

Introduction to the Data2Evidence Cohort Query Tool

Developing cohorts using Electronic Health Record Data
on the AI Ready Mount Sinai (AIR-MS) Platform

Ashwin Sawant, MD

Andrew Deonarine, MD

May 26, 2026



Hasso Plattner Institute for Digital Health at Mount Sinai

Cohort Query Tools

- Explore data
- Get counts for research proposals
- No coding needed

Cohort query tools for Mount Sinai data

	Data2Evidence	LEAF	ATLAS	SlicerDicer
Development	Data4Life	Nick Dobbins, Univ of Washington, ISMMS	OHDSI	Epic Inc.
License	Free and open-source software			Commercial
Tradeoff	Newer, actively developed High functionality Easy to use interface	Not actively developed Limited functionality	Complexity Higher functionality	For existing Epic Hyperspace users only Must not inadvertently access PHI without authorization
Data available	De-identified	De-identified	De-identified	All Epic data
Capabilities	Pre-defined stats Can customize visualization Pre-defined scenario wizards	Pre-defined stats	Customized stats	Customized stats and visualization

OHDSI, Observational Health Data Sciences and Informatics; PHI, Protected Health Information.

Cohort query tools for multi-site data

- Epic Cosmos
 - 304 million patients,
 - 2,070 hospitals, 47.1k clinics
- TriNetX
 - 117 million patients
 - Polished, commercially developed interface
 - Multi-institution data possible with institutional agreement/subscription
 - Mount Sinai currently doesn't have an active agreement
- Valuable for rare diseases research
- Useful as a validation cohort
- Regularly updated

<https://cosmos.epic.com/>

<https://trinetx.com/data-sets-analytics/>

Accessed May 26, 2026

Data modalities in AIR.MS

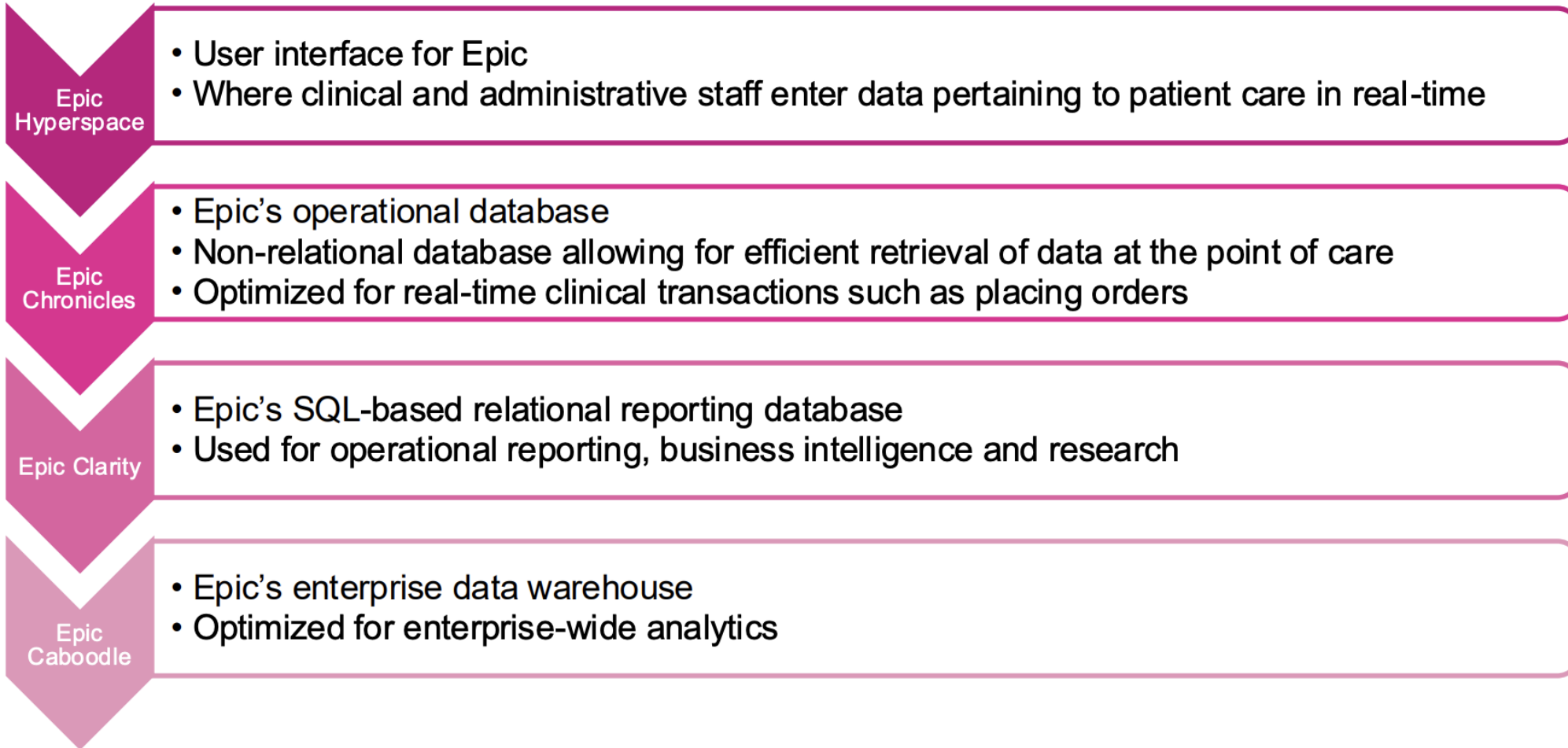
Currently Available Modalities:

- Mount Sinai Data Warehouse (MSDW), holds both protected health information (PHI) and DeID (de-identified) Observational Medical Outcomes Partnership (OMOP)-mapped electronic health records (EHR)
- Intensive care unit (ICU) datamart
- Pathology Metadata
- Radiology Metadata
- BioMe/Sinai Million
- Electrocardiogram (EKG)
- Echocardiography
- Gastroenterology (GI) Research Database

Work in progress: Electroencephalogram (EEG), Endoscopy, Colonoscopy, cardiac imaging reports, physiological waveform data (Bedmaster -> GE AirStrip in future?)

Clean up, add Minerva raw datasets

Epic is the Primary Source of Electronic Health Record Data at Mount Sinai

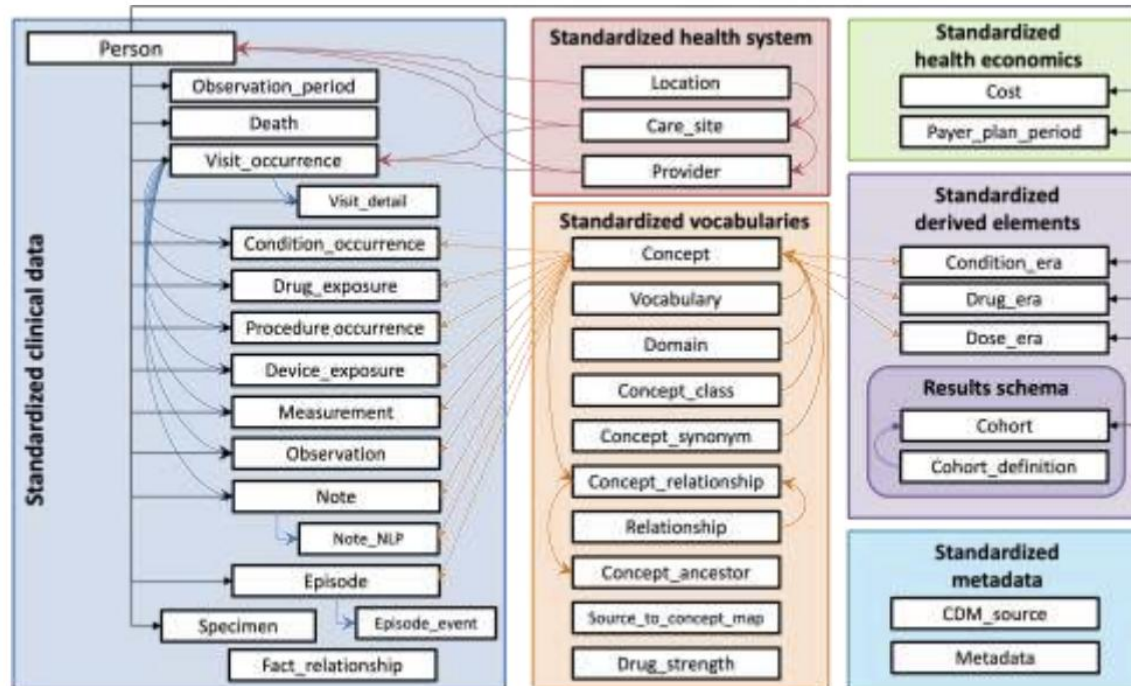


SQL, Structured Query Language.

OMOP Common Data Model Requirements



1. Standardize **data structure** via common format



2. Standardize **data content** via mapping EHR codes to standard healthcare vocabularies

OMOP Domain	Standard Vocabularies	Non-standard Vocabularies
Condition	SNOMED-CT	ICD-10-CM, ICD-9-CM
Drug	RxNorm, CVX	ATC, NDC, Multum
Measurement	LOINC	SNOMED-CT, Nebraska Lexicon
Procedure	CPT4, HCPCS, ICD-10-PCS	ICD-9-Proc
Observation	SNOMED-CT, LOINC	ICD-10-CM, ICD-9-CM
Race, Ethnicity	OMOP Race, OMOP Ethnicity	SNOMED-CT, Nebraska Lexicon
Provider (Specialty)	NUCC, Medicare Specialty	SNOMED-CT, Nebraska Lexicon
Route	SNOMED-CT	Nebraska Lexicon
Unit	UCUM	SNOMED-CT, Nebraska Lexicon

<https://ohdsi.github.io/CommonDataModel>

ATC, Anatomical Therapeutic Chemical; CPT, Current Procedural Terminology; CVX, Clinical Vaccines; HCPCS, Healthcare Common Procedure Coding System; ICD, International Statistical Classification of Diseases; LOINC, Logical Observation Identifiers Names and Codes; NDC, National Drug Code; NUCC, National Uniform Claim Committee; OHDSI, Observational Health Data Sciences and Informatics; OMOP, Observational Medical Outcomes Partnership; PCS, Procedure Coding System; PHI, Protected Health Information; SNOMED, Systematized Nomenclature of Medicine – Clinical Terms; UCUM, Unified Code for Units of Measure.

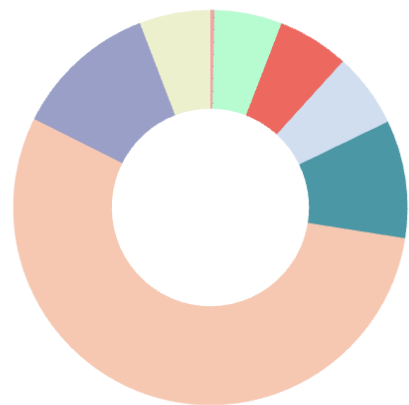
Data2Evidence Limitations

- Not all data modalities on Artificial Intelligence Ready Mount Sinai (AIR·MS) are currently available for search in Data2Evidence
- Radiology study data is older than a year
- Cannot filter with concept sets (also called value sets) at this time
- Filters may surface concepts in the dictionary table which are not used/mapped at Mount Sinai
 - Mitigate by searching for concepts of interest as a separate step before building cohorts
- Multiple filter criteria can degrade responsiveness
- Cannot currently download list of patients in the cohort

Follow along (needs VPN or on-site non-guest connection)

<https://d2e.airms.mssm.edu>

Datasets



Mount Sinai De-Identified Epic Electronic Health Record (EHR) Data

Click here to start analyzing AI Ready Mount Sinai (AIRMS) de-identified patient data derived from Mount Sinai's EHR, and stored in Observational Medical Outcomes Partnership (OMOP) format.

 Total subjects: 12,441,531  Date: 2023-06-08  Version: 5.3

 Data model: omop5-3



Dataset Overview



D2E Datathon Dataset Concepts Cohorts Wizards Account

Dataset info Data characterization Data quality

Datathon

The dataset consists of EHR data from ~5,000 patients with diagnoses relevant to the hackathon. The dataset is designed for participants to develop ECG representations that can identify patient matches, detect individuals outside the cohort, and extract clinical signals that provide information beyond the electronic health record. The dataset is a subset of the Mount Sinai De-Identified Epic Electronic Health Record (EHR) Data, structured with the OMOP Common Data Model format. Data includes patient demographics, conditions, encounters, procedures, vitals, medications, and lab results. We also provide the count of ECGs each patient has.

Metadata

Resource type	Dataset
Dataset ID	b74a57f2-026a-4f46-ba7f-3a59d5c67250
Entity Count Distribution	{"Observation Period Count": "4855", "Death Count": "604", "Visit Occurrence Count": "967358", "Visit Detail Count": "0", "Condition Occurrence Count": "1305352", "Drug Exposure Count": "2250423", "Procedure Occurrence Count": "2269754", "Device Exposure Count": "0", "Measurement Count": "17820345", "Observation Count": "2395372", "Note Count": "1521567", "Specimen Count": "0"}
Entity Count	28,535,630
Version	5.3
Schema Version	5.3
Latest Available Schema Version	5.3

>28M entities

Dataset Characterization

Dataset info

Data characterization

Data quality

Mount Sinai De-Identified Epic Electronic Health Record (EHR) Data

Select data characterization report

Show all reports

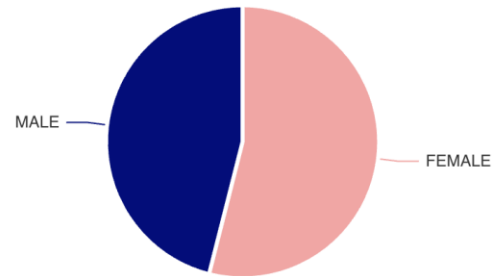
CDM Summary

Source name: CDMDEID

Number of persons:
12,441,531

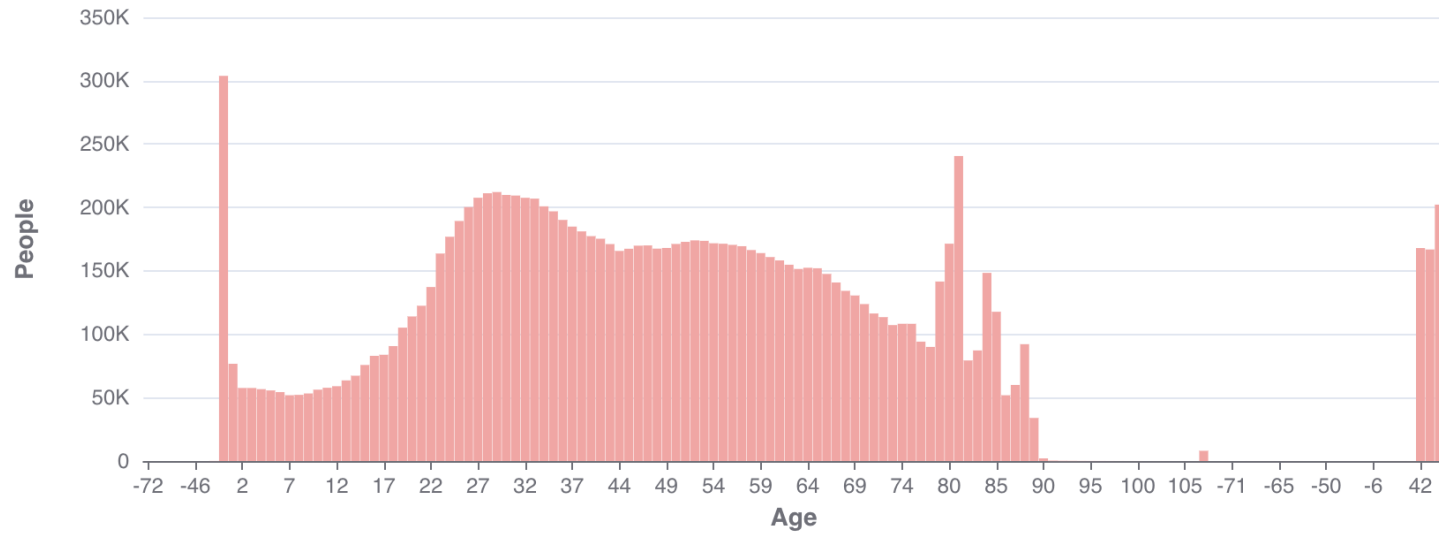
Gender

FEMALE MALE



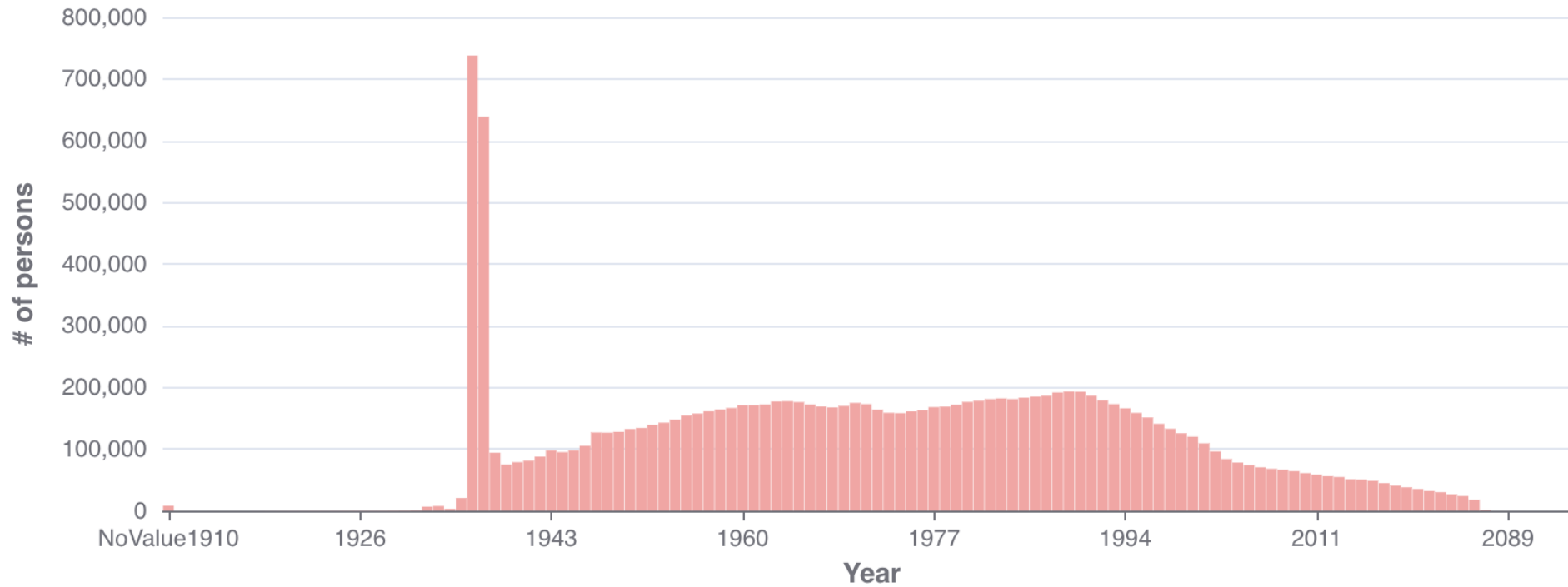
Dataset Characterization

Age at first observation



Dataset Characterization

Year of birth



Dataset Characterization

Measurement Prevalence

Treemap Table



Leukoc...	Alanine ...	Aspartat...	Body te...	Bilirub...	Body s...	Bl...	Bi...	pH...	La...	Ox...	Er...	Tr...	Pa...	Ox...	Glo
Platele...	Glucose ...	Carbon ...	Prothro...	Oxygen...	Neuro...										
				Respira...	Body h...	Glome...	Po...	Li...	H...	C...	Bl...	Cho	Plat	ABO	Thy
Creatin...	Order set														
Urea ni...	Platelets panel Blood Hematology and Cell Count Panels														
Potassi...	Platelets Number Concentration (count/vol) Moment in time Blood														
Sodium...	Platelets [#./volume] in Blood by Automated count														
	Prevalence: 0.00508														
	Number of persons: 63,250														
	Average records per person: 50.4														
MCV [E...]			Systol...	Nitrite...	Body m...	Specif...	Leuk...								
	Erythro...	Basophil...	Diastol...	Urobili...	Sodium...	Gluco...	Oxyg...								
Albumi...	Alkaline...	Eosinop...	Phosph...	Leukoc...	Glucos...	Trigly...	Rh [T...								
Bilirub...	Chloride...	Basophil...	Body w...	Bilirub...	Carbon...	Carbo...	Chlor...								

1.00 672.00



Box Size: Prevalence



Data Quality Dashboard (DQD)

Mount Sinai De-Identified Epic Electronic Health Record (EHR) Data

Overview

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	327	125	452	72%	286	5	291	98%	613	130	743	83%
Conformance	638	90	728	88%	102	4	106	96%	740	94	834	89%
Completeness	374	22	396	94%	12	5	17	71%	386	27	413	93%
Total	1,339	237	1,576	85%	400	14	414	97%	1,739	251	1,990	87%

570 out of 1,739 passed checks are not applicable, due to empty tables or fields.

28 out of 251 failed checks are SQL errors.

Corrected pass percentage for NA and Errors: 84% (1,169/1,392).

Concept Search



D2E ← **Datathon** ▼ Dataset **Concepts** Cohorts Wizards Account

Concept search Concept sets

Search →

ID	Code	Name	Vocabulary	Concept	Domain	Class	Validity	RC
			Filter by Vocabulary ×	Filter by Concept ×	Filter by Domain ×	Filter by Class ×	Filter by Validity ×	
45426237	9982.00		Read	Non-standard	Condition	Read	Invalid	0
45489629	9C3.12		Read	Non-standard	Observation	Read	Invalid	0
45426233	98Bn.00		Read	Non-standard	Condition	Read	Invalid	0
45429577	9E3.12		Read	Non-standard	Procedure	Read	Invalid	0
45422998	9DA.11		Read	Non-standard	Observation	Read	Invalid	0
45476340	92...11		Read	Non-standard	Observation	Read	Invalid	0
45486357	9N1F.13		Read	Non-standard	Observation	Read	Invalid	0
45489627	999Z.00		Read	Non-standard	Condition	Read	Invalid	0
45422994	98Bj.00		Read	Non-standard	Condition	Read	Invalid	0
45486335	9991.00		Read	Non-standard	Condition	Read	Invalid	0
45432915	9Ndr.00		Read	Non-standard	Condition	Read	Invalid	0
45486356	9K4.00		Read	Non-standard	Observation	Read	Invalid	0
45483694	K522		Read	Non-standard	Condition/Procedure	Read	Invalid	0
45423012	9F2.11		Read	Non-standard	Observation	Read	Invalid	0
45479742	9Nds.00		Read	Non-standard	Condition	Read	Invalid	0



Example Scenario

How many patients who identify as White do we have with multiple myeloma who were treated with bortezomib, and had a whole-body MRI?



Cohort Building

The screenshot shows the D2E Cohort Building interface. At the top, a navigation bar includes the D2E logo, a back arrow, a 'Datathon' dropdown menu, and tabs for 'Dataset', 'Concepts', 'Cohorts', 'Wizards', and 'Account'. The 'Cohorts' tab is selected. Below the navigation bar, the 'Cohorts' section is titled. Underneath, there is a 'Create Cohort:' section with a dark blue button labeled 'D2E', a light blue button labeled 'Compare', and a 'Shared' toggle switch. Below this, a message reads: 'You have not yet saved any filters. You can save your current filter settings.' Two blue callout arrows are present: one labeled '1' points to the 'Cohorts' tab, and another labeled '2' points to the 'D2E' button.



Filtering area

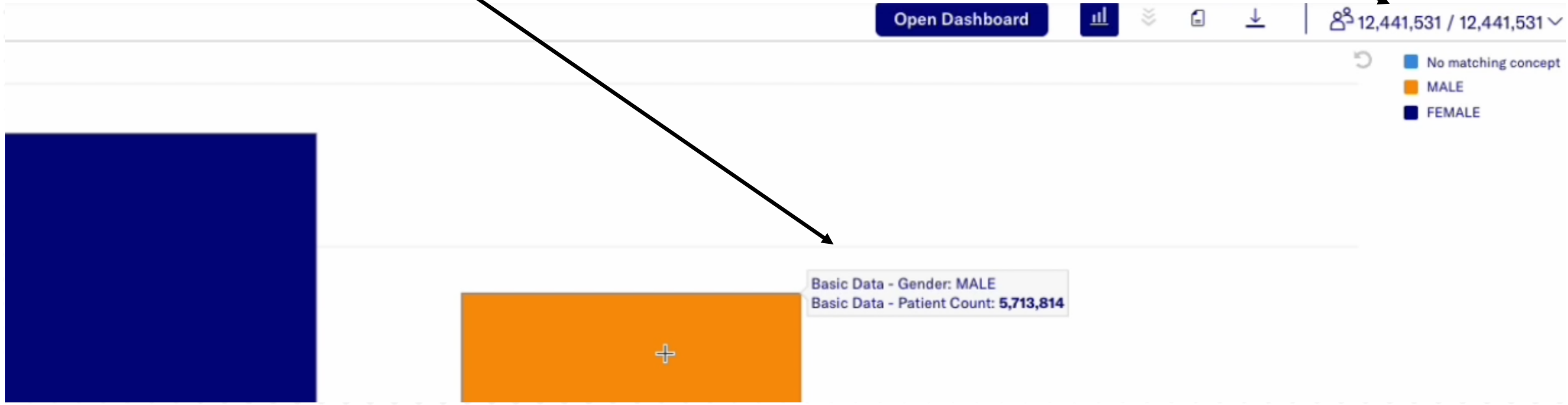
Cohort creator interface

The screenshot displays the Cohort creator interface. On the left, a red-bordered box highlights the 'Filtering area', which includes a sidebar with 'Basic Data' filters for Gender, Age, Race, and Ethnicity, and 'INCLUSION (0)' and 'Exclusion (0)' sections. The main area is divided into a 'Charting area' (green-bordered) and a 'Filtering area' (red-bordered). The charting area features a bar chart with a y-axis from 0 to 7M and an x-axis with categories: FEMALE, MALE, and No matching concept. The bars are colored dark blue for FEMALE and orange for MALE. A legend in the top right of the charting area identifies the colors: No matching concept (light blue), MALE (orange), and FEMALE (dark blue). The charting area also includes a 'Sort X-Axis By' dropdown set to 'Ascending', and 'Basic Data' dropdowns for 'Patient Count' and 'Gender'. The filtering area includes 'Add Filters' and 'Save' buttons at the bottom.

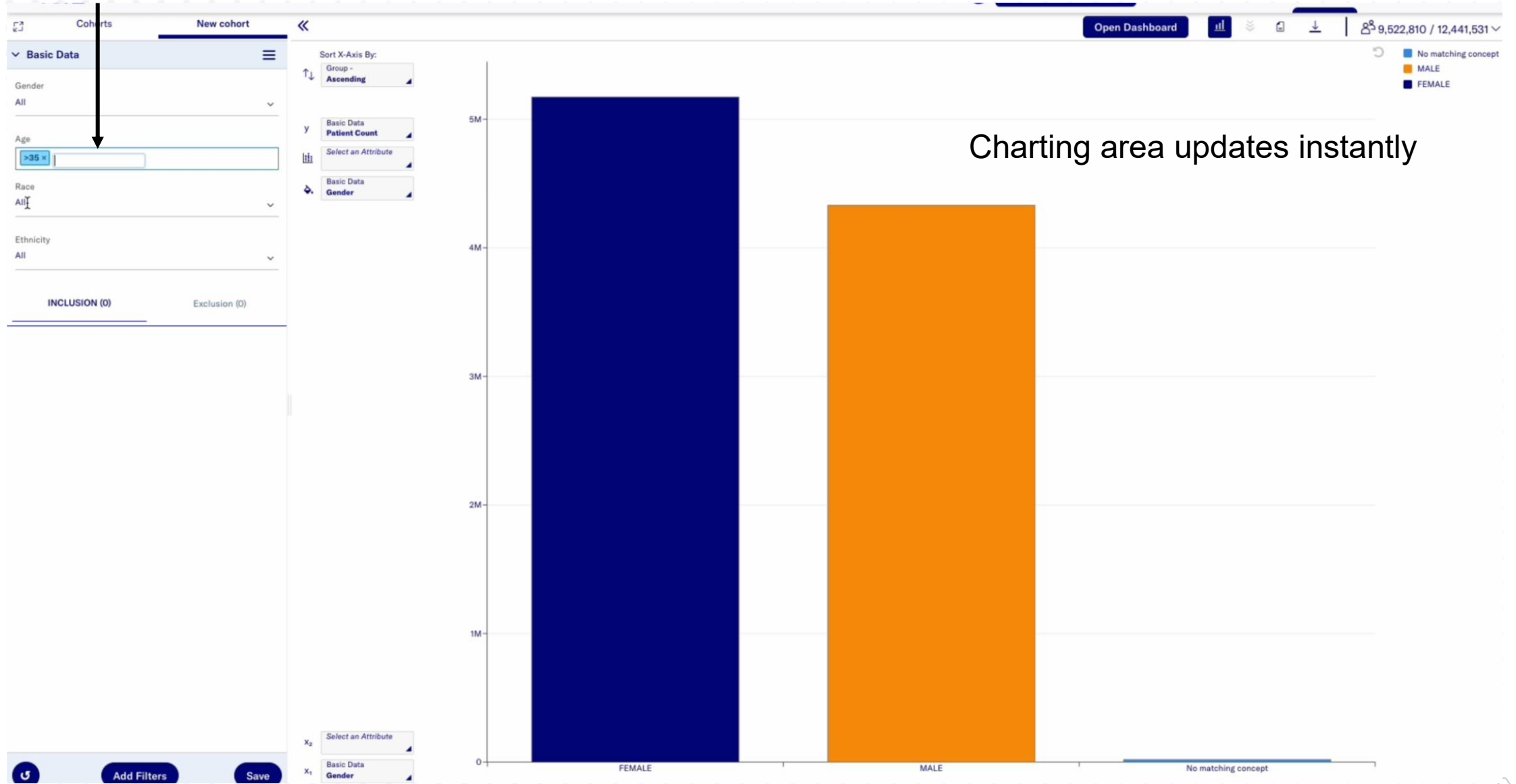
Category	Count (Approximate)
FEMALE	6.7M
MALE	5.7M
No matching concept	0.1M

Hover over bars to see counts

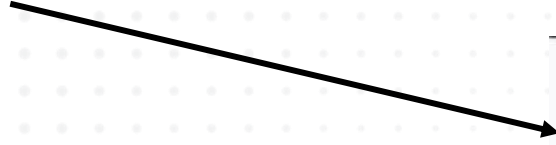
Current cohort size



Add filter Age > 35



Add more demographic filters



D2E

Cohorts **New cohort** <<

Basic Data

Gender

FEMALE x MALE x

Clear All

Age

>35 x

Race

White x

Clear All

Ethnicity

All

INCLUSION (1) Exclusion (0)

Condition filtering by ICD-10

Cohorts **New cohort** <<

Basic Data

Gender
FEMALE x MALE x
Clear All

Age
>35 x

Race
White x
Clear All

Ethnicity
All

INCLUSION (1) Exclusion (0)

Conditions (ICD10CM) A

Condition Source concept code
C90.d
C90.0 Loading suggestions...



INCLUSION (1) Exclusion (0)

Conditions (ICD10CM) A

Condition Source concept code

- Multiple myeloma not having achieved remission x
- Multiple myeloma in remission x
- Multiple myeloma in relapse x

Enter search term

C90.0 - Multiple myeloma

Search for medications by code

INCLUSION (2) Exclusion (0)

▼ Conditions (ICD10CM) A

Condition Source concept code

- Multiple myeloma not having achieved remission x
- Multiple myeloma in remission x
- Multiple myeloma in relapse x

Clear All

AND

▼ Medications A

Medication Source concept code

L01xg01

- L01xg01
- L01XG01 - bortezomib

We can flip the Boolean relationship between filter cards

The screenshot displays a search filter interface with two main sections: INCLUSION (2) and Exclusion (0). The INCLUSION section is expanded to show 'Conditions (ICD10CM) A' and 'Medications A'. Under 'Conditions (ICD10CM) A', there are three filter cards: 'Multiple myeloma not having achieved remission', 'Multiple myeloma in remission', and 'Multiple myeloma in relapse'. A red circle highlights a button with a double-headed arrow and a refresh icon, indicating the ability to toggle the Boolean relationship between the filter cards. The 'Medications A' section has one filter card: 'bortezomib'. Both sections have a 'Clear All' button.

INCLUSION (2) Exclusion (0)

Conditions (ICD10CM) A

Condition Source concept code

Multiple myeloma not having achieved remission x

Multiple myeloma in remission x

Multiple myeloma in relapse x

Clear All

Medications A

Medication Source concept code

bortezomib x

Clear All

Add temporal relationship between filters

INCLUSION (2) Exclusion (0)

> Conditions (ICD10CM) A

AND

Medications A

Medication Source concept code

bortezomib x

Clear All

Medications A - Conditions (ICD10CM) A +


Started


1 Days


After Start of


Conditions (ICD10CM) A


Filter by radiology studies

> Conditions (ICD10CM) A 



AND 

> Medications A 

AND 



▼ Imaging Research Warehouse A 

Imaging Modality

MAGNETIC RESONANCE  

Clear All


Body Part

WHOLE BODY  

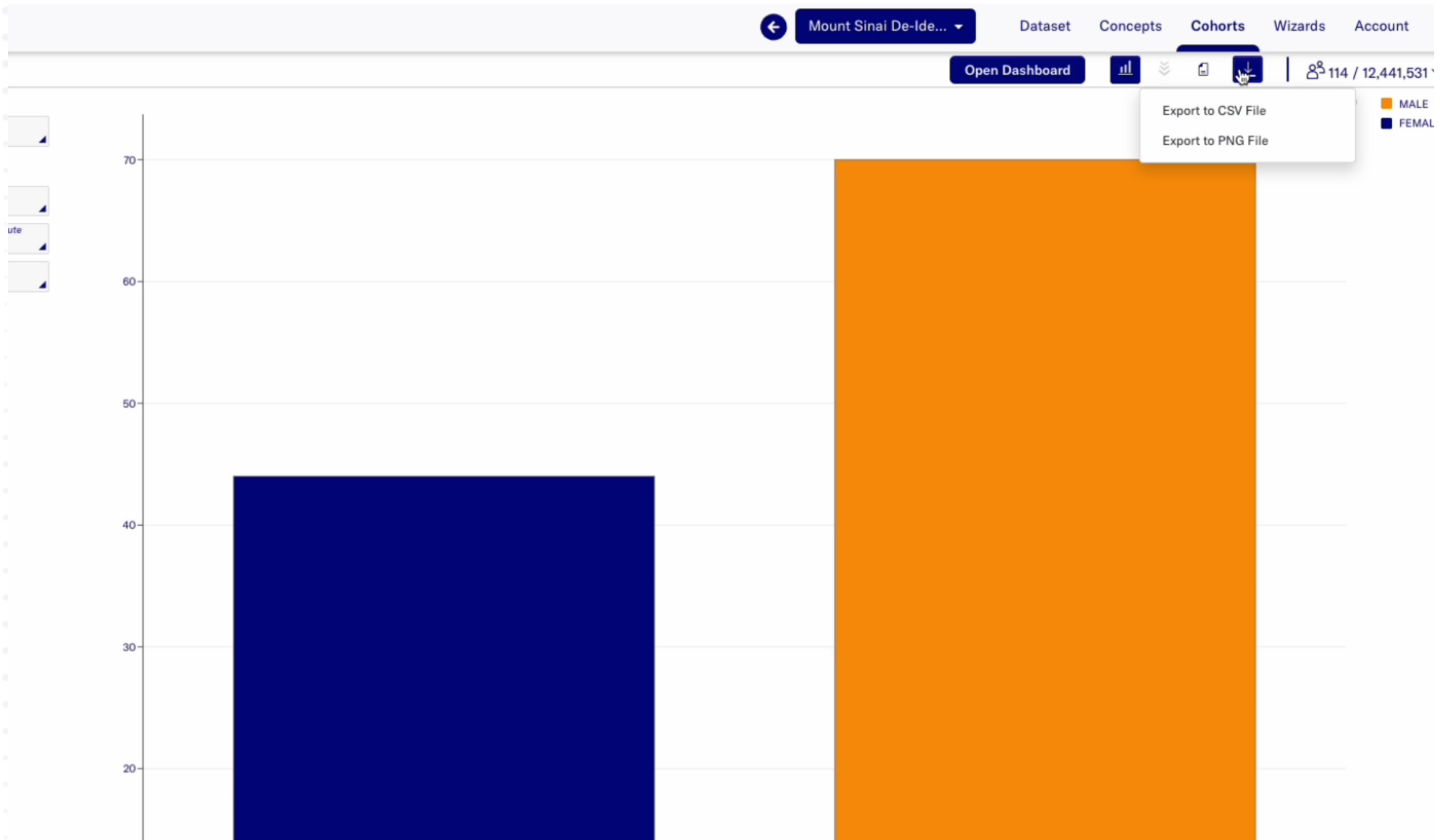
Clear All

Frequency

Study Description

All 

Download the chart



Download the SQL query

Dataset Concepts Cohorts Wizards Account

Dashboard | 114 / 12,441,531

Filter Summary ×

- Showing patients with:

Basic Data
Gender: FEMALE, MALE
Age: >35
Race: White

AND

Conditions (ICD10CM) A
Condition Source concept code: C90.00, C90.01, C90.02

AND

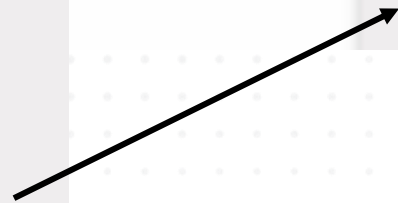
Medications A
Medication Source concept code: L01XG01
Started 1 Days after_startdate of Conditions (ICD10CM) A

AND

Imaging Research Warehouse A
Imaging Modality: MAGNETIC RESONANCE
Body Part: WHOLE BODY

Download SQL **Copy SQL**

Download SQL **Copy SQL**



Wizards in Data2Evidence

What are Wizards?

- These are pre-configured queries that can calculate common epidemiological measures, such as incidence, prevalence, mortality, etc.
- This will speed up calculations, and allow users to get quick insights with only a few mouse clicks
- Future wizards will include readmission rates, length of stay, etc.
- There **may be edge cases where the Wizards fail** - if you find these, please contact us and we'll improve it

Getting started

We've built some pre-configured scenarios to get you started

Calculate Incidence

This wizard will calculate the incidence (the number of new cases) for a condition or multiple co-occurring conditions, such as diabetes, stroke, or COVID-19. It will find patients with a particular diagnostic code or codes for a set of criteria you define for the patient cohort (dates, patient age, gender, weight, systolic blood pressure, etc).

Calculate Prevalence

Prevalence (the number of existing cases) for a particular condition or multiple co-occurring conditions, such as diabetes, stroke, or COVID-19 can be easily calculated using this wizard. Other criteria (such as patient age, weight, systolic blood pressure, etc) can also be set to further refine the prevalence rate calculation.

Calculate Mortality

This will determine the mortality rate for a specific condition or multiple co-occurring conditions. Mortality is determined by finding specific death dates that co-occur with a range of dates that you specify, as well as additional patient criteria you set (such as visit dates, patient age, gender, etc).

Cross sectional Demographics

The cross-sectional demographics analysis executed by this wizard will create a set of graphs describing a specific cohort of patients you define based on different criteria, such as age, gender, race, etc. This is useful to provide a quick exploration of data and generate hypotheses.

More wizards will appear here in the future

Calculate Incidence

This wizard will calculate the incidence (the number of new cases) for a condition or multiple co-occurring conditions, such as diabetes, stroke, or COVID-19. It will find patients with a particular diagnostic code or codes for a set of criteria you define for the patient cohort (dates, patient age, gender, weight, systolic blood pressure, etc).

Note: this is a very rough approximation that is just a starting point for a more comprehensive analysis.

Age Range*: **>=60**

Gender: **Male**

Ethnicity:

Race:

Height:

Weight:

BMI:

Resp Rate:

Pulse Rate:

Systolic Blood Pressure:

Diastolic Blood Pressure:

Let's calculate diabetes Incidence for males >= 60

Systolic Blood Pressure:

Diastolic Blood Pressure:

Years*: -

Condition 1*: **include descendants**

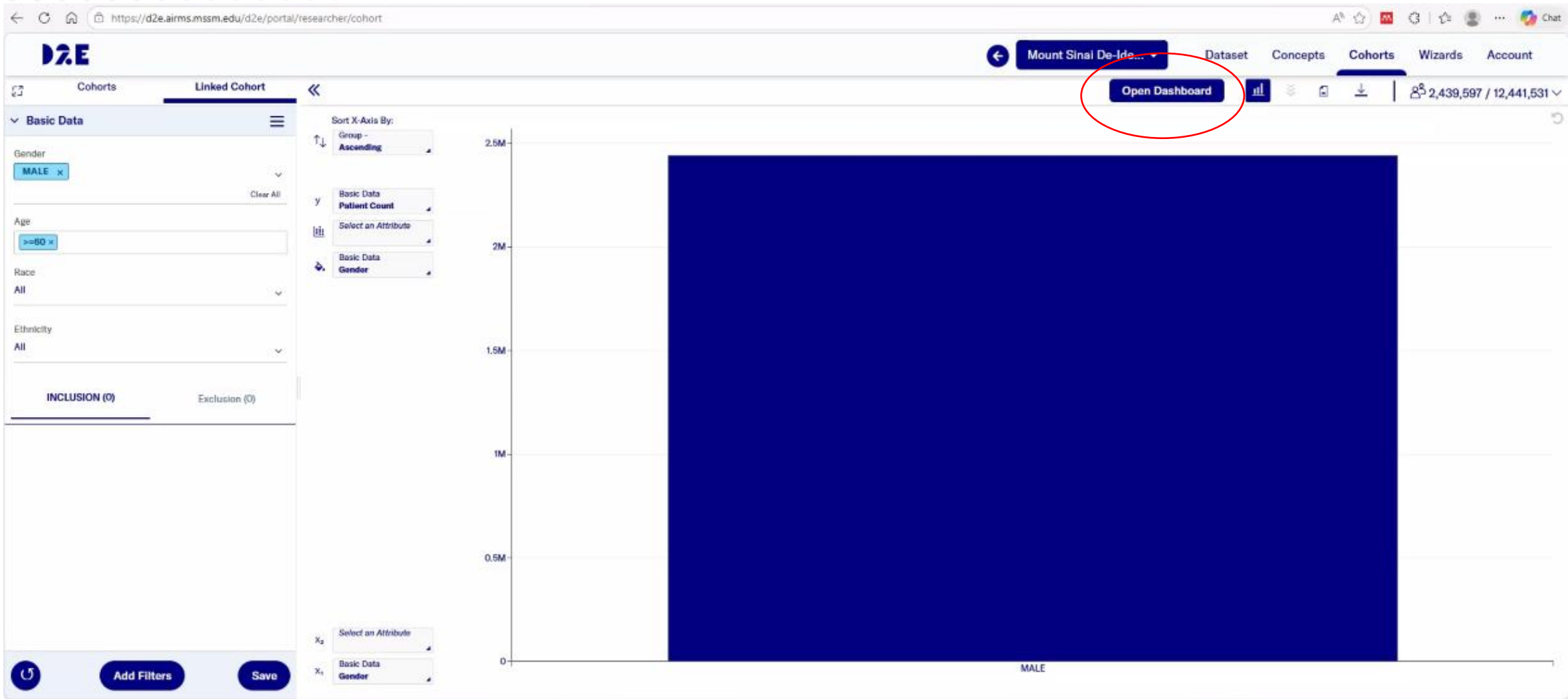
Condition 2:

Condition 3:

Condition 4:

Condition 5:

Make sure to click this





Open Dashboard

2,439,597 / 12,441,531

Basic Data

Gender: MALE

Age: >=80

Race: All

Ethnicity: All

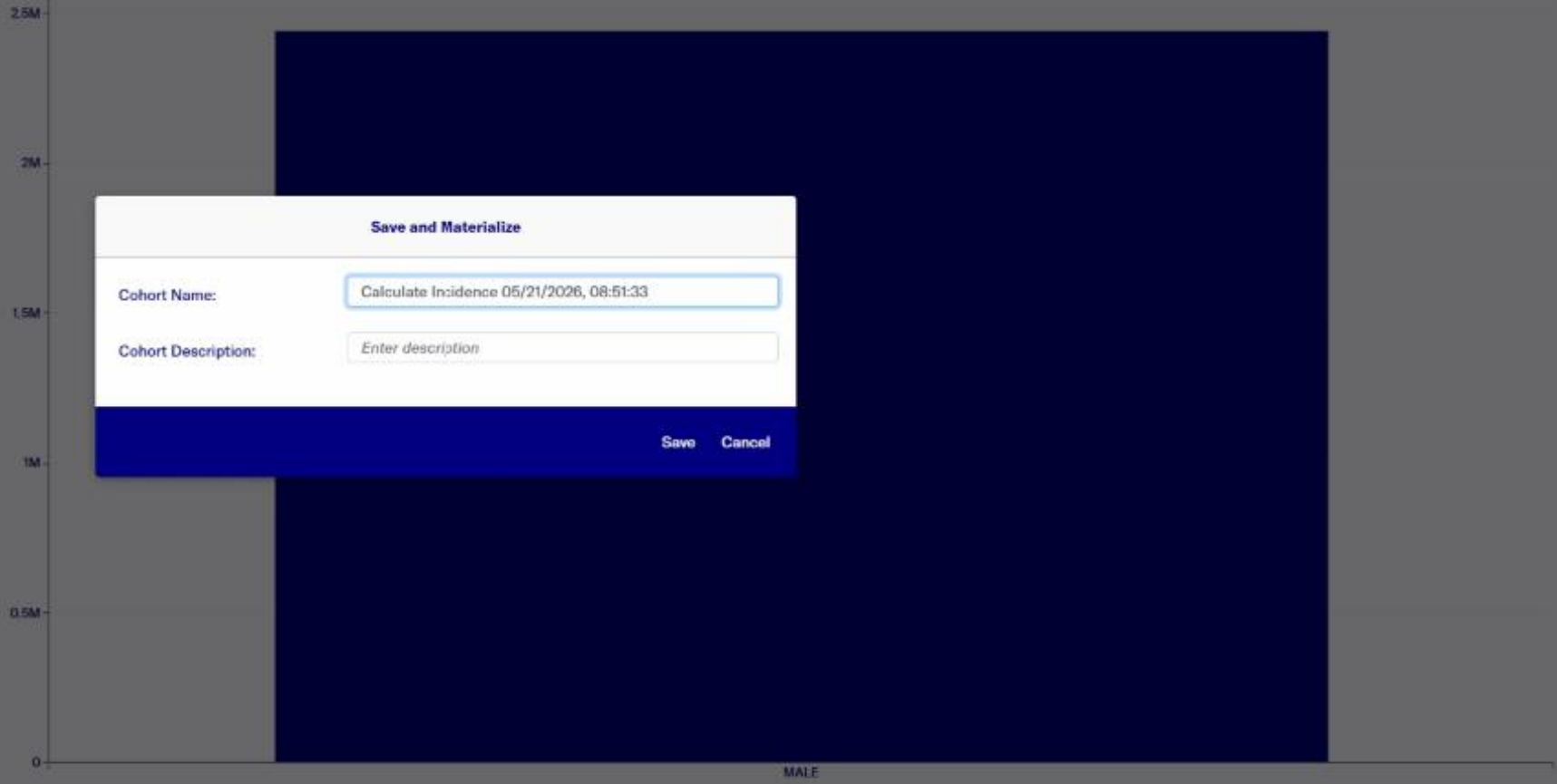
INCLUSION (0) Exclusion (0)

Sort X-Axis By:
Group - Ascending

Basic Data
Patient Count

Select an Attribute

Basic Data
Gender



Save and Materialize

Cohort Name: Calculate Incidence 05/21/2026, 08:51:33

Cohort Description: Enter description

Save Cancel

Select an Attribute

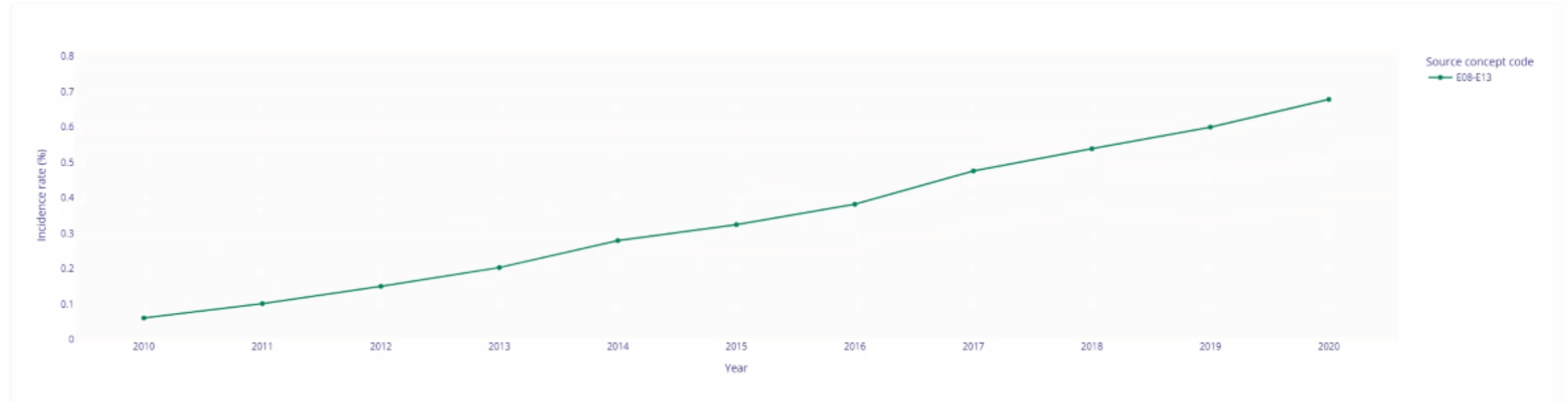
Basic Data
Gender



Dashboard

Incidence rate dashboard

Calculated as the percentage of the cohort that had these conditions as an incidence during the years 2010 to 2020 based on condition start date, and did not have the conditions a year prior.





Getting started

We've built some pre-configured scenarios to get you started

Calculate Incidence

This wizard will calculate the incidence (the number of new cases) for a condition or multiple co-occurring conditions, such as diabetes, stroke, or COVID-19. It will find patients with a particular diagnostic code or codes for a set of criteria you define for the patient cohort (dates, patient age, gender, weight, systolic blood pressure, etc).

Calculate Prevalence

Prevalence (the number of existing cases) for a particular condition or multiple co-occurring conditions, such as diabetes, stroke, or COVID-19 can be easily calculated using this wizard. Other criteria (such as patient age, weight, systolic blood pressure, etc) can also be set to further refine the prevalence rate calculation.

Calculate Mortality

This will determine the mortality rate for a specific condition or multiple co-occurring conditions. Mortality is determined by finding specific death dates that co-occur with a range of dates that you specify, as well as additional patient criteria you set (such as visit dates, patient age, gender, etc).

Cross sectional Demographics

The cross-sectional demographics analysis executed by this wizard will create a set of graphs describing a specific cohort of patients you define based on different criteria, such as age, gender, race, etc. This is useful to provide a quick exploration of data and generate hypotheses.

More wizards will appear here in the future

Calculate Prevalence

Prevalence (the number of existing cases) for a particular condition or multiple co-occurring conditions, such as diabetes, stroke, or COVID-19 can be easily calculated using this wizard. Other criteria (such as patient age, weight, systolic blood pressure, etc) can also be set to further refine the prevalence rate calculation.

Note: this is a very rough approximation that is just a starting point for a more comprehensive analysis.

Age Range*: e.g. >=60, [50-80] **>=60**

Gender: Search Gender **Male**

Ethnicity: Search Ethnicity

Race: Search Race

Height: cm

Weight: kg

BMI: kg/m²

Resp Rate: bpm

Pulse Rate: bpm

Systolic Blood Pressure: mmHg

Pulse Rate: bpm

Systolic Blood Pressure: mmHg

Diastolic Blood Pressure: mmHg

Years*: 2010 - 2020

Condition 1*: Diabetes mellitus (E08-E13) Include descendants

Condition 2*: Essential (primary) hypertension (I10) Include descendants

Condition 3: Search Condition 3

Condition 4: Search Condition 4

Condition 5: Search Condition 5

Back

Let's calculate diabetes and hypertension prevalence for males >= 60

https://d2e.aims.mssm.edu/d2e/portal/researcher/cohort

D2E Mount Sinai De-Ide... Dataset Concepts Cohorts Wizards Account

Cohorts **Linked Cohort**

Basic Data

Gender: MALE x Clear All

Age: >=60 x

Race: All

Ethnicity: All

INCLUSION (0) Exclusion (0)

Sort X-Axis By: Group - Ascending

Y: Basic Data Patient Count

Select an Attribute

Basic Data Gender

X₂: Select an Attribute

X₁: Basic Data Gender

Open Dashboard

Group	Patient Count
MALE	~2,439,597

2.5M

2M

1.5M

1M

0.5M

0

MALE

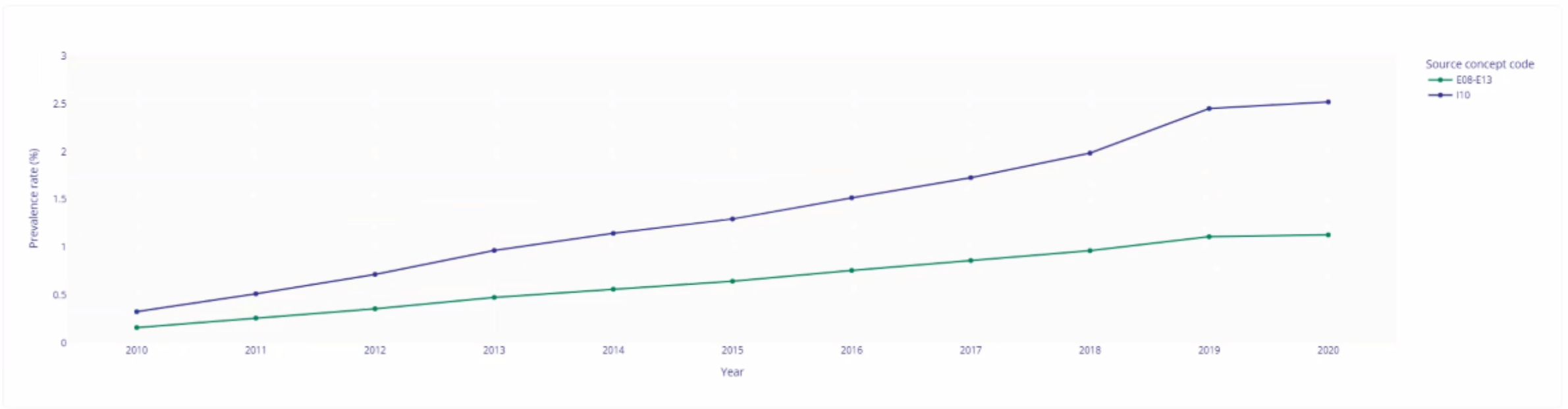
Refresh Add Filters Save



Dashboard

Prevalence rate dashboard

Calculated as the percentage of the cohort that had these conditions during the years 2010 to 2020 based on condition start date





Thank you!