performance of propensity score matched TBSS with laboratory value based outcomes as well as natural language processing (NLP) derived outcomes in an outcome tree for an applied example where there is an

expected adverse effect that could be detected in billing codes, however laboratory values and clinical notes may add greater sensitivity for detecting sub-clinical effects. The example that we will focus on is dipeptidyl peptidase-4 inhibitors (DPP-4is) versus second generation sulfonylureas, where the expected adverse effect is hypoglycemia. This example was selected collaboratively with the FDA.

We will compare the results from applying TBSS with a propensity score matched cohort where 1) the outcome tree is based only on billing coded data, 2) after adding NLP derived outcomes to the outcome tree, and 3) after adding laboratory measurements to the outcome tree.

6. Research Question and Objectives

Table 2 Methods research questions and objective (exploratory)

Objective:	To compare the results from applying TBSS with a propensity score matched cohort of DPP-4is versus 2nd generation sulfonylurea users where 1) the outcome tree is based only on billing coded data, 2) after adding NLP derived outcomes to the outcome tree, and 3) after adding laboratory measurements to the outcome tree.
Hypothesis:	This is an exploratory, signal identification study where we aim to see whether the addition of NLP and laboratory defined outcomes increases the ability of TBSS to detect the anticipated signal of increased hypoglycemia risk with 2nd generation sulfonylureas versus DPP4-inhibitors.
Population *(mention key inclusion-exclusion criteria):*	Patients that are 18 years or older, with type 2 diabetes mellitus, and new initiators of DPP-4is or 2nd generation Sulfonylureas (SUs)
Exposure:	2nd generation SUs (Glipizide, Glyburide, Glimepiride, Gliclazide)
Comparator:	DPP-4is (Sitagliptin, Saxagliptin, Linagliptin, Alogliptin)
Outcome:	All claims diagnosis code, NLP, and lab defined outcomes included in the outcome tree used for TBSS signal identification. Hypoglycemia is an anticipated true effect. We are evaluating whether the different outcome trees would facilitate identification of this effect in a screening activity.
Time *(when follow up begins and ends):*	Follow-up starts one day after initiation of therapy until death, disenrollment, 180 days after initiation, end of study period, switch, discontinuation.
Setting:	We will explore trees that include outcomes using inpatient/emergency department data only versus all care settings
Main measure of effect:	Relative risk and relative difference

7. Research Methods

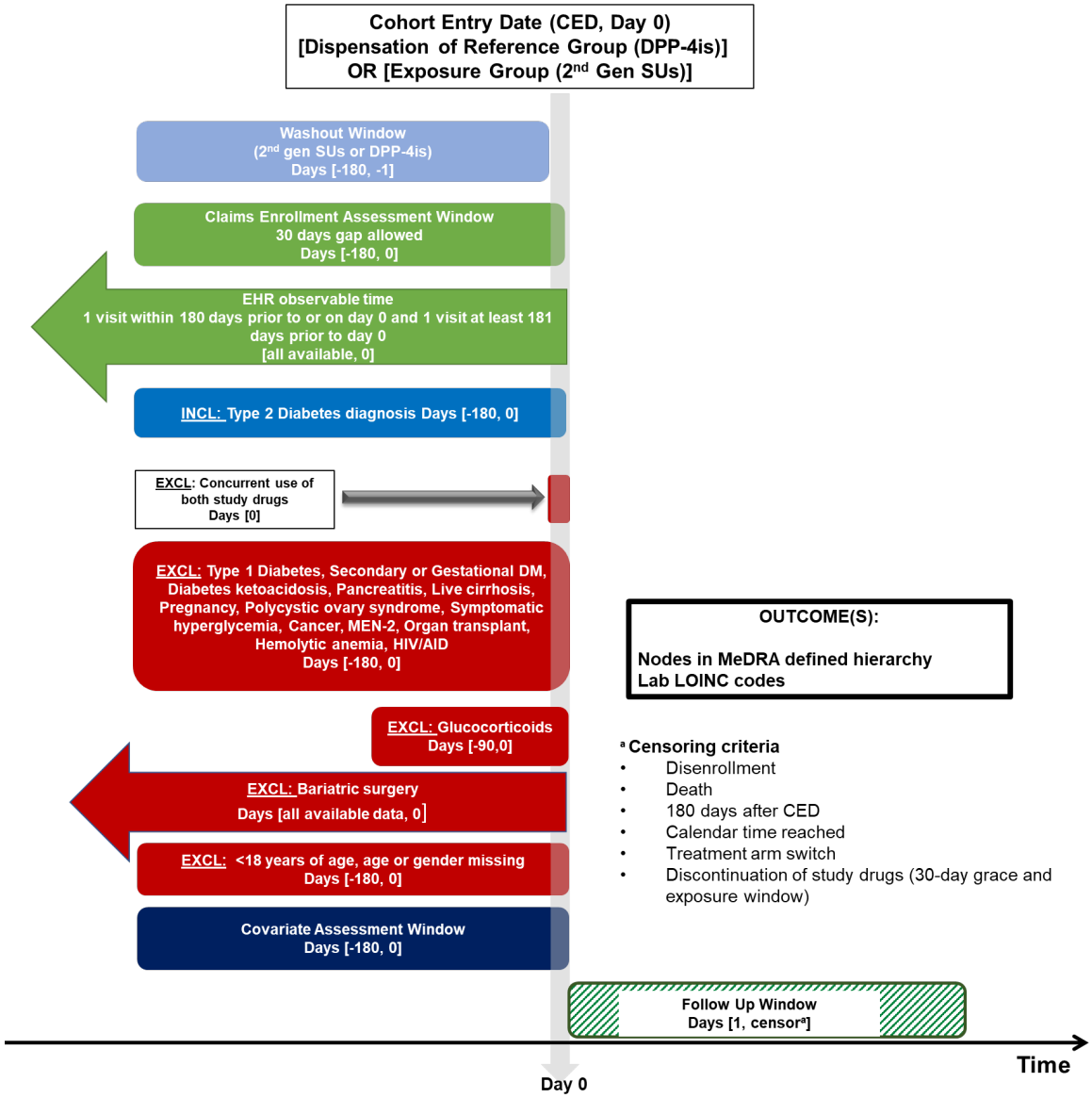
7.1. Study design

Research design (e.g. cohort, case-control, etc.): New user, active comparator cohort study

Rationale for study design choice: The propensity score matched new user, active comparator cohort design is currently being used by the FDA Sentinel Program in active surveillance activities.[8-14] In this context, a cohort is generally defined by incident exposure to a new drug or an appropriate comparator drug. The initiators of each drug are then matched based on a propensity score, which is a summary measure capturing the probability of being an initiator of the new drug of interest rather than the comparator drug based on multiple baseline characteristics.

This study design identifies comparable patient populations and is appropriate for evaluation of safety starting from the time of initiation of treatment. The index date for study entry will be defined based on initiation of either the 2nd generation SUs or DPP-4is after a specified washout period of 180 days during which the patient has anticipated observability in the source database and no recorded use of either study drugs. We require 180 days continuous enrollment in claims data prior to the index date in order to capture baseline characteristics, identified via diagnoses and procedures during healthcare encounters. This new initiator design establishes a clear temporal sequence from treatment exposure to outcome and allows capture of events that occur shortly after treatment initiation; these events would be missed if prevalent users were allowed to enter the cohort. Cohorts will be restricted to patients with the relevant indication for type II diabetes. Patients with other types of diabetes will be excluded.

7.2. Study design diagram



7.3. Setting

7.3.1 Context and rationale for definition of time 0 (and other primary time anchors) for entry to the study population

Time 0 is the date of initiation of DPP-4is or 2nd gen SUs after a 180-day washout when the patients were not exposed to either DPP-4is or 2nd generation SUs.

Table 3 Operational Definition of Time 0 (index date) and other primary time anchors

Study population name(s)	Time Anchor Description (e.g. time 0)	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type[2]	Diagnosis position	Incident with respect to…	Measurement characteristics/ validation	Source of algorithm
Exposure: 2nd gen SUs	Date of incident dispensation for SUs	Single	Incident	[-180, -1]	n/a	NDC	n/a	Any formulation of DPP4 inhibitor or 2nd generation SU	n/a	Investigators review of generic names
Reference: DPP-4is	Date of incident dispensation for DPP-4is	Single	Incident	[-180, -1]	n/a	NDC	n/a	Any formulation of DPP4 inhibitor or 2nd generation SU	n/a	Investigators review of generic names

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable

[2]See appendix 5 for listing of clinical codes for each study parameter

7.3.2 Context and rationale for study inclusion criteria:

We require six months of medical and prescription enrollment coverage (allowing 30-day gaps) prior to time 0 to ensure that patients have observable time in the data where contact with the healthcare system will allow capture of clinical codes to measure inclusion-exclusion criteria and covariates. We include patients with type 2 diabetes.

Table 4. Operational Definitions of Inclusion Criteria

Criterion	Details	Order of application	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Observable time (claims)	Medical and prescription coverage (up to 30-	Before selection of index date	[-180, 0]	n/a	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is	n/a	n/a

	day gaps allowed)								
Observable time (EHR)	At least 1 visit within 180 days prior to or on index and 1 visit any time prior to 180 day look back	Before selection of index date	[-∞, -0	n/a	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is	n/a	n/a
Type 2 Diabetes	At least 1 diagnosis in any care setting	Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is	n/a	https://pubmed.ncbi.nlm.nih.gov/36745425/

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix 5 for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.3.3 Context and rationale for study exclusion criteria

We excluded patients under the age of 18, if they have had diagnosis of type 1 diabetes, secondary diabetes, gestational diabetes, diabetic ketoacidosis, selected other comorbidities as specified below, or if age/gender is missing in the baseline assessment window.

Table 5. Operational Definitions of Exclusion Criteria

Criterion	Details	Order of application	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Age missing or less than 18 years		Before selection of index date	[0, 0]	n/a	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		n/a
Gender missing		Before selection of index date	[0, 0]	n/a	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		n/a

Type I Diabetes	Contraindication to SU [12,13]	Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/36745425/
Diabetes Ketoacidosis	Contraindication to SU [12, 13, 14]	Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/36745425/
Pancreatitis	If pancreatitis is suspected, Sitagliptin needs to be discontinued [15]	Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigators review of diagnosis codes
Liver cirrhosis[6]	Higher risk of hypoglycaemia among SUs users with alcoholism/cirrhosis	Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34256144/ and https://pubmed.ncbi.nlm.nih.gov/22674685/
Gestational Diabetes		Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34729891/
Secondary Diabetes		Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34729891/
Pregnancy		Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigators review of diagnosis and procedure codes
Polycystic ovary syndrome		Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34953068/
Symptomatic hyperglycaemia		Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34953068/ and Investigator review of ICD10 codes
Bariatric surgery		Before selection of index date	[-all available data, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigators review of procedure codes

History of MEN-2		Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigators review of diagnosis codes
Cancer		Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34729891/
Organ transplant		Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34729891/
Use of glucocorticoids		Before selection of index date	[-90, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		Investigators review of NDC codes
Haemolytic anaemia		Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		Investigators review of diagnosis codes
HIV/AIDS	Patients are too sick and on many other drugs	Before selection of index date	[-180, 0]	Any	ICD9-CM, ICD10-CM, NDC	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34729891/ https://hivinfo.nih.gov/understanding-hiv/fact-sheets/fda-approved-hiv-medicines

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix 5 for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4. Variables

7.4.1 Context and rationale for exposure(s) of interest

We focus on new initiators of drugs which are commonly used anti-diabetic drugs at a similar stage of type 2 diabetes.

Algorithm to define duration of exposure effect:

Assuming the biological effect of the therapies lasts for 30 days after the days' supply, we allow 30 days gap between dispensation's + days supply and the subsequent dispensation (grace period) and also add 30 days at the end of the final days' supply in a treatment episode (exposure risk window). A stockpiling

algorithm will be used to account for dispensings in the exposure or comparator group with overlapping days of supply. Any overlap of supply between dispensings will be corrected by pushing the start date of the second dispensing to occur following the end of the days supplied for the first dispensing. Only the first episode for each person will be included in the analysis.

Table 6. Operational Definitions of Exposure

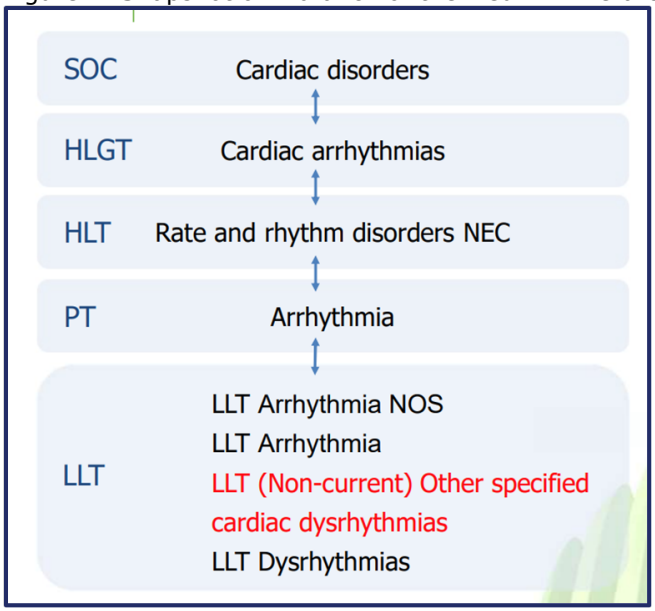
Exposure group name(s)	Details	Washout window	Assessment Window	Care Setting[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Incident with respect to…	Measurement characteristics/ validation	Source of algorithm
Exposure	2nd Gen Sus	[-180, -1]	[1, censor]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is	DPP-4is OR 2nd Gen SUs	n/a	Investigators review of generic names
Comparator	DPP4-is	[-180, -1]	[1, censor]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is	DPP-4is OR 2nd Gen SUs	n/a	Investigators review of generic names

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix 5 for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4.2 Context and rationale for outcome(s) of interest

The "tree" in TBSS refers to a classification system that hierarchically groups coded clinical concepts into clinically related categories. Example clinical coding systems that can be used include the International Classification of Diseases (ICD), Multi-Level Clinical Classification (MLCC)[15] for ICD codes, or the Medical Dictionary for Regulatory Activities (MedDRA) classification system, where each grouping of clinically related concepts represents an outcome "node" in the hierarchical tree. Use of a hierarchical tree structure increases power to detect safety signals in clinically related diagnosis groupings that would not be feasible by evaluating individual diagnoses.

Figure 2. Snapshot of 1 branch of the MedDRA hierarchy



SOC = system organ class
HGLT = high level group term
HLT = high level term
PT = preferred term
LLT = lowest level term

MedDRA: This analysis will use the MedDRA hierarchy as the basis for an outcome tree. MedDRA is a standardized terminology used by regulatory authorities and the biopharmaceutical industry in pre- and post-marketing surveillance.[16] It contains concepts including diseases, diagnoses, signs, and symptoms, among others. It also includes a robust hierarchy describing relationships between its concepts at varying levels of granularity. In this project, we will use MedDRA to encode structured and unstructured EHR data for use with TBSS. In addition to the inbuilt hierarchy which TBSS can utilize, MedDRA can also be used with standard biomedical natural language processing (NLP) tools to identify concepts mentioned in clinical text and normalize them to appropriate levels of specificity.

Mapping ICD-10-CM codes to MedDRA: Mapping ICD-10-CM codes used in EHR diagnosis codes and claims data will require the mapping of ICD codes to MedDRA concepts. The VUMC team will create a portable pipeline to produce outcome counts for a hierarchical MedDRA tree based on structured claims + EHR data combined with NLP-extracted data from unstructured clinical notes. Several mappings between the vocabularies currently exist, but the quality and completeness of these mappings is not certain. We will therefore combine multiple existing mappings, including those from MedDRA, the National Library of Medicine (NLM), and those present in the Unified Medical Language System (UMLS), to create a more-complete mapping. To improve current efforts, we will also utilize existing mappings from ICD-10-CM to SNOMED-CT (an ontology of concepts used for clinical notes documentation) and SNOMED to MedDRA, using SNOMED as an intermediary when necessary. To validate our new mapping, we will enlist a chart reviewer to manually review our results as much as possible. We will also create the necessary MedDRA hierarchy files for use with TBSS. The ICD-MedDRA mapping will be used to translate VUMC diagnosis codes (and, potentially, claims data from Tennessee Medicaid obtained from the Sentinel Innovation Center DI7 project, if possible) to MedDRA concepts. Details of the mapping are provided in appendix 1.

Mapping NLP extracted terms to MedDRA: Full text notes will require proper formatting and data-cleaning before being processed by our NLP tools. We will write code to accomplish this with a focus on portability and scalability. Additionally, some types of EHR notes contain more relevant information than others; for example, documents containing only discharge instructions are likely to contain disease descriptions or warning to seek care if certain symptoms are experienced. We will therefore need to accurately identify each note's type (which is not always apparent) and determine which note types are of most value. We will therefore leverage the work of DI7 to map site-specific note titles to generic note types which can be excluded or included depending on study needs. Once the notes are cleaned, appropriately labeled, and processed, we will combine the extracted MedDRA concepts with the mapped MedDRA claims/diagnosis code data described above. Depending on analysis, it may be advisable to require extracted MedDRA concepts to appear more than once in a note/patient to ensure veracity. Details of the mapping are provided in appendix 2.

After mapping ICD-10 and NLP extracted terms to the MedDRA tree, the tree will be "pruned" to remove branches related to congenital, familial and genetic disorders, cancer, pregnancy, and social circumstances, which are not expected to contain outcomes that may be causally related to the exposure and comparator of interest.

Mapping laboratory values to MedDRA: A subset of labs relevant to the type II diabetes and/or the target outcome of hypoglycemia will be extracted and included in the outcome tree. Results will be extracted as continuous lab values, and lab result ranges and reporting units will be standardized. These will be included in the outcome tree as their own leaf level without mapping. As a proof of concept, we will also map out-of-range values for a small subset of lab tests to MedDRA concepts, which can then be included as standard MedDRA codes similar to the ICD and NLP-extracted concepts (i.e., high serum potassium values will be mapped to the concept for Hyperkalemia). Details of the mapping are provided in appendix 3.

Outcome tree: We will combine outcomes derived from multi-modal data. This will result in the creation of three separate outcome trees: 1) MedDRA concepts from claims/billing codes, 2) MedDRA concepts from claims/billing codes + NLP mapped concepts from unstructured notes, 3) MedDRA concepts from claims/billing codes + NLP mapped concepts from unstructured notes + selected laboratory data (mapped and native continuous values). These data will be formatted as necessary, with MedDRA preferred terms and dates of occurrence, to be used as input for TBSS. Details of the outcome tree and node specific washout to define incident outcomes are provided in appendix 4.

We will explore how results vary depending on how the outcomes are defined, including defining outcomes based on incident IP or ED encounters versus incident outcomes defined in any care setting. The washout window for all outcomes will be 180 days prior to and including the index date. For analyses where the outcome tree includes only claims data, the washout will be applied using claims data only. For analyses where the outcome tree includes claims and NLP extracted data, the washout will use the same sources. Likewise for analyses where the outcome tree includes claims, NLP and lab-based data, the washout will be applied across all 3 data types.

Table 7. Operational Definitions of Outcome

Outcome name	Details	Primary outcome?	Type of outcome	Washout window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source of algorithm
Nodes in MedDRA defined hierarchy		n/a	binary	[-180, 0]	1) IP, ED vs 2) Any	ICD9-CM, ICD10-CM, NLP extracted concepts, targeted	Any	Exposure: 2nd Gen SUs Reference: DPP-4is		Nodes defined by ICD-10 codes and NLP extracted terms mapped to MedDRA according to

						extracts of lab results that are out of range				according to pipeline developed by VUMC team
Laboratory LOINC codes for diabetes related labs - continuous	Continuous values (change in value between most recent value prior to index and unique measurement during follow up that is within plausible range of expected values for lab test)	n/a	Continuous	n/a (most recent value [-180, 0])	n/a	LOINC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		Defined by reported lab value
Laboratory LOINC codes for diabetes related labs-binary	Binary values after mapping abnormal test result values to MedDRA	n/a	binary	[-180, 0]	n/a	LOINC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		Defined by reported lab value

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix 5 for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4.3 Context and rationale for follow up

We focus on an as-treated analysis, therefore, in addition to censoring on death, disenrollment, and end of data, we censor patients upon discontinuation or switching of treatment.

Table 8. Operational Definitions of Follow Up

	Select all that apply	Specify
Follow up start	Day 1	
Follow up end[1]		
Date of outcome	No	Follow up distributions will be checked to see if they are equivalent between the compared treatment arms. All outcomes will be counted if they meet incidence criteria and occur within the outcome agnostic follow up window.
Date of death	Yes	
End of observation in data	Yes	Allowing 30-day gaps in enrollment
Day X following index date *(specify day)*	Yes	Day 180
End of study period (specify date)	Yes	End of data availability is:
End of exposure *(specify operational details, e.g. stockpiling algorithm, grace period)*	Yes	**Stockpiling algorithm:** If next refill occurs before end of days' supply, then count overlapping days at the end of the subsequent dispensing's day supply. **Grace period:** 30 days gaps is allowed between last date of supply and next dispensing **Exposure risk window:** Add 30 days at the end of the treatment episode ***Note:*** *Exposure here includes DPP4-is and 2nd Gen SUs.* *Patients in the DPP4-is arm are censored upon discontinuing either DPP-4is. Likewise for the 2nd generation arm, censoring occurs upon discontinuation the 2nd generation SU*
Date of add to/switch from exposure *(specify algorithm)*	Yes	Date when exposed group is dispensed comparator drug or vice versa ***Note****: 2nd gen SUs can be switched within the class of 2nd gen SU and DPP-4is can be switched within the class as well).*
Other date *(specify)*	No	

[1] Follow up ends at the first occurrence of any of the selected criteria that end follow up.

7.4.4 Context and rationale for covariates (confounding variables and effect modifiers, e.g. risk factors, comorbidities, comedications)

We identified demographic, comorbidities, healthcare utilization and frailty related risk factors of outcome that were associated with the exposure and reference group.

Table 9. Operational Definitions of Covariates

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Age	(cohort entry year - year of birth)	Continuous	[0,0]	n/a	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		Master beneficiary file
Gender	Male, Female	Categorical	[0, 0]	n/a	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		Master beneficiary file
Region	Northeast, South, Midwest, West	Categorical	[0, 0]	n/a	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		Master beneficiary file
Race	White, Black, Asian, Hispanic, Other or Missing	Categorical	[0, 0]	n/a	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		Master beneficiary file
Calendar year	Date of cohort entry-	Categorical	[0, 0]	n/a	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		n/a
Baseline hbA1c	Most recent value	Continuous with missing indicator and count of labs measured in assessment window (latter not in PS)	[-180, 0]	any	lab	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		
Hypoglycemia		Binary and days with hypoglycemia diagnosis (latter not	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/35043165/

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
		included in PS)							
Smoking	Defined by claims OR smoking from vital status file of EHR is currently smoking	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002 and Investigator choice of codes
Obesity or Overweight		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002 and Investigator choice of codes
Hypertension		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002
Hyperlipidaemia		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002
Myocardial infarction	Defined by acute MI/ old MI	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002
Angina	Defined by unstable angina/stable angina	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34570599/ and RCT DUPLICATE Investigator review of codes
Coronary atherosclerosis		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Other forms of chronic ischemic heart disease		Binary	[-180, 0]	any	ICD9-CM,	any	Exposure: 2nd Gen SUs Reference:		RCT DUPLICATE Investigator review of codes

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
					ICD10-CM		DPP-4is		
History of CABG/PTCA		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Any Stroke	Defined by any stroke/ tia/ late effects of cvd	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002 and RCT DUPLICATE Investigator review of codes
Heart failure		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/35043165/
Peripheral artery disease	Defined by diagnosis or surgery	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Atrial fibrillation		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002 and RCT DUPLICATE Investigator review of codes
Other cardiac dysrhythmia	Not used in PS/Fine stratification model	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002 and RCT DUPLICATE Investigator review of codes
Cardiomyopathy		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Diabetes mellitus without mention of complications		Binary	[-180, 0]	any	ICD9-CM,	any	Exposure: 2nd Gen SUs Reference:		RCT DUPLICATE Investigator review of codes

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
					ICD10-CM		DPP-4is		
Diabetic with complications	Defined by diabetic foot/ diabetic nephropathy/ diabetic neuropathy/ diabetic retinopathy/ unspecified complications / peripheral circulatory disorders	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/35043165/ and RCT DUPLICATE Investigator review of codes
Lower limb amputations		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/36745425/ and RCT DUPLICATE Investigator review of codes
Hyperosmolar hyperglycaemic nonketotic syndrome		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
CKD stage 1-2		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/35043165/
CKD stage 3-5		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/35043165/
ESRD		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002
CKD stage NOS		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Miscellaneous renal disease	Not used in PS/Fine stratification model	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Dementia	Defined by Alzheimer and other dementia disease and dementia medications	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002
Parkinson's disease	Defined by diagnosis codes/ antiparkinson medications	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM, NDC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002
Delirium		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/35043165/ and RCT DUPLICATE Investigator review of codes
Psychosis		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/35043165/ and RCT DUPLICATE Investigator review of codes
Anxiety		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/35043165/
Depression		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/35043165/
Obstructive sleep apnoea		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
COPD		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Asthma		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002
Osteoporosis	Defined by diagnosis codes/ antiosteoporosis medications	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM, NDC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Osteoarthrosis		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Syncope		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34705014/
Falls		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://pubmed.ncbi.nlm.nih.gov/34705014/
NASH/NAFLD		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Alcohol abuse or dependence[6]	Higher risk of hypoglycaemia among SUs users with alcoholism/cirrhosis	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002
Disorders of gallbladder. Biliary tract and pancreas		Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Vertebral and non-vertebral fractures	Not used in PS/Fine stratification model	Binary	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Combined morbidity index [7, 8]		Continuous	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		
Frailty index [9]		Continuous	[-180, 0]	any	ICD9-CM, ICD10-CM	any	Exposure: 2nd Gen SUs Reference: DPP-4is		
Antihypertensive medications	Defined ACEs/ARBs/Beta blockers/Calcium channel blockers/Thiazides/Diuretics	Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Anti-arrhythmics		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Digoxin		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Anticoagulants		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Anticonvulsants		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Anti-lipidemia medications	Defined by Statins/other lipid-lowering medications	Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
NSAIDS (without aspirin)		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Aspirin	Not used in PS/Fine stratification model	Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Opioids		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Antidepressants		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Anxiolytics/ Hypnotics		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Antipsychotics	Both typical and atypical	Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Benzodiazepine		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
PPIs		Binary	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Metformin prior use		Binary	[-180, -1]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Metformin concurrent use	Either started new prescription or days' supply overlap	Binary	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Alpha glucosidase prior use		Binary	[-180, -1]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference:		RCT DUPLICATE Investigator review of codes

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
							DPP-4is		
Alpha glucosidase concurrent use	Either started new prescription or days' supply overlap	Binary	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Amylin prior use		Binary	[-180, -1]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Amylin concurrent use	Either started new prescription or days' supply overlap	Binary	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
GLP1 agonists prior use		Binary	[-180, -1]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
GLP1 agonists concurrent use	Either started new prescription or days' supply overlap	Binary	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Meglitinides prior use		Binary	[-180, -1]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Meglitinides concurrent use	Either started new prescription or days' supply overlap	Binary	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
SGLT2 inhibitors prior use		Binary	[-180, -1]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
SGLT2 inhibitors concurrent use	Either started new	Binary	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs		RCT DUPLICATE Investigator review of codes

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
	prescription or days' supply overlap						Reference: DPP-4is		
1st gen-SUs prior use		Binary	[-180, -1]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
1st gen-SUs concurrent use	Either started new prescription or days' supply overlap	Binary	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Insulin prior use		Binary	[-180, -1]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Insulin concurrent use	Either started new prescription or days' supply overlap	Binary	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Thiazolidinediones prior use		Binary	[-180, -1]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Thiazolidinediones concurrent use	Either started new prescription or days' supply overlap	Binary	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
Number of concurrent anti-diabetic medications	Count the number of concurrent anti-diabetic medication class	Continuous	[0, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator review of codes
EHR continuity		Categorical	[-180, 0]	IP, OP	ICD9-CM, ICD10-CM	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		*Lin et al* PMID: 32099479

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Number of prescriptions		Continuous	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		
Number of generics		Continuous	[-180, 0]	n/a	NDC	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		
Number of AV visits		Continuous	[-180, 0]	OP		n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		
Number of OA visit		Continuous	[-180, 0]	OP		n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		
Number of ED visit		Continuous	[-180, 0]	ED		n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		
Number of hospital visit		Continuous	[-180, 0]	IP		n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		
Bone mineral density test		Binary	[-180, 0]	any	CPT/HCPC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator choice of codes
PSA or Prostate exam	Only among males	Binary	[-180, 0]	any	CPT/HCPC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002 and RCT DUPLICATE Investigator choice of codes
Mammogram	Only among females	Binary	[-180, 0]	any	CPT/HCPC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Pap smear	Only among females	Binary	[-180, 0]	any	CPT/HCPC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		https://dev.sentinelsystem.org/projects/AP/repos/sentinel-analytic-packages/browse?at=refs%2Fheads%2Fcder_sir_wp002
Metabolic blood chemistry test		Binary	[-180, 0]	any	CPT/HCPC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator choice of codes
Flexible Sigmoidoscopy or colonoscopy or CT virtual colonoscopy		Binary	[-180, 0]	any	CPT/HCPC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator choice of codes
Flu vaccine		Binary	[-180, 0]	any	CPT/HCPC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator choice of codes
Pneumococcal vaccine		Binary	[-180, 0]	any	CPT/HCPC	any	Exposure: 2nd Gen SUs Reference: DPP-4is		RCT DUPLICATE Investigator choice of codes
Low-income subsidy indicator	"LIS_MONTHS" where low income subsidy = 1 if LIS_MONTHS value > 0 OR "DUAL_STUS_CD" where low income subsidy = 1 if 'DUAL_STUS_CD" = 01 or 02 or 03 or 04 or 05 or 06 or 08 from MBSF file	Binary	[-180, 0]	any	n/a	n/a	Exposure: 2nd Gen SUs Reference: DPP-4is		https://resdac.org/cms-data/variables/part-d-low-income-subsidy-lis-months and https://resdac.org/cms-data/variables/medicare-medicaid-dual-eligibility-code-january
Medicare vs Medicaid vs. dual enrollee	Variable available in	Categorical	[-180, 0]	any	n/a	n/a	Exposure: 2nd Gen SUs Reference:		Variable name in RIF: DUAL_ELGBL_CD_01

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
	original RIF from ResDAC						DPP-4is		

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix 5 for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.5. Data analysis

7.5.1 Context and rationale for analysis plan

For this proof-of-concept comparison of signal identification analysis options, we will vary the outcome tree to explore the impact of adding EHR based information from NLP and labs. By varying the outcome tree, we will be able to describe the ability of these enhanced trees to identify an expected adverse signal (hypoglycemia) and the pattern of statistical alerts that arise in the selected example. We will also vary the care setting from which outcomes will be identified – either requiring outcomes to occur in an inpatient or emergency department setting OR allowing outcomes to occur in any care setting. This will allow us to compare the impact of selecting more stringent criteria for defining incident outcomes (IP/ED) against more inclusive criteria (any care setting). The former may capture more serious outcomes and be more likely to be new occurrences; however, may miss less severe but real adverse outcomes. The latter may identify more outcomes of lower severity, however more of these may reflect recording of non-incident health events that were discussed during the healthcare encounter. Details of how the data can be analysed with the TBSS R package using binomial, Poisson and Gaussian test statistics are described in Russo et al[17], including tutorials and a comparison of results when implementing the same analyses with the same data using the open-source R TBSS package versus the original, proprietary TreeScan™ software. Traditionally, when using TreeScan™, the investigators choose a level of the hierarchy at which to define incident outcomes. Then, after applying a washout of *n* days to identify outcomes that are incident in nodes at that level, the incident event defining diagnoses are traced to the leaf level, and these leaf level diagnoses are included in the count file as input for analyses. This approach of identifying incident outcomes does not work when the hierarchical tree is multi-axial, meaning that there are multiple pathways for lower-level nodes to be aggregated up a tree. In order to accommodate the multi-axial pathway of ontologies that are commonly used in bioinformatics, TBSS methods have been expanded to allow node-specific washouts for defining incident outcomes across the hierarchical outcome tree. In the same manner as the original TBSS methods, the global null hypothesis is that there is no increase in risk in any nodes across the tree. This is tested against an alternative hypothesis that there is at least one node for which there is an elevated risk with exposure. The test statistic for the global null is the most extreme test statistic across all nodes in the hierarchical outcome tree. Under the null hypothesis, the treatment labels are exchangeable, however the distribution of test statistics is unknown. Therefore, in the original TreeScan™ the distribution of p-values under the null are obtained by permuting the labels of treatment and comparator in the leaves, aggregating counts in the tree up to the level at which incidence is defined, and then computing test statistics. With node-specific washouts, the permutation algorithm is modified such that exposure vs comparator labels are permuted based on patient vectors of incident outcomes rather than at the leaf level. This means that all incident outcomes for a patient are permuted together as a cluster. After permuting the exposure labels, the counts of events for each node are recomputed, and the test statistics are calculated. The distribution of the test statistics under the null hypothesis is generated by permuting the data multiple times (e.g. 999 times), and the global p-

value is calculated by the rank of the test statistic for the observed data compared to the 999 permutations under the null (p-value = rank/(# permutations +1). Details of inputs and outputs necessary for TBSS analyses with node specific washout and permutation are provided in appendix 4. The complete set of algorithms and code lists used to extract the analytic cohort are provided in the Query Report Package (QRP) inputs provided in appendix 5.

Table 10. Analysis specification

	Outcome Tree	***Care Setting for Incident Outcomes***	***TBSS model***	***Confounding Adjustment***
Analysis 1	Claims	IP/ED	Poisson	PS fine stratification
Analysis 2	Claims+NLP	IP/ED	Poisson	PS fine stratification
Analysis 3	Claims+NLP+labs (mapped binary only)	IP/ED	Poisson	PS fine stratification
Analysis 4	Claims	Any	Poisson	PS fine stratification
Analysis 5	Claims+NLP	Any	Poisson	PS fine stratification
Analysis 6	Claims+NLP+labs (mapped binary only)	Any	Poisson	PS fine stratification
Analysis 7	Claims	IP/ED	Gaussian	PS fine stratification
Analysis 8	Claims+NLP	IP/ED	Gaussian	PS fine stratification
Analysis 9	Claims+NLP+labs (mapped binary and continuous)	IP/ED	Gaussian	PS fine stratification
Analysis 10	Claims	Any	Gaussian	PS fine stratification
Analysis 11	Claims+NLP	Any	Gaussian	PS fine stratification
Analysis 12	Claims+NLP+labs (mapped binary and continuous)	Any	Gaussian	PS fine stratification

Hypothesis:	N/A, this is a methods comparison study that evaluates TBSS, a hypothesis free signal identification method
Exposure contrast:	2nd generation SUs versus DPP-4is
Outcome:	Outcomes included in the 3 variants of an outcome tree: 1) MedDRA concepts from claims/billing codes, 2) MedDRA concepts from claims/billing codes + NLP mapped concepts from unstructured notes, 3) MedDRA concepts from claims/billing codes + NLP mapped concepts from unstructured notes + selected laboratory data (mapped and native

	continuous values will be evaluated simultaneously for the Gaussian test statistic, only the mapped binary values will be evaluated for the Poisson test statistic)
Analytic software:	FDA Sentinel's Cohort Identification and Descriptive Analysis tool (CIDA), R TBSS, Python
Model(s): ***(provide details or code)***	Unconditional binomial test statistic versus unconditional Poisson test statistic vs gaussian test statistic
Confounding adjustment method	***Name method and provide relevant details, e.g. bivariate, multivariable, propensity score matching (specify matching algorithm ratio and caliper), propensity score weighting (specify weight formula, trimming, truncation), propensity score stratification (specify strata definition), other.***
	Nearest-neighbor propensity score matching, 1:1 match ratio, caliper of 0.2 standard deviations of the PS on the logit scale Fine stratification weights with 1% asymmetric trimming, 50 strata defined by percentiles of the PS in the exposure group. Weight formula for ATT estimates: Exposed group weight = 1; comparator group weight = $\frac{N_{exposed\ in\ strata\ i}/N_{total\ exposed}}{N_{unexposed\ in\ strata\ i}/N_{total\ unexposed}}$
Missing data methods	***Name method and provide relevant details, e.g. missing indicators, complete case, last value carried forward, multiple imputation (specify model/variables), other.***
	For binary baseline characteristics or outcomes defined by either occurrence of a diagnosis code, presence of a mention in a note, or mapping abnormal lab values to MedDRA, we will assume that if there is no evidence of an event, that the event did not occur. For continuous laboratory measures, we will do a complete case analysis for the purposes of this project. In the future, multiple imputation or other approaches may be applied in conjunction with TBSS methods.
Subgroup Analyses	***List all subgroups***
	n/a

7.6. Data sources

7.6.1 Context and rationale for data sources

Reason for selection: Mass General Brigham (MGB) and Vanderbilt University Medical Center (VUMC) are two members of the Sentinel Innovation Center Development network. MGB has linked Medicare claims data (2013-2020) and Medicaid claims (2000-2018) deterministically to EHRs for patients receiving care at the MGB healthcare system (linkage success rate 99.2%). The claims data include inpatient, outpatient, skilled nursing and rehabilitation claims, and drug dispensing claims, linked with clinical assessment files, including MDS35, OASIS36, and IRF-PAI. MGB is the largest healthcare provider system in Massachusetts, comprising >40 healthcare facilities across the full care continuum. The EHR databases contain information on patient demographics, medical diagnosis and procedures, medications, vital signs, smoking status, body mass index (BMI), immunizations, laboratory data,

various clinical notes and reports. MGB investigators have direct access to full-text operative notes for these patients, including ambulatory notes, discharge summaries, and various specialty reports (e.g., cardiology, pulmonary). The claims databases contain information on demographics, enrollment start and end dates, medical diagnosis, dispensed medications, and performed procedures. The linked claims-EHR dataset includes a total of 1.2 million patients; of whom nearly 40% (~484,000) have at least one healthcare encounter in the most recent year. Finally, these patients have been deterministically linked to state vital statistic files from Massachusetts, Connecticut, Vermont, and New Hampshire, to track long term outcomes of all-cause and cause-specific mortality.

Vanderbilt University Medical Center (VUMC) has EHR data and clinical notes for over 3 million patients, with 55,000 admissions, 70,000 emergency department visits, and 40,000 surgeries annually. VUMC has linked Tennessee Medicaid (TennCare) claims data (2000-2021) deterministically linked to EHR data for patients receiving care in the Vanderbilt Health System. Vanderbilt Health is a growing health system with 1,741 licensed hospital beds across seven hospitals and more than 180 clinics in Tennessee and neighboring states. It is anchored by Vanderbilt University Medical Center, one of the largest academic medical centers in Southeast. EHR data was extracted from the Research Derivative (RD), a database of clinical and related data derived from VUMC's clinical systems and restructured for research. Data is repurposed from VUMC's enterprise data warehouse, which includes data from StarPanel, VPIMS, and ORMIS (Operating Room Management Information System), EPIC, Medipac, and HEO among others. The medical record number and other person identifiers are preserved within the database. Data types include reimbursement codes, clinical notes and documentation, nursing records, medication data, laboratory data, encounter and visit data, among others. Output includes structured data points, such as ICD 9 or 10 codes and encounter dates, semi-structured data such as laboratory tests and results, or unstructured data such as physician progress reports. The TennCare claims databases contain information on demographics, enrollment start and end dates, medical diagnosis, dispensed medications, and performed procedures. The linked claims-EHR dataset includes approximately 725,000 patients with data in both the EHR and TennCare databases. TennCare patients with a non-missing date of death were probabilistically linked to Tennessee Vital death data using name, date of birth, sex, address, SSN, and Medicaid date of death; death and cause of death were also sourced from EHR.

Strengths of data source(s): Both MGB and VUMC have data assets that include Medicare/Medicaid claims data linked to electronic health records and state death data. This will enable assessment of the performance of TBSS with hierarchical trees that include outcomes derived from multi-modal data streams. The claims linkage will enable us to ensure a baseline window of enrollment/observation over which to assess baseline characteristics as well as determine a window of observable time for claims-based outcomes. The EHR data will provide the opportunity to capture signs, symptoms, and changes in vital signs or labs that may not reach the level of severity that would become a billable diagnosis. Exposure to newly initiated and dispensed medications of interest will be captured through claims, which may be a better reflection of true exposure than prescribed medications from the EHR due to primary non-adherence.[18] Eligibility criteria will be captured through a mix of claims and EHR data. The linkage to state death data will provide more complete capture of censoring related to death and supplement the claims and EHR recorded deaths.

Limitations of data source(s): Both MGB and VUMC data assets are secondary data collected as part of routine care, not for research purposes. Due to fragmentation of the healthcare system, EHR data from a specific healthcare system may only capture part of a patients' healthcare experience. In contrast, insurance-based claims databases will capture the entirety of the billable patient experience. Relatedly, while enrollment windows for the claims portion of the MGB and VUMC databases can be used to determine 'observable' windows of patient healthcare experience, this is more challenging to determine with EHR data. There may also be (high) missingness on variables of interest, including eligibility criteria, confounders, and potential outcomes.

Data source provenance/curation: Both the MGB and VUMC databases will be transformed from the raw form in which they are received to the Sentinel Common Data Model (CDM). These data transformations will reflect the prior work from the Sentinel Operations to update the claims-based Sentinel CDM as well as the Sentinel Innovation Center's projects focused on incorporating EHR data, including: DI2 Representation of unstructured data across Common Data Models, DI3 Identification and mitigation of structured EHR source data mapping issues, and DI7 creating a Sentinel linked EHR-claims development network.[19]

Table 12. Metadata about data sources and software

	Data 1	Data 2
Data Source(s):	Mass General Brigham Medicare linked to EHR	Vanderbilt University Medical Center Medicaid linked to EHR
Study Period:	10/1/2015 – 12/31/2020	10/1/2015 – 12/31/2020
Eligible Cohort Entry Period:	10/1/2016 – 12/31/2020	10/1/2016 – 12/31/2020
Data Version (or date of last update):	Medicare DUA 56141, 1/1/2007 to 12/31/2020, most recent data received 2022-12-31 Research Patient Data Registry[1] (RPDR) 2017P002659, 2000 – 2021, most recent data received 2022-05-10 Death data: 1991-2021	Tennessee Medicaid: 2000 – 2021. Most recent data received 2023-06-28. VUMC Research Derivative (RD): 2010-2023. Most recent data received: 2024-04-26. Death data: Tennessee Department of Health 2000-2021. Most recent data received 2023-05-03. VUMC RD 2010-2023. Most recent data received: 2024-04-26.
Data sampling/extraction criteria:	Patients aged 65 and above were identified in the RPDR warehouse. These patients were deterministically linked to Medicare data. Medicare data for linkable patients were extracted. Patients with any record in RPDR-Medicare linked data were identified and probabilistically linked to MA state death data.	EHR and Medicaid patients were probabilistically linked using EHR supplied values. Medicaid data were extracted for successfully linked patients. Medicaid data patients with a non-missing date of death were probabilistically linked to Tennessee Vital death data.
Type(s) of data:	*Medicare:* enrollment files, inpatient claims, outpatient claims, carrier claims, and other claims files (home health agency, hospice, skilled nursing facility, and durable medical equipment), and Part D (prescription drug coverage) files. *RPDR:* patient demographics, medical problems, medications, vital signs, smoking status, BMI, immunizations, laboratory data, radiology reports, cardiology and other specialty reports, ambulatory clinic notes, progress reports, admission notes,	*TN Medicaid (TennCare):* enrollment files, inpatient claims, outpatient claims, physician claims, nursing home claims, pharmacy (outpatient) claims, durable medical equipment claims, dental claims and revenue data. *VUMC RD:* patient demographics, medical problems, medications, vital signs, BMI, immunizations, laboratory data, radiology reports, cardiology and other specialty reports, ambulatory

	and discharge summaries *State death data:* death and cause of death	clinic notes, progress reports, admission notes, and discharge summaries *Death data:* death and cause of death
Data linkage:	Medicare and RPDR data were deterministically linked using Health Insurance Claims (HIC) number or Medicare Beneficiary Identifiers (MBIs), the linkage success rate was 99.9%. Medicare and RPDR data were probabilistically linked to death data using name, sex, and date of birth, 76.2% of patients with date of death not missing in RPDR or Medicare were linked to MA death. Additional deaths may be linked in pending linkages with death data from surrounding states.	EHR and Medicaid data were probabilistically linked using EHR supplied data values for last name, first name, middle initial, birthdate, sex, race, ssn, address, zip code, and date of death. EHR provided linking values for 5,156,356 patients. There were 1,100,867 successful links between the EHR and Medicaid data. Medicaid patients with a non-missing date of death were probabilistically linked to Tennessee Vital death data using name, date of birth, sex, address, SSN, and Medicaid date of death.
Conversion to CDM*:	Data are converted to the Sentinel CDM v8.1.2	Data are converted to the Sentinel CDM version: 8.1.0
Software for data management:	SAS 9.4	SAS 9.4

*CDM = Common Data Model

[1]RPDR is a centralized clinical data registry that contains data from MGB EHRs

7.7. Data management

MGB data management procedures

Tracking data: The Data Use Agreement (DUA) Custodian (Principal Investigator) and the Division Operations Manager (Winta Tekle) are responsible for receiving and creating a record of new data associated with the given request. The Operations Manager is responsible for all indexing and archiving of documents and electronic media related to data that the Division receives and will log data location, data contents, and associated DUA and Institutional Review Board (IRB) numbers in the Division's centralized tracking system.

Handling & Storage: Senior Programmers will perform QA/QC via SAS/R/Aetion software. In-scope data will be stored utilizing Mass General Brigham approved information systems – i.e., Division servers and Mass General Brigham network storage.

Archiving: Data retention and any removal will be executed by the Division Operations Manager (Winta Tekle) and Data Manager (Todd MacGarvey). The Division follows Mass General Brigham enterprise record retention policies and follow the terms of the agreement through which the in-scope data was received.

Information security: The Division follows Mass General Brigham's Enterprise Information Security Program (EISP). The EISP helps by providing assurance that Mass General Brigham information and information systems are protected from unauthorized access, use, disclosure, duplication, modification, or destruction in order to maintain their confidentiality, integrity, and availability. To that end, the EISP policies, standards and procedures create an information security framework that is aligned with the recommendations of the International Organization for Standardization's (ISO) publication 27001 and the National Institute of Standards and Technology's (NIST) publication 800-53 Family of Controls and MGB regulatory and legal requirements (as a HIPAA covered entity).

Mass General Brigham workforce members are required to complete new workforce privacy and information security training upon hire and annually thereafter. Refresh training is provided as appropriate. Mass General Brigham workforce members are required to review and sign a Confidentiality Agreement upon hire and annually thereafter. Additionally, we follow Enterprise Information Security and Privacy policies, including Managing Workforce Members Information Security Responsibilities Policy, which was developed to help assure that Mass General Brigham workforce members are competent and eligible to support the information security responsibilities associated with their role. Mass General Brigham workforce, including Division staff, are required to comply with Enterprise Information Security and Privacy Policies.

Facilities: The research team uses a highly secure, state-of-the-art, computing facility housed at Mass General Brigham's Corporate Data Centers in Needham and Marlborough, Massachusetts as well as Amazon Web Services (AWS). We maintain redundant storage for maximal data integrity and high-speed data access.

The Division uses Mass General Brigham corporate provisioned servers to analyze and store data in connection with this project. There are strict access controls enforced by technical means to ensure that only study staff who have been approved to conduct data analyses and contracted data engineers are able to access data (e.g., Mass General Brigham Authentication (Active Directory) utilized; auditing, logging, and monitoring enabled; firewalls enabled; etc.). All Division servers used in this study are accessible to only authorized staff only through the MGB network and utilizing VPN as appropriate (i.e., users must be on the network and logged into VPN to access). Mass General Brigham has a network information security monitoring team and in-scope servers are enrolled in enterprise security tools (e.g., vulnerability scanning service, antivirus, etc.).

The Division also uses AWS. In-scope AWS services / resources have undergone review and assessment by Mass General Brigham's Information Security Risk Assessment Team/ InfoSec Risk Management Team as required and in alignment with NIST 800-30: Guide for Conducting Risk Assessments. In-scope AWS servers are for our exclusive use (reserved instances), are covered under our organization's Business Associate Agreement with AWS. In-scope servers are housed in anonymous facilities that are not branded as AWS facilities. Physical access is strictly controlled both at the perimeter and at building ingress points by professional security staff utilizing video surveillance, intrusion detection systems, and other electronic means.

Security attributes and controls for the Division's use of in-scope AWS in connection with this project include: encryption at rest and in transit (industry standard AES-256 encryption, SSL/TSL), anti-malware, latest OS updates and patches as well as anti-virus; multi-factor authentication (MFA), Identity and Access Management, including SSO; auditing and logging; principle of least privilege, including for traffic and ports (e.g., deny all default configurations), use of VPC enabled Step functions, etc.; separation between environments; minimum necessary (out of the box AWS roles are not used as

they can be overly permissive); all data would reside in the US domestic regions – use only US east HIPAA compliant region; performance monitoring enabled and reviewed as per internal processes; passwords compliant with enterprise password policy and, as appropriate, more protective password requirements; etc. Our Division's AWS use is also subject to ongoing evaluation and monitoring as well – e.g., vulnerability scanning and management, access review, etc.

A data file list will be maintained in Amazon S3 and tracked through Cloudtrail and Cloudwatch. Reporting of data file availability for any file stored in an S3 bucket will be controlled by MGB.

Amazon EC2, Amazon EBS, and Amazon VPC are integrated with AWS CloudTrail, a service that provides a record of actions taken by a user, role, or an AWS service in Amazon EC2, Amazon EBS, and Amazon VPC. CloudTrail captures all API calls for Amazon EC2, Amazon EBS, and Amazon VPC as events, including calls from the console and from code calls to the APIs. MGB IT/security staff have full access to CloudTrail to ensure that CloudTrail monitoring is available at all times. Amazon CloudWatch Events delivers a near-real-time stream of system events that describe changes in AWS resources. Amazon CloudTrail can log, continuously monitor, and retain account activity related to actions across the MGB AWS infrastructure. CloudTrail provides event history of MGB AWS account activity. This event history simplifies security analysis, resource change tracking, and troubleshooting and complies with requirements to ensure the data file inventory is controlled by MGB.

Backups: Backups are created using industry recognized Cryptographic mechanisms (e.g., 256-bit AES encryption) as appropriate and required. Data are backed up within the MGB network from the MGB Needham data center to the MGB Marlborough data center via replication/mirroring and volume-level snapshots. AWS Backup supports both instance-level backups as Amazon Machine Images (AMIs) and volume-level backups as separate snapshots based on the resource tags. AMIs allow MGB IT and security staff to create backups of all S3 and EC2 instances. This backup approach allows elimination of EC2 instances when access to CMS data files is no longer needed. Use of temporary instances will allow highest security control to be applied regarding for whom and when CMS data files are accessible.

In addition to the safeguards listed above, the Division follows enterprise privacy and information security policies and maintains internally procedures in connection with this project to safeguard the data in connection with this project as appropriate and required including: Physical Security and Environmental Controls for Electronic Information Policy; IT Access Control Standards for Users policy; Safeguarding Fax Copiers Printers Telephone Use and Pagers; Physical Removal and Transport of Protected Health Information and Personal Information policy, etc.

VUMC data management

The Vanderbilt University Medical Center serves as a collaborating data partner to the FDA Sentinel System, contributing Tennessee Medicaid (TennCare) data through the VUMC Department of Health Policy. The Department of Health Policy Research Data Core Team is directed by Dr. Grijalva (Co-investigator) and managed by Dr. Chad You (Senior Database Administrator). The Research Data Core Team currently provides the infrastructure, administrative support, and expertise in the use of large real-world databases that includes Tennessee Medicaid data augmented with Medicare and Hospital Discharge Data System records and birth, death and fetal death certificates. Furthermore, the Research Data Core Team has leveraged this data infrastructure to create the TN Medicaid Mother-Child Linked Cohort (MCLC) data platform. The TN Medicaid MCLC is a ready-to-use and continuously updated data platform consisting of linked data for mothers and babies in TN Medicaid that includes enrollment data, healthcare claims and pharmacy data, birth, fetal death and death certificate data, and hospital-based data from an all-payer statewide registry. This resource is updated annually with State

Vital Records and encompasses >600,000 linked mother-child dyads between 2000 and 2021. Faculty in the Department of Health Policy have used the TN Medicaid MCLC extensively to conduct studies that inform public health and clinical practice during the last 30 years, including impactful work in the areas of pharmacoepidemiology, maternal and child health, and substance use and mental health.

The Research Data Core Team has comprehensive data security procedures encompassing data safety, data storage and access, study personnel training and assurances, and data destruction. The availability of multiple linked databases with personal identifiers requires scrupulous attention to protect confidentiality. All files, including any original data transmitted to the VUMC file share or transmitted via CD/DVD media will be uploaded onto a secured VUMC file share which is Dell EMC - Isilon Network Attached Storage (NAS) and maintained by VUMC IT. The file share is behind the Vanderbilt Protected Health Information (PHI) firewall, with the VUMC Network Infrastructure serving as the principal means of safeguarding all electronic information (same security level as VUMC EHR). The physical hardware is in the primary data center located in VUMC. The building containing these cabinets, file share, and servers are protected by electronic key access as determined by access control lists. The building containing these cabinets, file share, and servers are protected by electronic key access as determined by access control lists. The facility has two power feeds as well as a backup generator and Universal Power Supply. It also has two internet sources from different Internet Service Providers. A system for cooling and humidity control is in place. and the facility has fire emergency controls such as the emergency power off, and fire suppression system. Vanderbilt data centers are staffed 24/7 by human monitors. Access is only granted to those with a documented reason for access to the facility. The file share is fully compliant with the HIPAA Security and Privacy Rules and VUMC defined guidelines. VUMC requires reasonable and appropriate physical, administrative, and technical safeguards to protect PHI from accidental or intentional destruction, alteration or loss from events such as sabotage, corruption, theft, environmental disasters, malicious software threats, and/or inappropriate disclosure. Files and data are presented for the file share clients thru the Server Message Block (SMB) protocol and a SMB share. The Isilon cluster, i10FILE has been joined to the Active Directory Domain and authentication thru the service is required to access the share. All external computer access is username and password-protected via VPN (Firepass from F5/BIG-IP Edge). In addition, Multi-Factor Authentication (MFA) has been set up as an extra layer of security. The VPN service plus MFA protect the data and preserves data confidentiality and integrity when information is transmitted from a remote location. Points of internal access to these databases are monitored and maintained by a firewall and router. The router has access lists and commands to deny or permit traffic. Passwords must be updated yearly, computers automatically lock after fifteen minutes, and employees agree to lock computer screens when leaving their workspace. VUMC Encryption Policies state that AES-256-gcm encryption methods are supported for IPsec VPN tunnels terminated to the VUMC infrastructure. Devices storing data with classifications which require encryption will use Full-Disk Encryption (FDE). Both Windows and MacOS operating systems support FDE.

The Research Data Core Team has several other safeguards in place to reduce the risk of informational harm and PHI disclosure to ensure the protection of subjects. Investigators are expected to follow VUMC specific regulations and security policies designed to ensure the confidentiality of all sensitive information. All investigators understand that information about patients or employees is confidential and protected from unauthorized disclosure by law. Improper disclosure of information to anyone not authorized to receive it may result in criminal charges and a fine under the Privacy Act of 1974, Title 5 United States Code (U.S.C.) 552(a), 38 U.S.C. Section 5701, Confidential Nature of Claims, and 38 U.S.C. Section 7332, Confidentiality of Certain Medical Records. These procedures are likely to reduce to the maximum extent any possible breach of confidentiality.

All data files with identifiable data will be stored and processed using a physically secure computer system and using the secure file share described above. Identifiable data will be stored in a folder on the secure file share with restricted access. After linkage, analytic files will be processed in such a manner that

the ultimate analytical files do not contain individual identifiers, but rather only a unique study ID. Additionally, analytic files will be stored in a separate folder on the secure file share with access restricted only to the individuals conducting the analysis.

7.8. Quality control

We will use extensively validated QRP code produced by the Sentinel Operations Center to extract an analytic cohort for the case example for this project. The open source TBSS software has been tested against the original TreeScan™ software using the same input files for functions that are available in both. Additional validation of the TBSS R package will be conducted through checks on simulated data with known characteristics. We will use a freely available and widely used software (MetaMap) as the basis for extraction of NLP concepts from clinical notes. We will validate extracted terms as much as time/effort permits by assessing face validity of samples of the extracted data and spot-checking that extracted concepts match the intended meaning. Tables that describe baseline characteristics in crude, matched, or weighted cohorts (see appendix 7) will be reviewed to identify anomalies in the data for variables that are relevant to the analysis and adjustments will be made to operational/implementation algorithms as needed.

7.9. Study size and feasibility

A query report package was run in the MGB data to obtain anticipated sample sizes (without looking at outcomes). The QRP attrition tables are provided in Appendix 6. Baseline characteristics of the crude cohort, 1:1 matched, and fine stratified cohorts are provided in Appendix 7.

8. Limitations of the methods

- It is infeasible to adjust for all potential confounders for the thousands of outcomes evaluated in a TBSS analysis. Residual confounding can lead to spurious alerts or decreased ability to detect true associations. It is challenging to identify an optimal set of confounders to adjust for that are measurable in the data, general risk factors for multiple potential outcomes, and not strong instruments for any outcomes. Prior work has demonstrated that partial adjustment via a general propensity score that includes proxies for general comorbidity and frailty can provide broad confounding coverage.[7]
- A single common risk window to screen for potential adverse event signals may not be optimal across all potential outcomes. We plan to use a relatively short 6 month follow up window. This precludes detection of adverse events that require longer term exposure or follow up.
- The washout for all outcomes was limited to [-180, 0], consistent with the enrollment requirement, but shorter than has typically been done for prior TBSS analyses (400 days) due to limited sample size that could be obtained with longer washout. This shorter washout may result in more events being classified as “incident outcomes”. However, if the exposure and comparator cohorts are well balanced, the groups should be equally affected.
- Additionally, the definition of T2DM was more relaxed than in prior Sentinel studies, requiring only 1 diagnosis of diabetes at baseline rather than the more stringent 1 inpatient or 2 outpatient diagnoses. Requiring only 1 diagnosis, in addition to starting an antidiabetic drug was chosen to maximize the limited sample size available in each data partner.

- The hierarchy of outcomes across the MedDRA tree is not based on validated algorithms. While useful for screening activities, these may be neither sensitive nor specific, making them sub-optimal for capturing a particular outcome of interest. Relatedly, the MedDRA hierarchy aggregates outcomes based on similarity in one dimension or another, however these groupings may or may not be relevant in different clinical contexts.
- While TBSS methods will account for multiple comparisons when screening across thousands of outcomes in a tree to fix the rate of false positives for a study, this will decrease power compared to analysing a single (or handful) of pre-specified hypotheses. Importantly, while p-values are a useful metric on which to evaluate and rank statistical alerts, they are not the only dimension to evaluate when prioritizing potential signals to follow up on. For example, magnitude of absolute differences may indicate public health significance.
- Inclusion of NLP extracted signs, symptoms and events that do not lead to billable diagnoses may increase the probability of observing statistical alerts for non-specific syndromes or patterns of outcomes that must be carefully interpreted through both a clinical and methodological lense.
- Additionally, accurate NLP on clinical text is an ongoing area of research. While the methods and tools we will employ are well-validated and widely used, it is likely that some negated concepts will be incorrectly recognized as present, some affirmatively-mentioned concepts will be missed, and that certain words, phrasing, and clinical concepts will be missed or misclassified due to synonymy, the complexity of human language, and/or the unique semantic and structural challenges posed by clinical notes. Overall, we expect NLP-extracted data to be accurate and not to adversely affect the study, but information extracted from clinical text will not be 100% accurate.
- We are evaluating the methods in a single proof-of-concept example where there is an expected signal for hypoglycemia, however the “truth” is not known across all outcomes being screened.

9. Protection of human subjects

FDA Sentinel projects are considered public health activities rather than research, therefore this project does not require human subjects review.

10. Reporting of adverse events

This is a non-interventional methods evaluation project using routinely collected data. We will screen for potential adverse effects of drugs and all results will be reported to the FDA members of the workgroup as well as in public reports.

11. References

1. Kulldorff M, Dashevsky I, Avery TR, et al. Drug safety data mining with a tree-based scan statistic. *Pharmacoepidemiology and drug safety*. May 2013;22(5):517-23. doi:10.1002/pds.3423
2. Kulldorff M, Fang Z, Walsh SJ. A tree-based scan statistic for database disease surveillance. *Biometrics*. Jun 2003;59(2):323-31.
3. Maro JK, A; Baker, Ml Daschevsky, I; Nguyen, M; Scott, J; Kulldorff, M. TreeScan Power (PRISM). US Food and Drug Administration (FDA). Updated November 13, 2014. Accessed June 26, 2017, https://www.sentinelinitiative.org/sentinel/methods/336
4. Wang SV, Maro JC, Baro E, et al. Data mining for adverse drug events with a propensity score matched tree-based scan statistic. *Epidemiology*. Aug 1 2018;doi:10.1097/EDE.0000000000000907

5. Brown JS, Petronis KR, Bate A, et al. Drug Adverse Event Detection in Health Plan Data Using the Gamma Poisson Shrinker and Comparison to the Tree-based Scan Statistic. *Pharmaceutics*. Mar 14 2013;5(1):179-200. doi:10.3390/pharmaceutics5010179
6. Yih WM, JC; Dashevsky, I; Anderson, S; Baker, MA; Mba-Jonas, A; Russek-Cohen, E; Shoaibi, A; Yan, L; Kulldorff, M. Evaluation of HPV9 (Gardasil9) Vaccine Safety Surveillance Using the TreeScan Data Mining Method Surveillance Protocol. US Food and Drug Administration (FDA). Updated June 30, 2016. Accessed June 26, 2017, https://www.sentinelinitiative.org/vaccines-blood-biologics/assessments/evaluation-hpv9-gardasil9-vaccine-safety-surveillance-using
7. Wang SV, Maro JC, Gagne JJ, et al. A General Propensity Score for Signal Identification Using Tree-Based Scan Statistics. *American journal of epidemiology*. Jul 1 2021;190(7):1424-1433. doi:10.1093/aje/kwab034
8. Gagne JJ HX, Hennessy S, Leonard CE, Chrischilles EA, Carnahan RM, Wang SV, Fuller C, Iyer A, Katcoff H, Woodworth TS, Archdeacon P, Meyer TE, Schneeweiss S, Toh S. . Successful comparison of US Food and Drug Administration Sentinel analysis tools to traditional pharmacoepidemiologic approaches. . *Clinical Pharmacology and Therapeutics* ((in press))
9. Go AS, D; Cheetham, TC; Toh, D; Reichman, M; Graham, D; Southworth, MR; Rongmei, Z; Houstoun, M; Wu, Y; Mott, K; Gagne, JJ. Mini-Sentinel Product Assessment: A Protocol for Assessment of Dabigatran Version 3. Food and Drug Administration, Sentinel Program. Updated March 27, 2015. http://www.mini-sentinel.org/work_products/Assessments/Mini-Sentinel_Protocol-for-Assessment-of-Dabigatran.pdf
10. Carnahan RG, JJ; Nelson, J; Fireman, B; Wang, SV; Shoaibi, A; Reichman, M; Zhang, Rongmei; Levenson, M; Graham, D; Tiwari, R; Southworth, MR; Archdeacon, P; Chakravarty, A; Goulding, M; Izem, R; Brown, J; Fuller, C; Rogers, C; Toh, Darren; Chrischilles, E. Mini-Sentinel Prospective Routine Observational Monitoring Program Tools (PROMPT): Rivaroxaban Surveillance Version 3. Food and Drug Administration, Sentinel Program. Updated October 29, 2015. http://www.mini-sentinel.org/work_products/Assessments/Mini-Sentinel_PROMPT_Rivaroxaban-Surveillance-Plan.pdf
11. Toh S, Reichman ME, Houstoun M, et al. Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. *Archives of internal medicine*. Nov 12 2012;172(20):1582-9. doi:10.1001/2013.jamainternmed.34
12. Leonard CR, ME; Toh, D; Kulldorff, M; Nelson, JC; Gagne, JJ; Ouellet-Hellstrom, RP; Moeny, DG; Mott, KA; By, K; Wang, SV; Hennessy, S. . Mini-Sentinel Prospective Surveillance Plan: Prospective Routine Observational Monitoring of Mirabegron. . June 2014
13. Toh S, Avorn J, D'Agostino RB, Sr., et al. Re-using Mini-Sentinel data following rapid assessments of potential safety signals via modular analytic programs. *Pharmacoepidemiology and drug safety*. Oct 2013;22(10):1036-45. doi:10.1002/pds.3478
14. Workgroup. P. Taxonomy for monitoring methods within a medical product safety surveillance system: year two report of the Mini-Sentinel Taxonomy Project Workgroup. . http://www.mini-sentinel.org/work_products/Statistical_Methods/Mini-Sentinel_Methods_Taxonomy-Year-2-Report.pdf
15. Quality AfHRa. Clinical Classifications Software (CCS) 2015. Healthcare Cost and Utilization Project (HCUP) Accessed 2021/01/28. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf
16. MedDRA Hierarchy. Medical Dictionary for Regulatory Activities. Accessed 2017-09-09, https://www.meddra.org/how-to-use/basics/hierarchy
17. Russo M, Wang SV. An open-source implementation of tree-based scan statistics. *Pharmacoepidemiology and drug safety*. 2024/03/01 2024;33(3):e5765. doi:https://doi.org/10.1002/pds.5765
18. Adams AJ, Stolpe SF. Defining and Measuring Primary Medication Nonadherence: Development of a Quality Measure. *Journal of managed care & specialty pharmacy*. May 2016;22(5):516-23. doi:10.18553/jmcp.2016.22.5.516
19. Sentinel Innovation Day. Accessed 1/30/2024, 2024. https://www.sentinelinitiative.org/sites/default/files/documents/2022_Sentinel_Innovation_Day_Presentation_0.pdf

12. Appendices

Appendix 1. Process and file layout for mapping ICD-10-CM codes to MedDRA
Appendix 2. Process and file layout for mapping NLP extracted terms to MedDRA
Appendix 3. Process and file layout for mapping laboratory values to MedDRA
Appendix 4. Process and file layout for outcome tree and count files for TBSS analyses using node specific washout and permutation
Appendix 5. Query Report Package (QRP) with detailed code algorithms for exposure, eligibility criteria, outcome, and covariates
Appendix 6 QRP attrition tables
Appendix 7 Baseline characteristics of the crude cohort, 1:1 matched, and fine stratified cohorts