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

HPI·MS

Mount Sinai

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

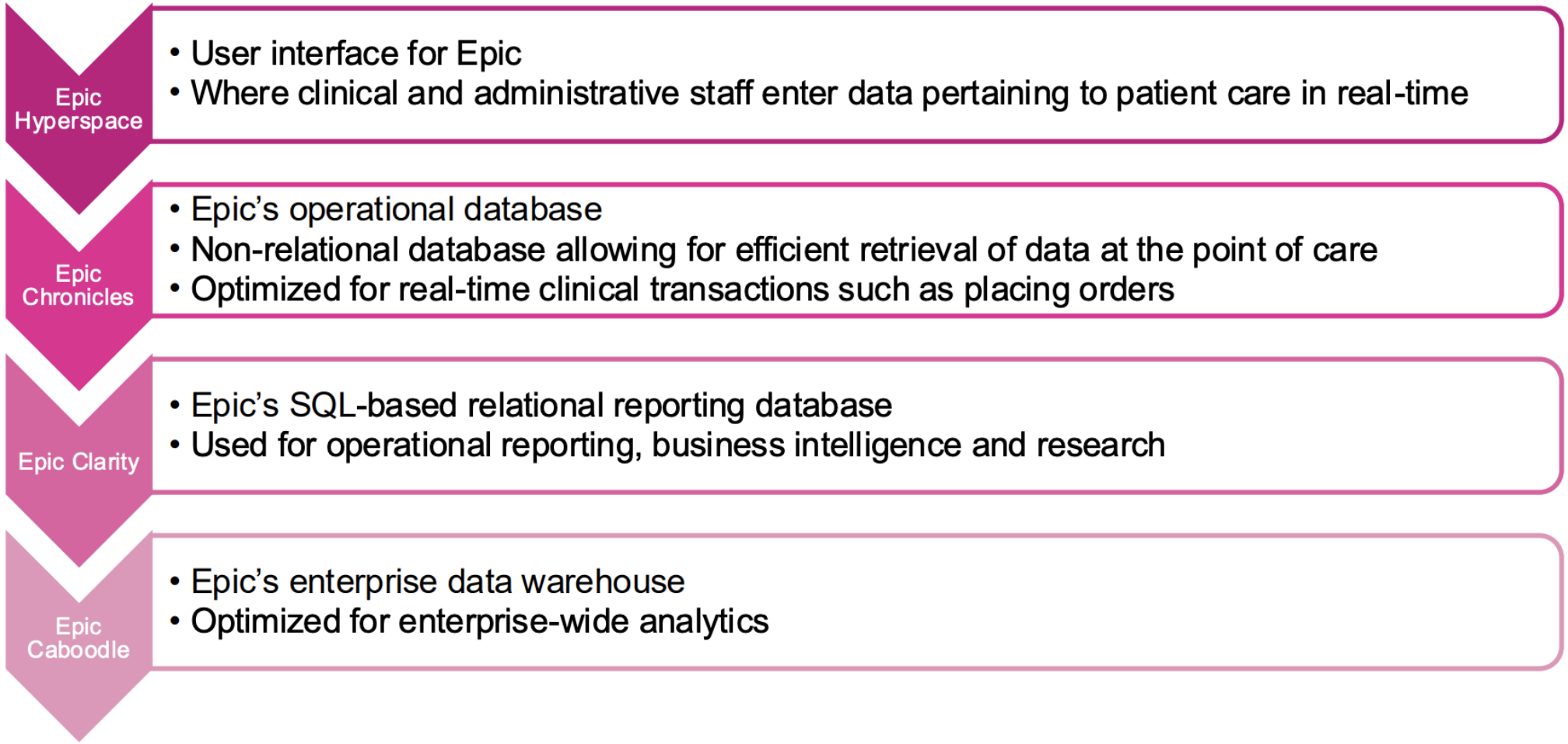
Currently Available Modalities:

- 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

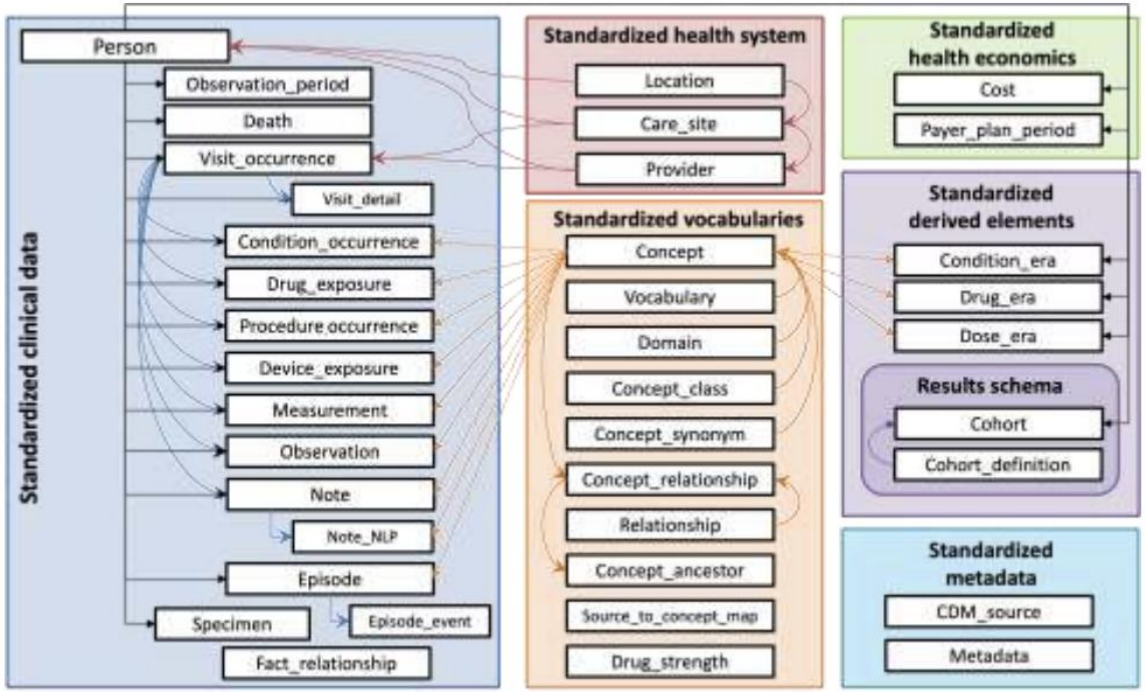
Epic is the Primary Source of Electronic Health Record Data at Mount Sinai



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.

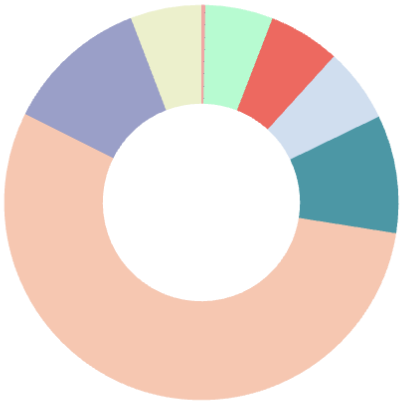
Data2Evidence Limitations

- 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



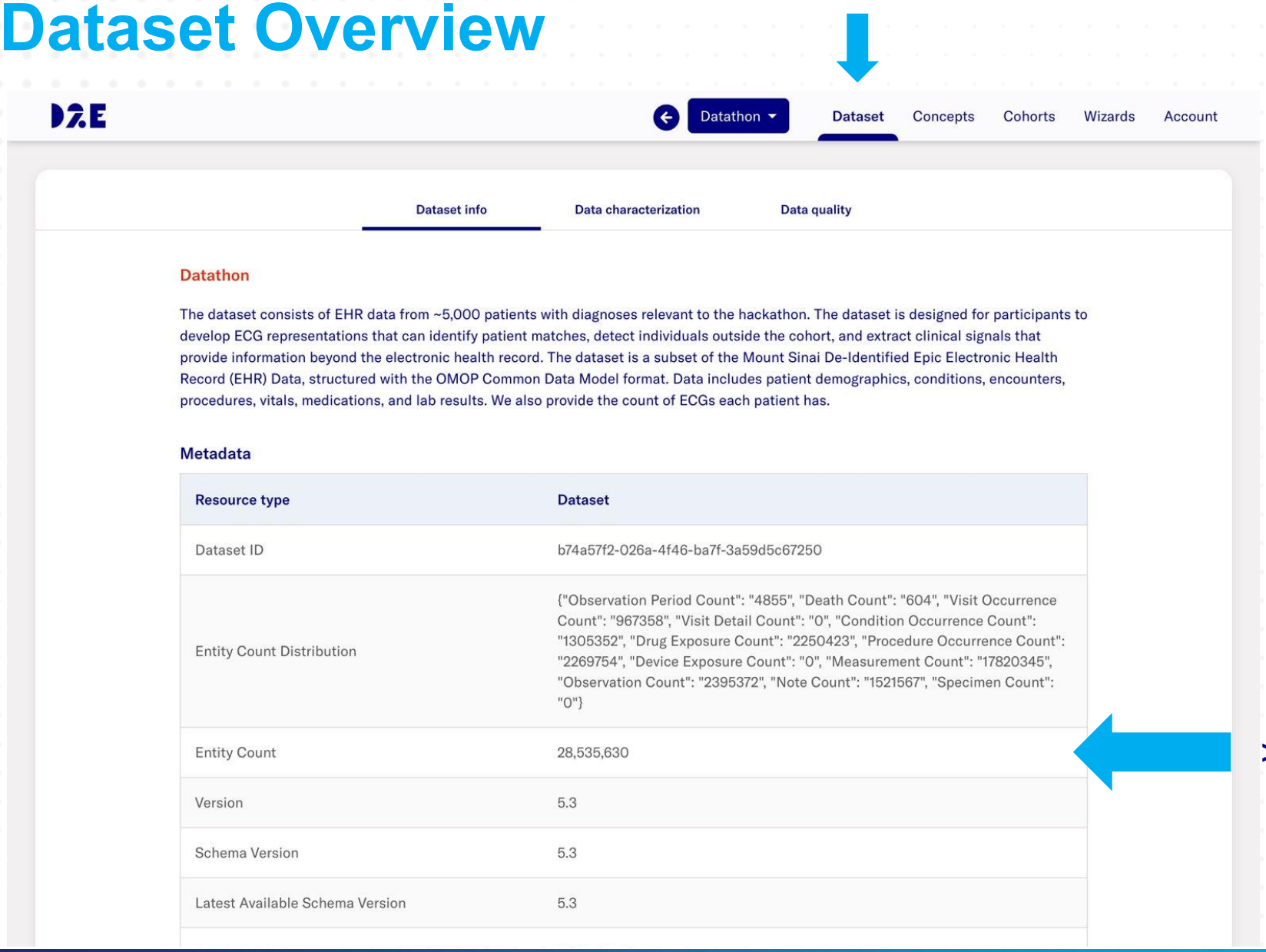
Mount Sinai De-Identified Epic Electronic Health Record (EHR) Data

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



>28M entities

Slide credit: Pablo Guerrero, Tim Walz, Data4Life

Dataset Characterization

Dataset info | Data characterization | Data quality

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

Select data characterization report

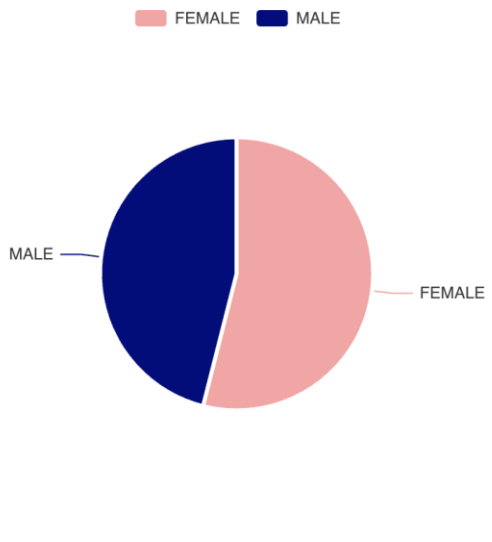
Show all reports

CDM Summary

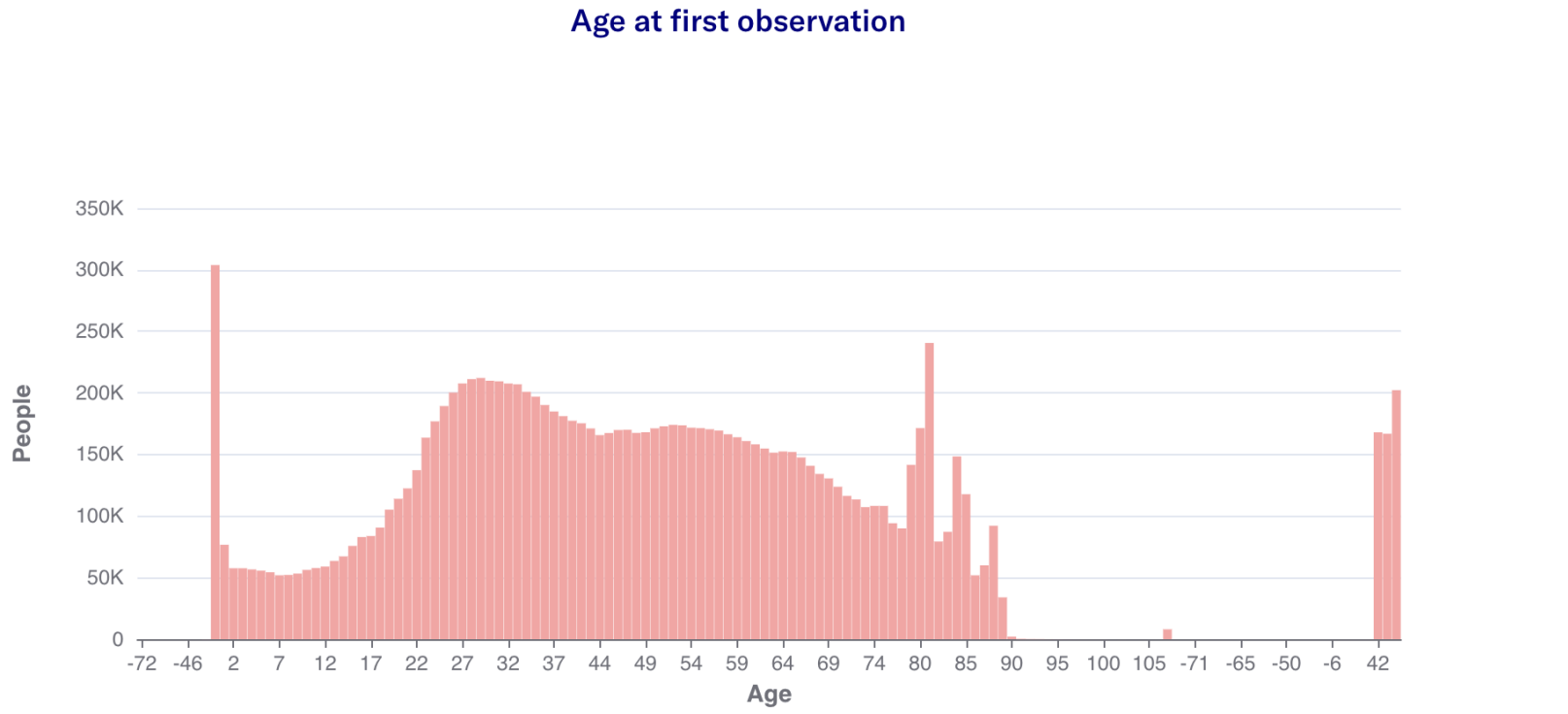
Source name: CDMDEID

Number of persons: 12,441,531

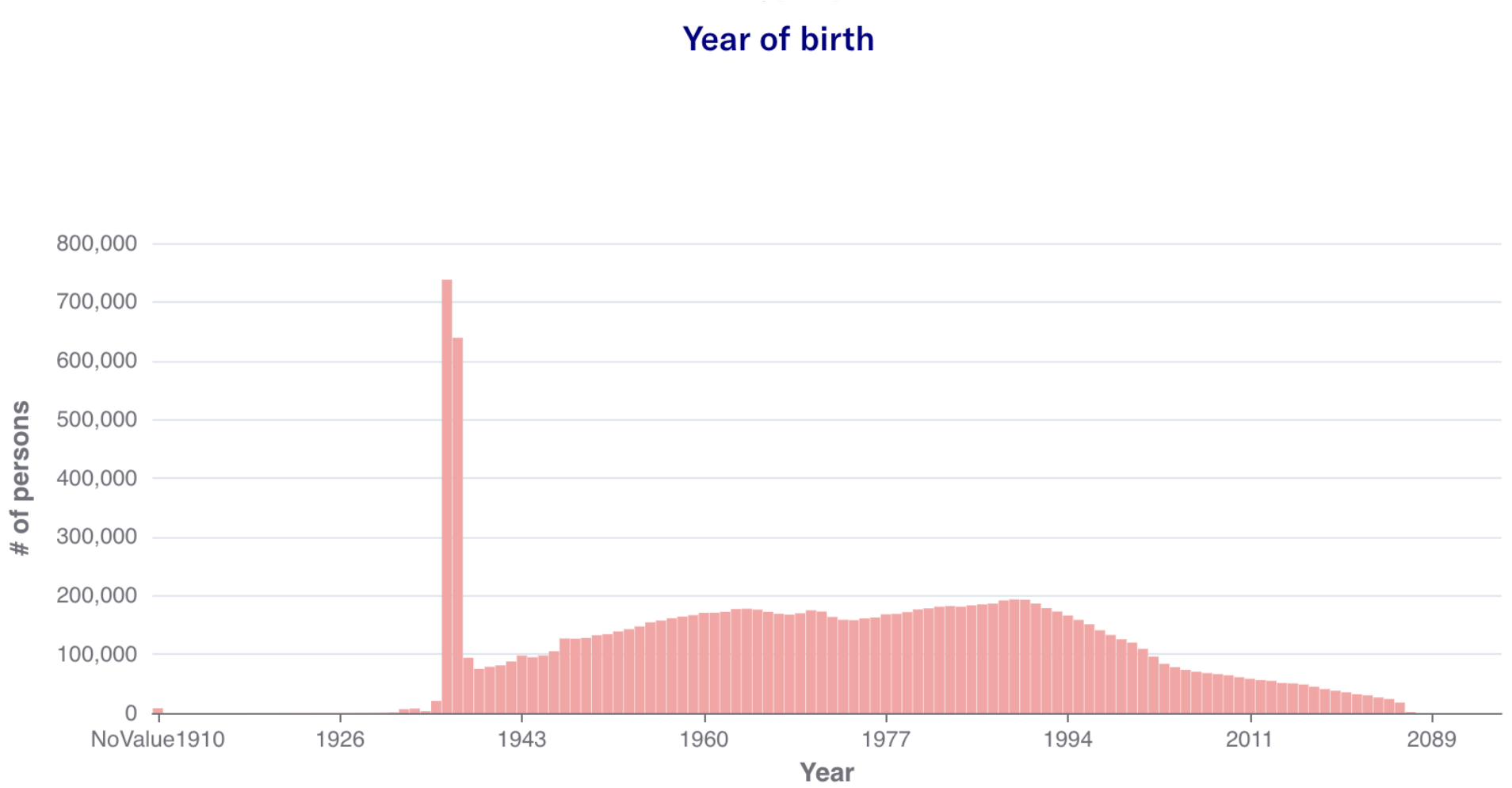
Gender



Dataset Characterization



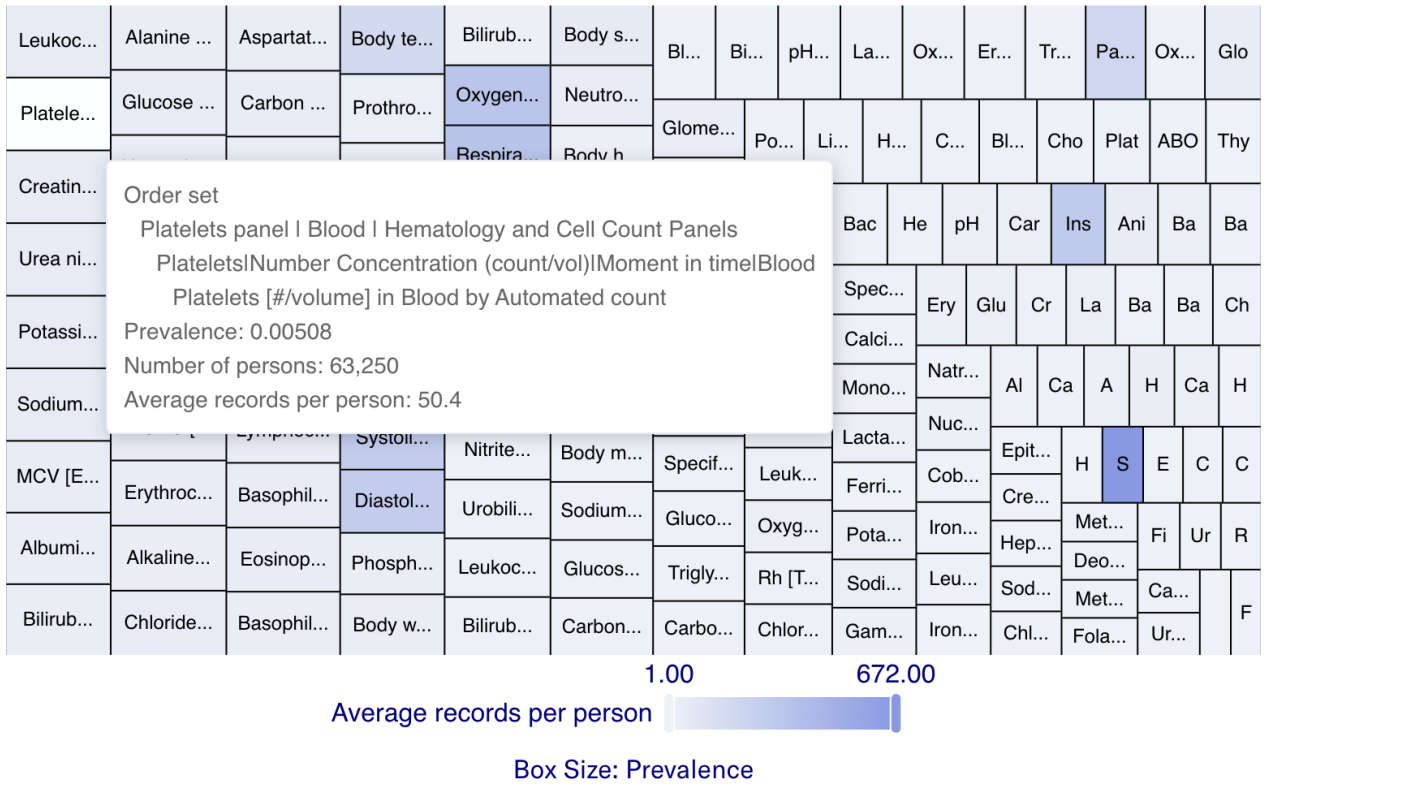
Dataset Characterization



Dataset Characterization

Measurement Prevalence

Treemap | Table



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%

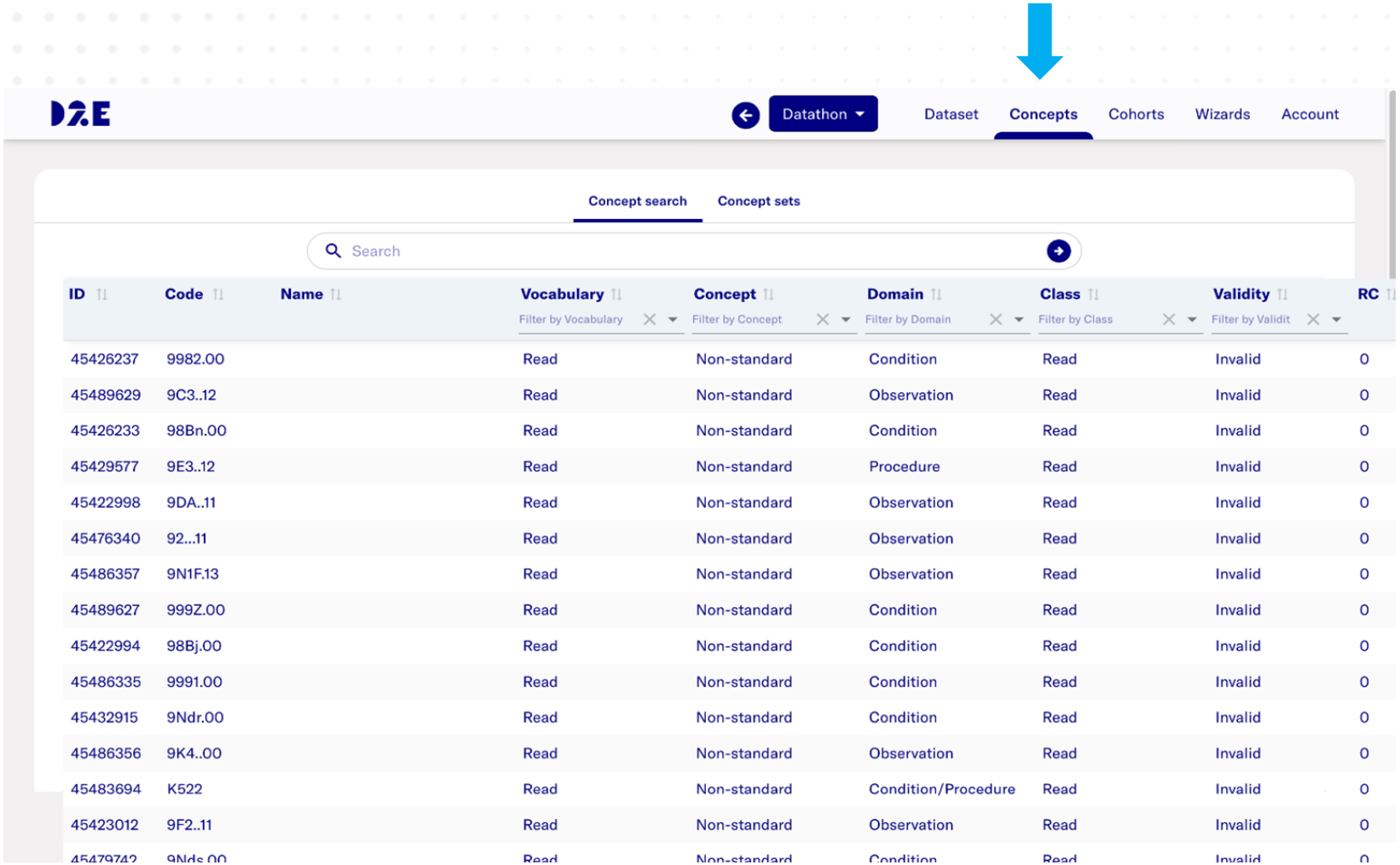
570 out of 1,739 passed checks are not applicable, due to empty tables or fields.

28 out of 251 failed checks are SQL errors.

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

Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 2016;4(1):18. doi:10.13063/2327-9214.1244

Concept Search



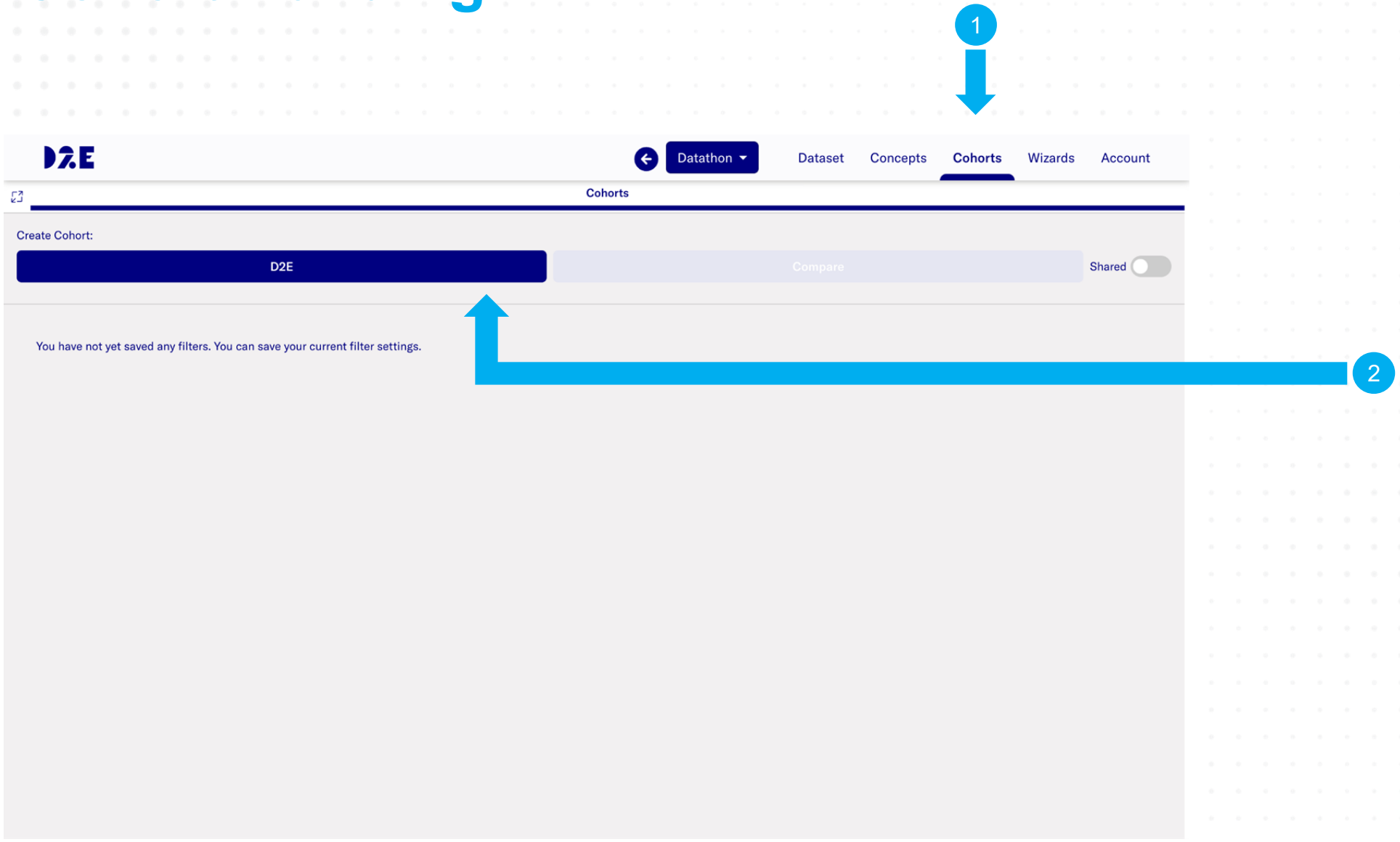
ID	Code	Name	Vocabulary	Concept	Domain	Class	Validity	RC
			Filter by Vocabulary	Filter by Concept	Filter by Domain	Filter by Class	Filter by Validit	
45426237	9982.00		Read	Non-standard	Condition	Read	Invalid	0
45489629	9C3..12		Read	Non-standard	Observation	Read	Invalid	0
45426233	98Bn.00		Read	Non-standard	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

Slide credit: Pablo Guerrero, Tim Walz, Data4Life

Example Scenario

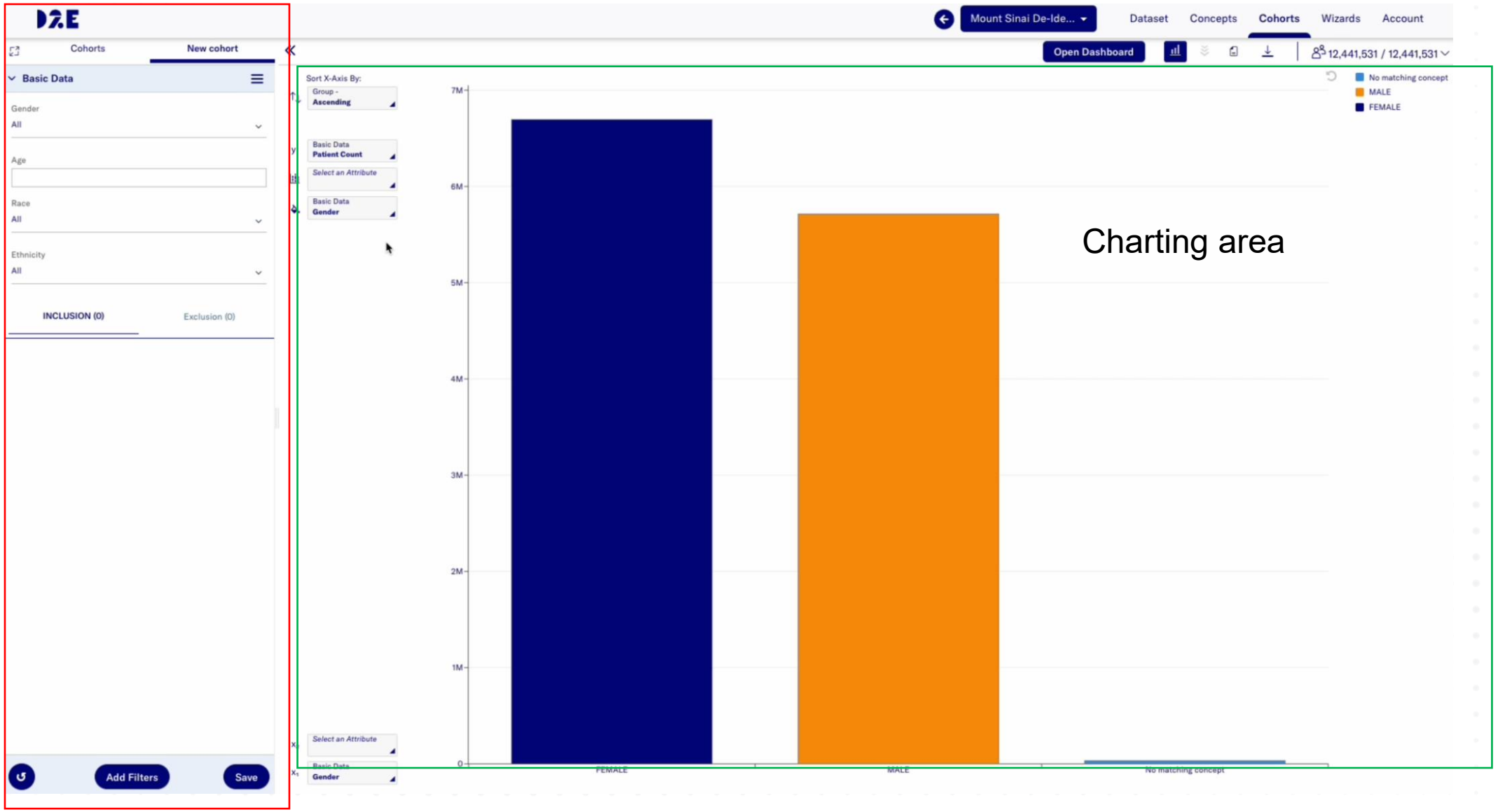
How many patients who identify as White do we have with multiple myeloma who were treated with bortezomib, and had a whole-body MRI?

Cohort Building



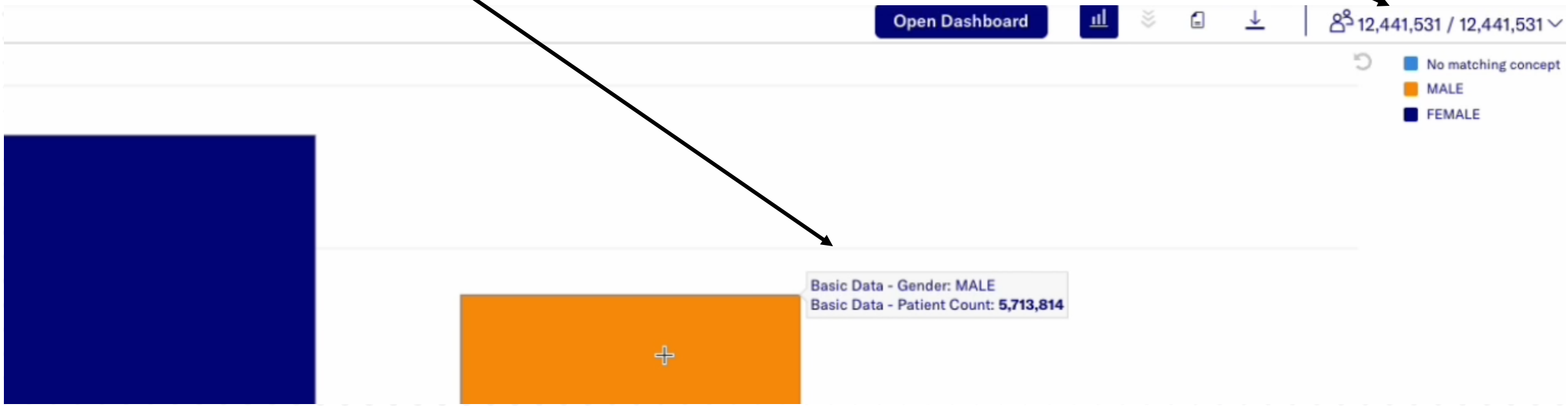
Cohort creator interface

Filtering area

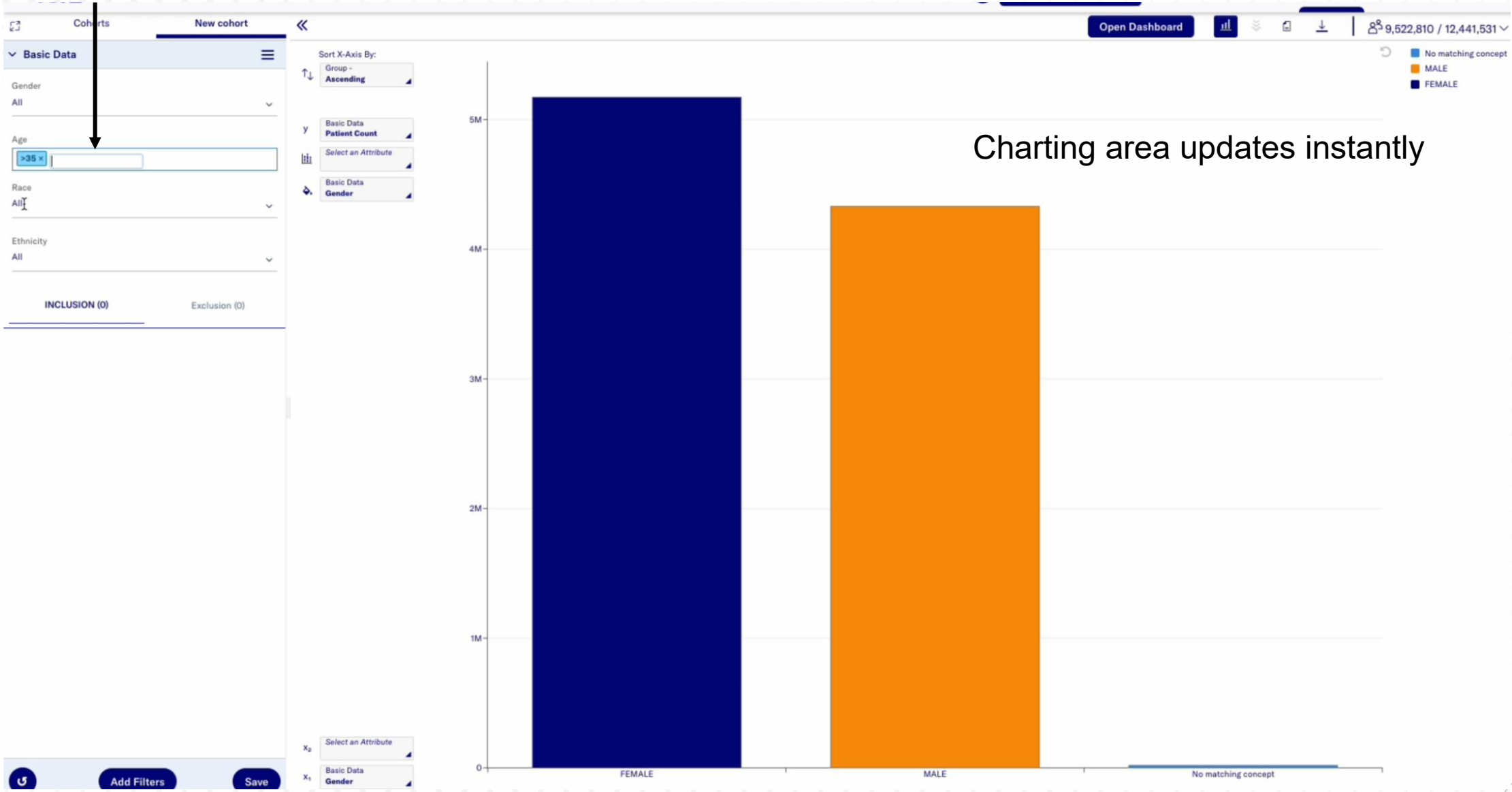


Current cohort size

Hover over bars to see counts

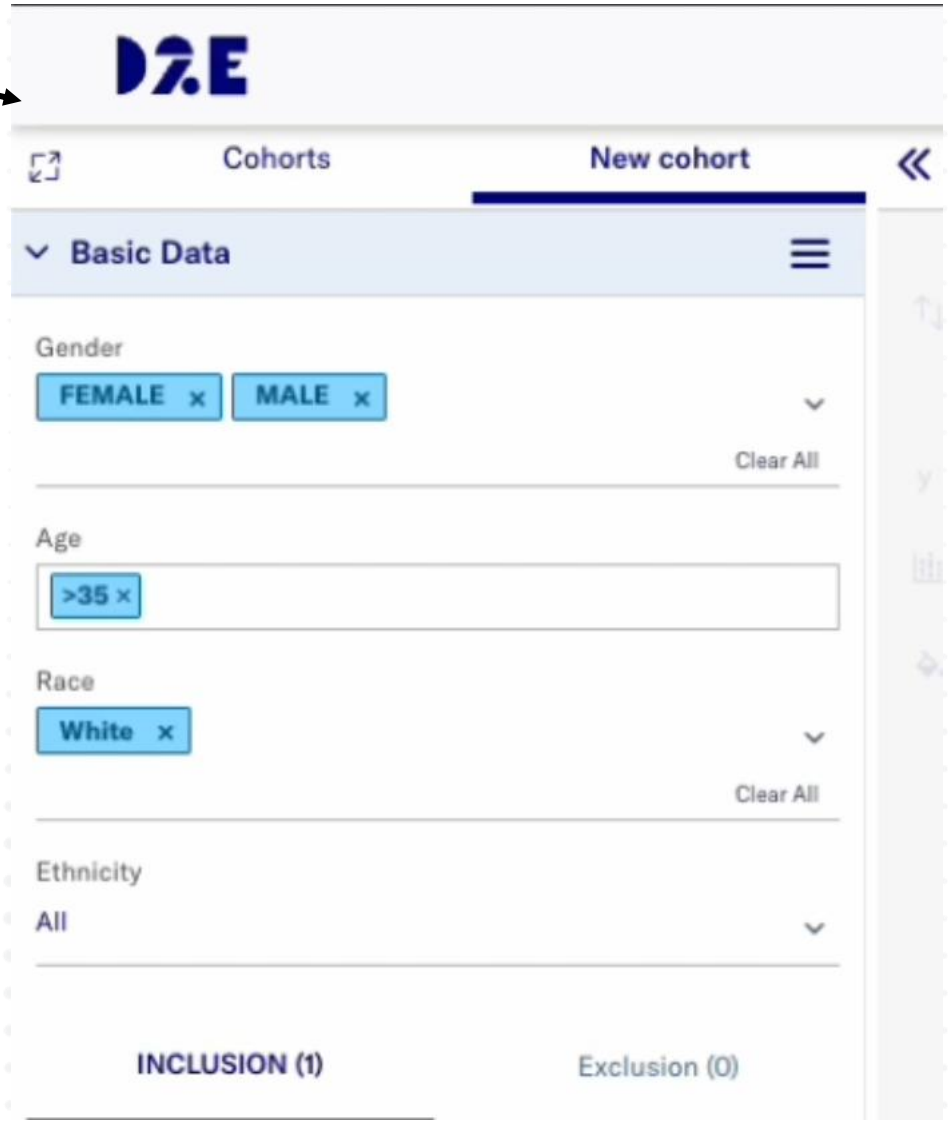


Add filter Age > 35

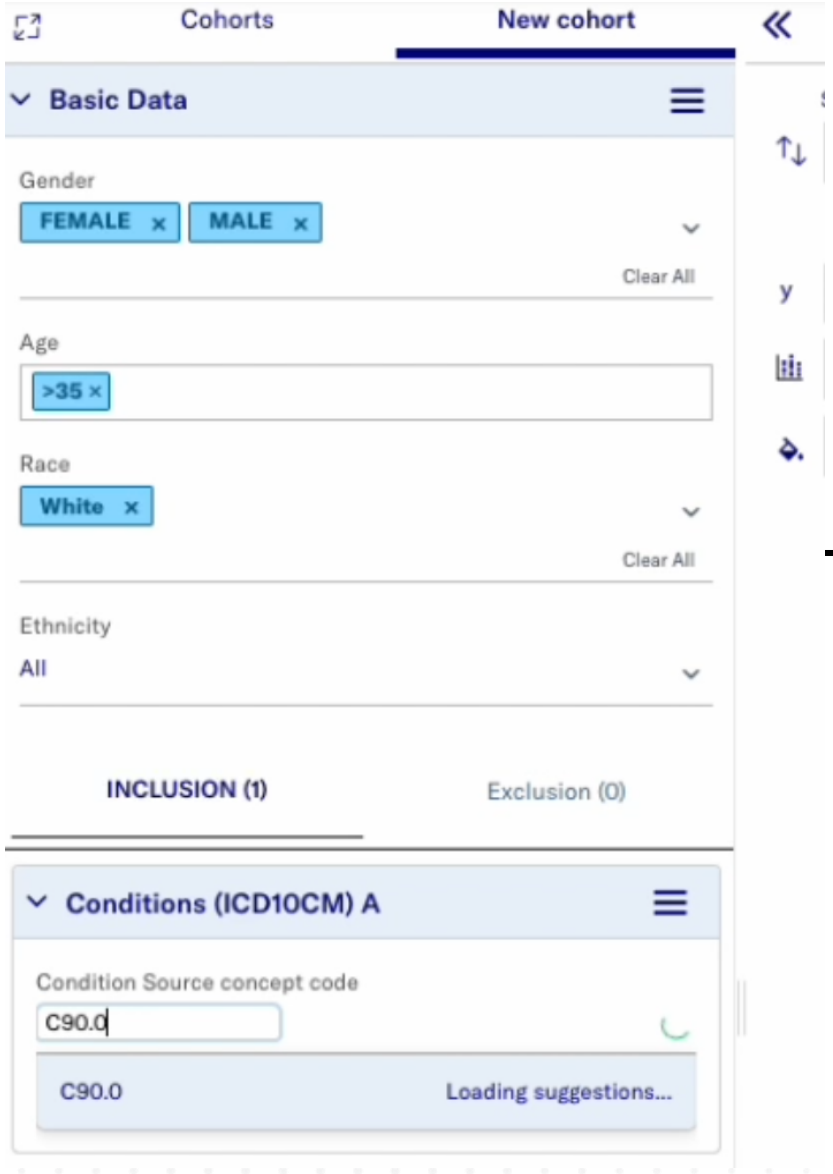


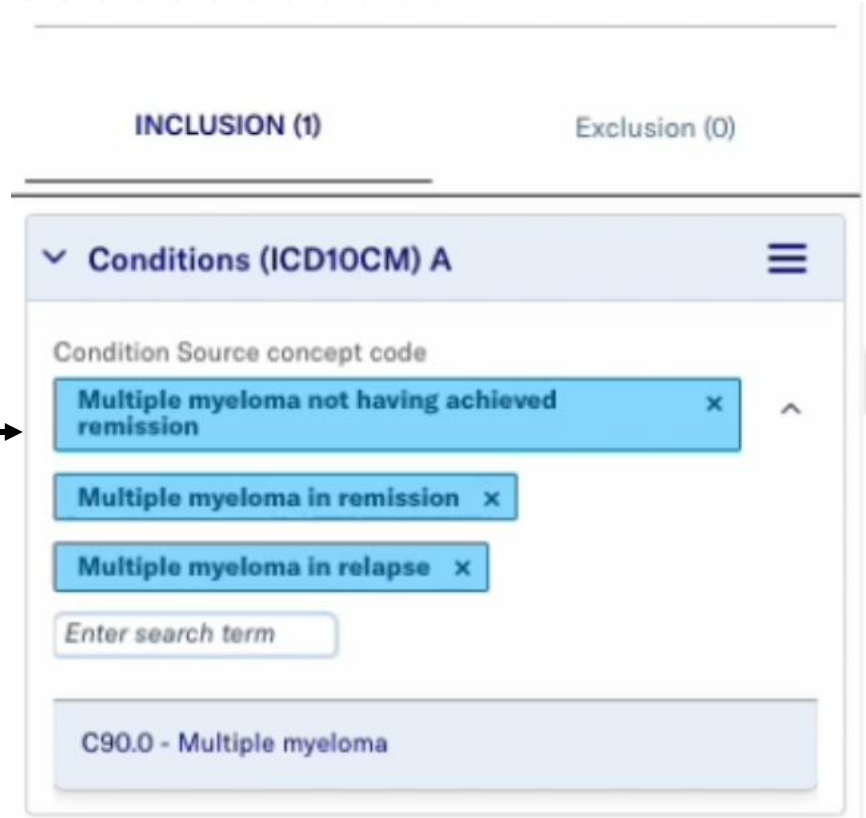
Charting area updates instantly

Add more demographic filters

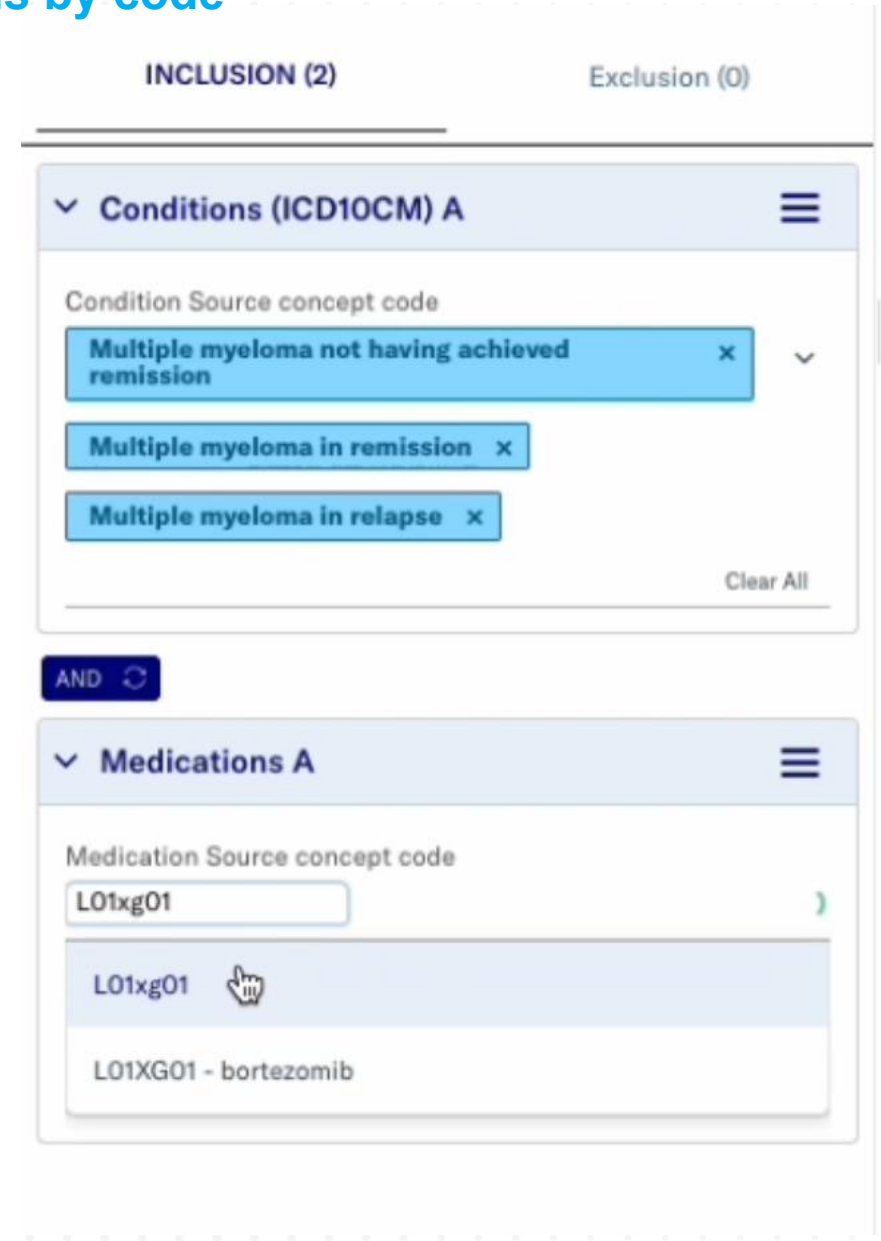


Condition filtering by ICD-10

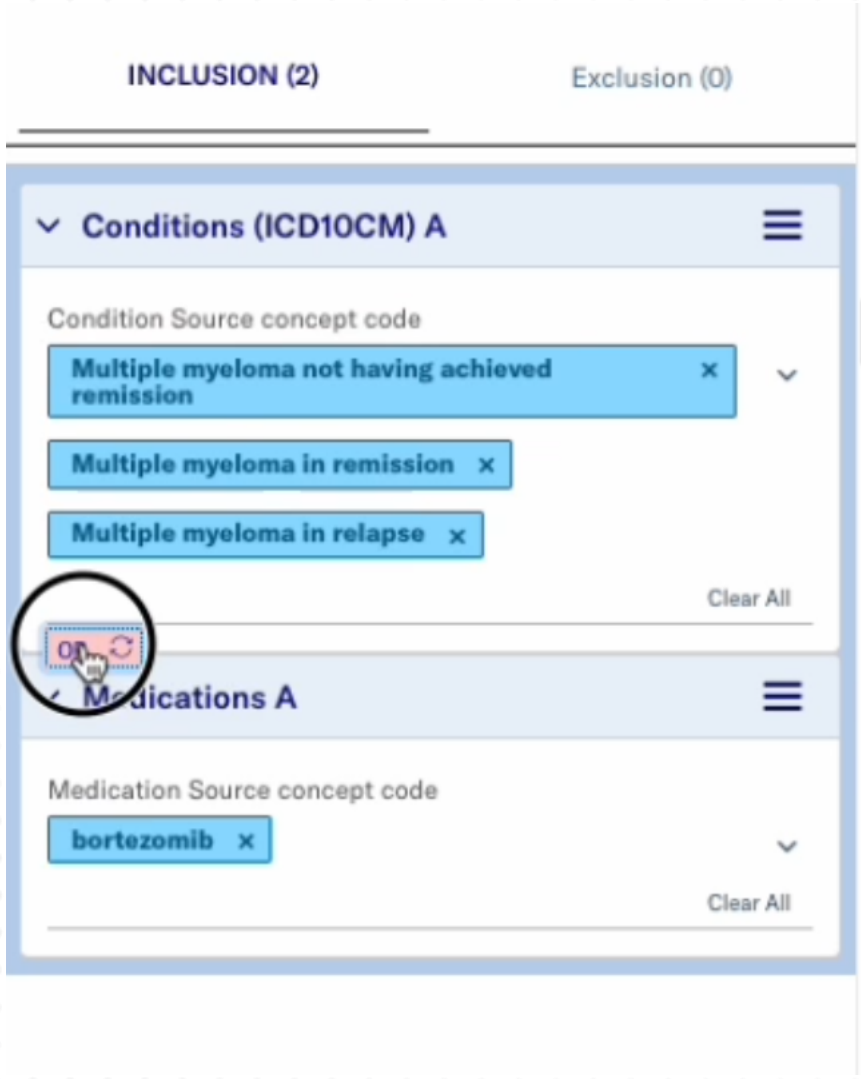




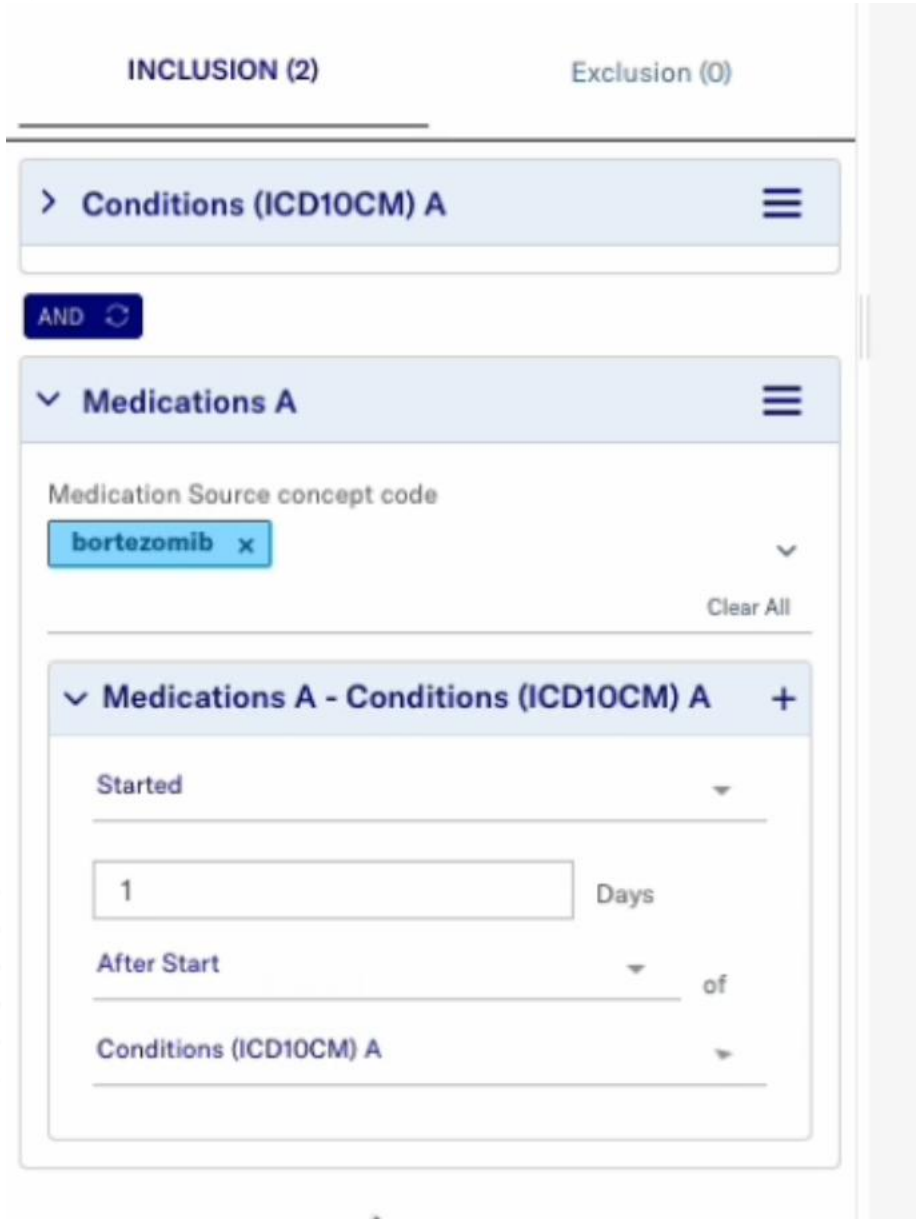
Search for medications by code



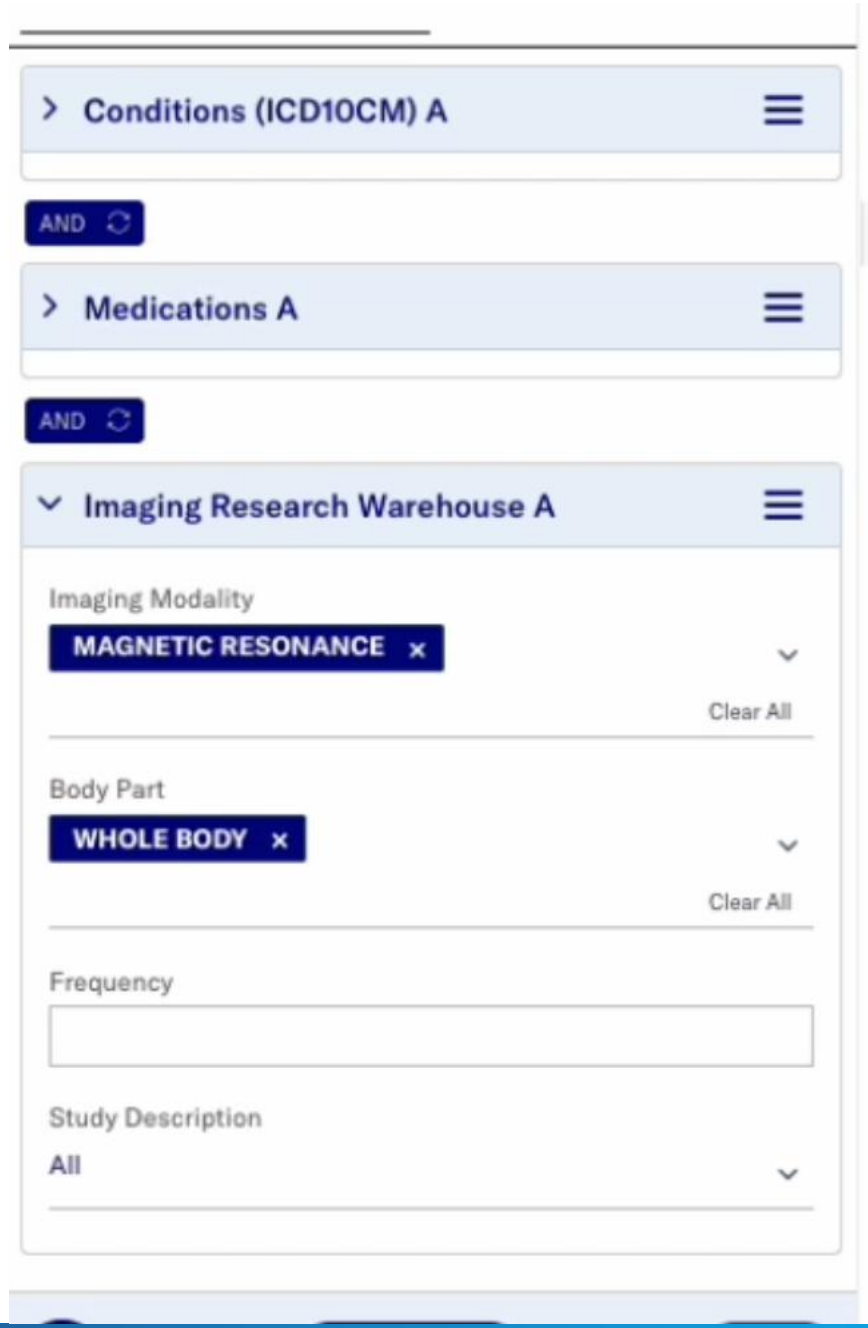
We can flip the Boolean relationship between filter cards



Add temporal relationship between filters

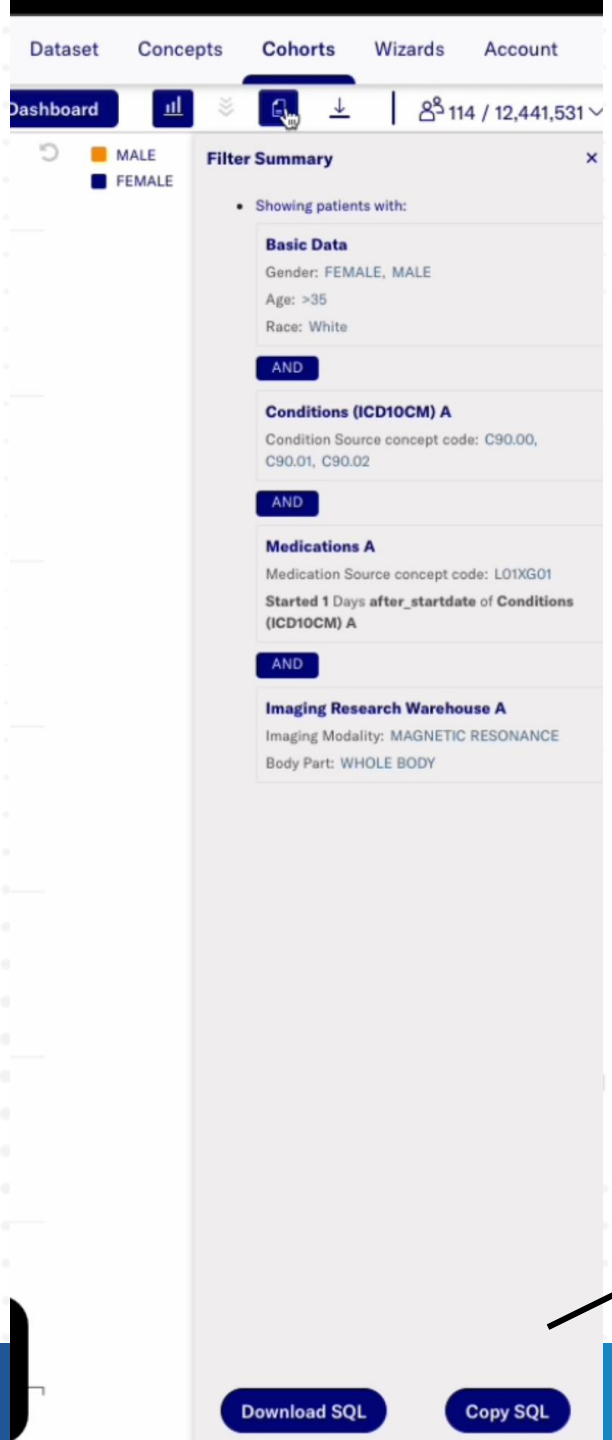


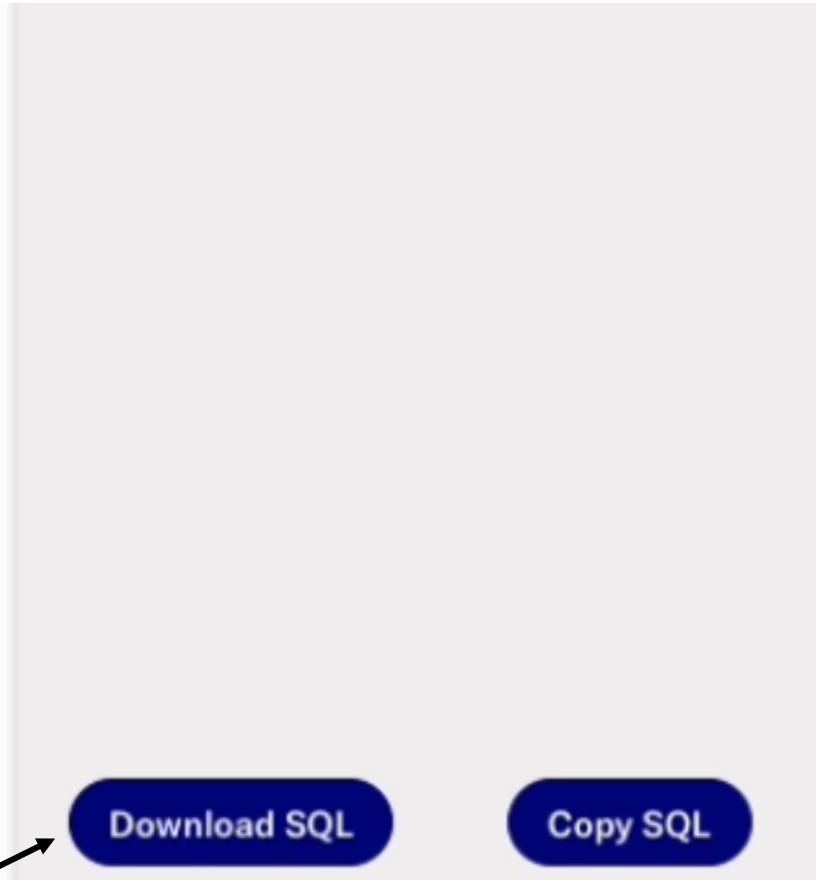
Filter by radiology studies



Download the chart

Download the SQL query

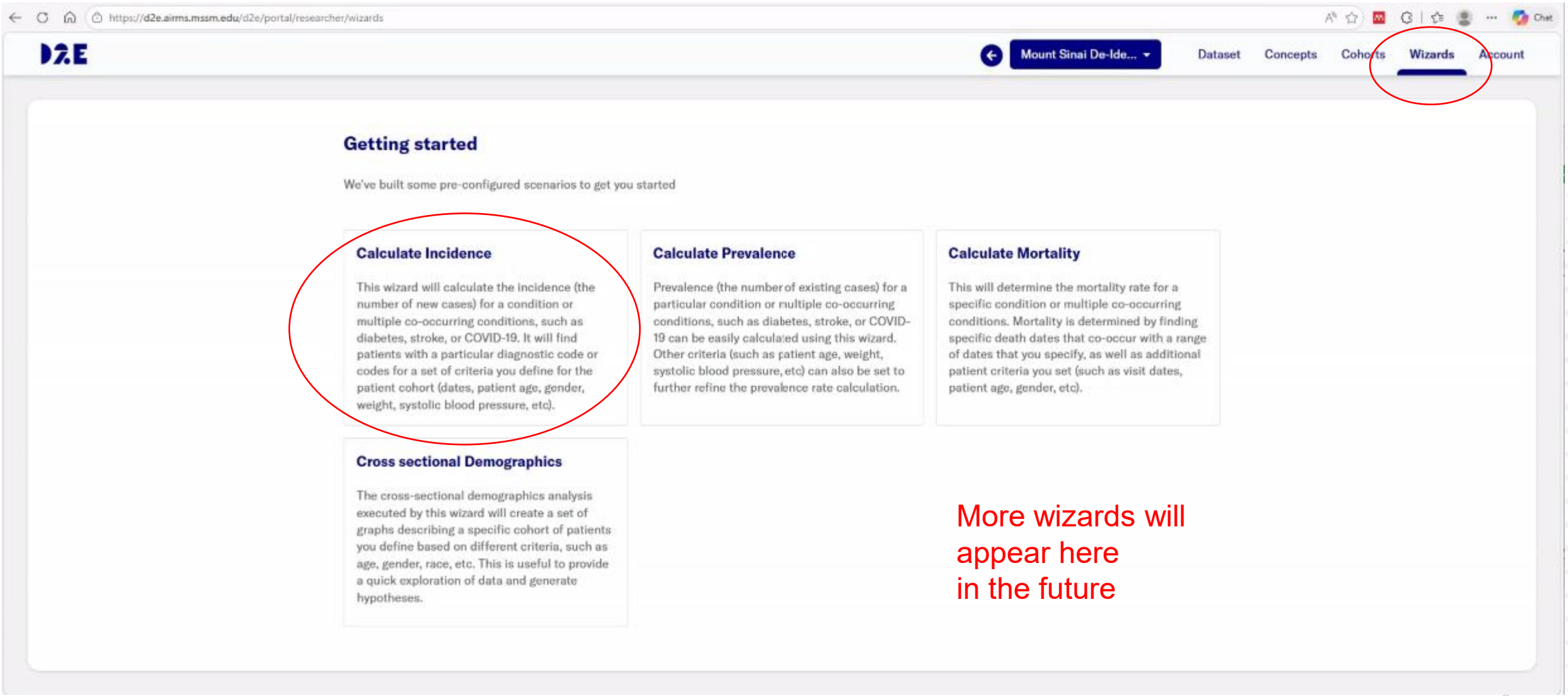




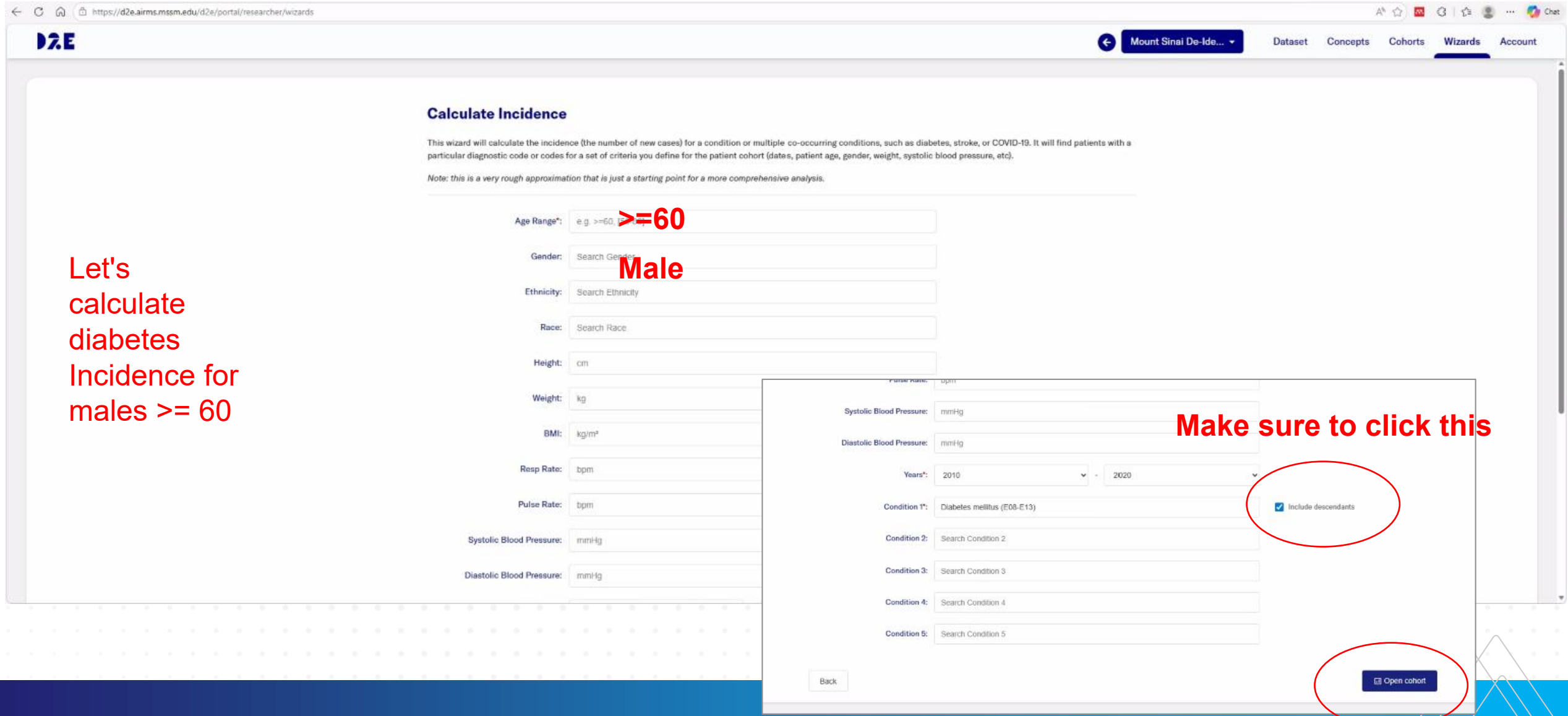
Wizards in Data2Evidence

What are Wizards?

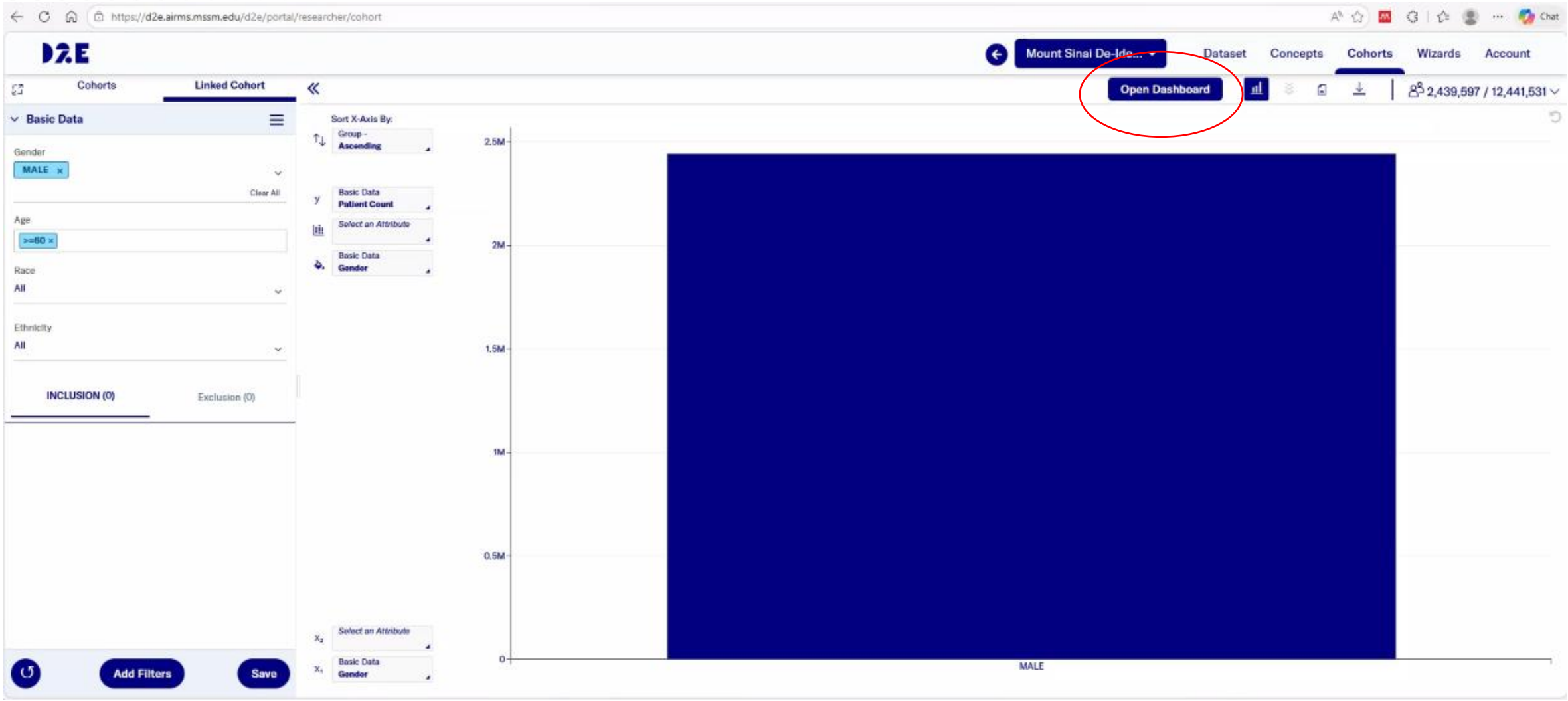
- These are pre-configured queries that can calculate common epidemiological measures, such as incidence, prevalence, mortality, etc.
- This will speed up calculations, and allow users to get quick insights with only a few mouse clicks
- Future wizards will include readmission rates, length of stay,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

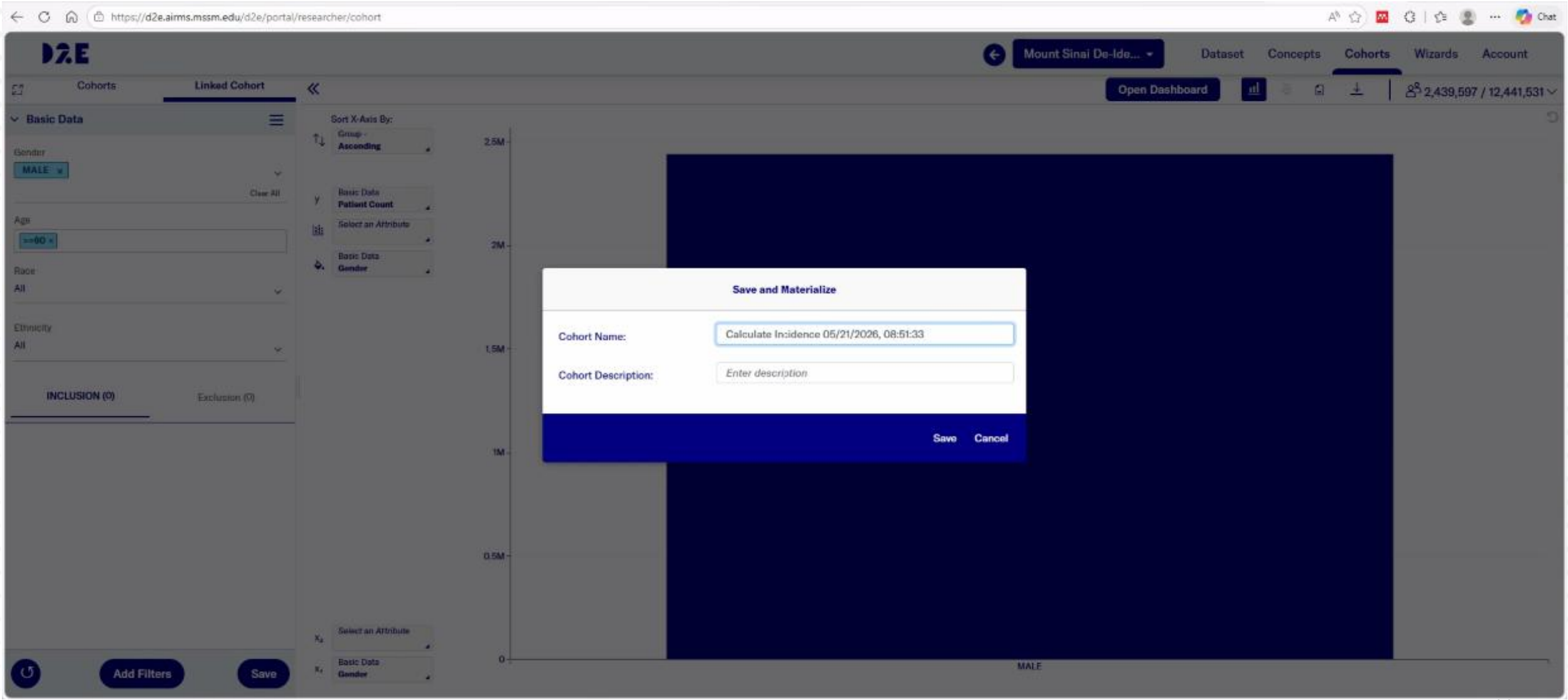

https://d2e.airms.mssm.edu/d2e/portal/researcher/wizards
D2E
Mount Sinai De-Ide...
Dataset
Concepts
Cohorts
Wizards
Account
Getting started
We've built some pre-configured scenarios to get you started
Calculate Incidence
This wizard will calculate the incidence (the number of new cases) for a condition 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

Let's calculate diabetes Incidence for males >= 60



Make sure to click this


https://d2e.airms.mssm.edu/d2e/portal/researcher/cohort
D2E
Mount Sinai De-Ide...
Dataset
Concepts
Cohorts
Wizards
Account
Cohorts
Linked Cohort
Open Dashboard
2,439,597 / 12,441,531
Basic Data
Gender
MALE
Clear All
Age
>=60
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
Select an Attribute
Basic Data Gender
Add Filters
Save
2.5M
2M
1.5M
1M
0.5M
0
MALE


https://d2e.airms.mssm.edu/d2e/portal/researcher/cohort
D2E
Mount Sinai De-Ide...
Dataset
Concepts
Cohorts
Wizards
Account
Cohorts
Linked Cohort
Open Dashboard
2,439,597 / 12,441,531
Basic Data
Gender
MALE
Clear All
Age
>=60
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
2.5M
2M
1.5M
1M
0.5M
0
MALE
Add Filters
Save

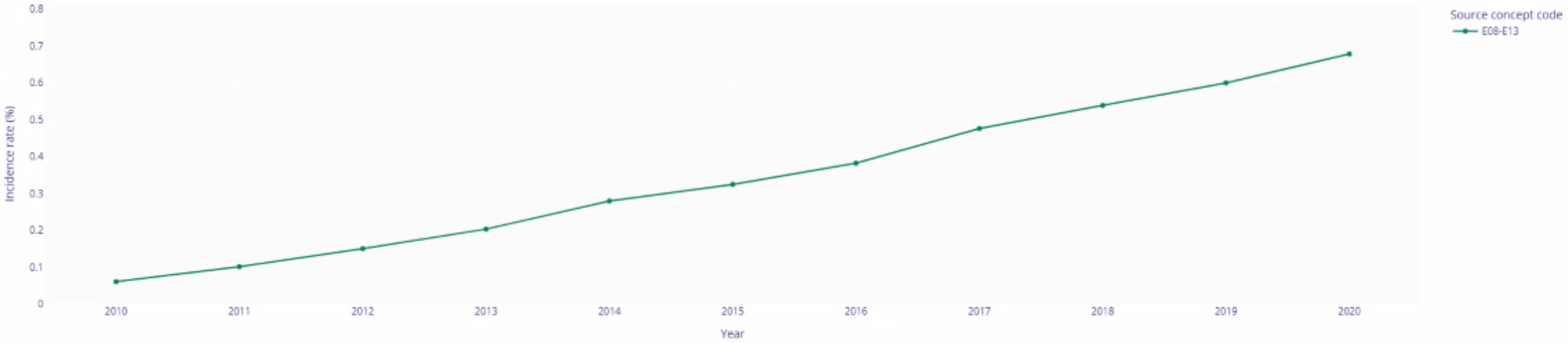
https://d2e.airms.mssm.edu/d2e/portal/researcher/cohort

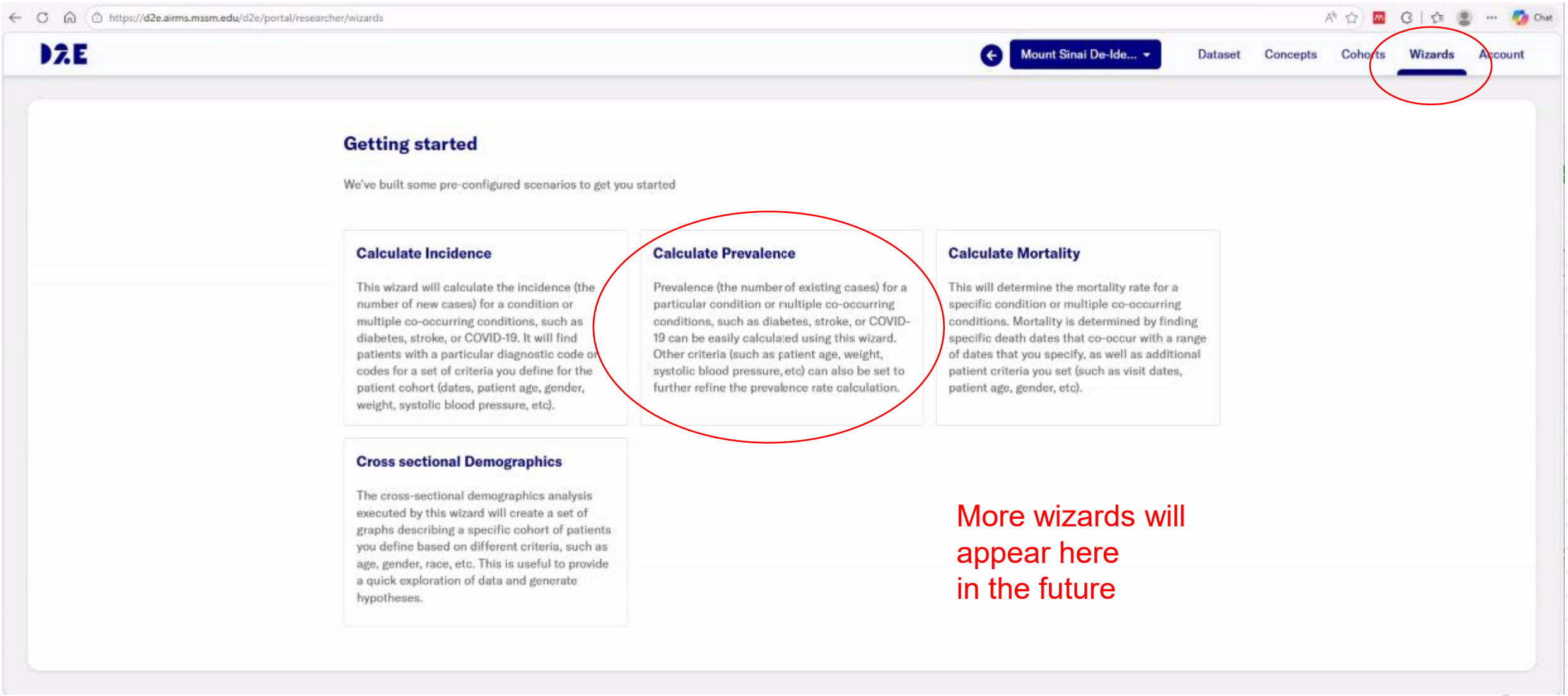
Chat

Dashboard

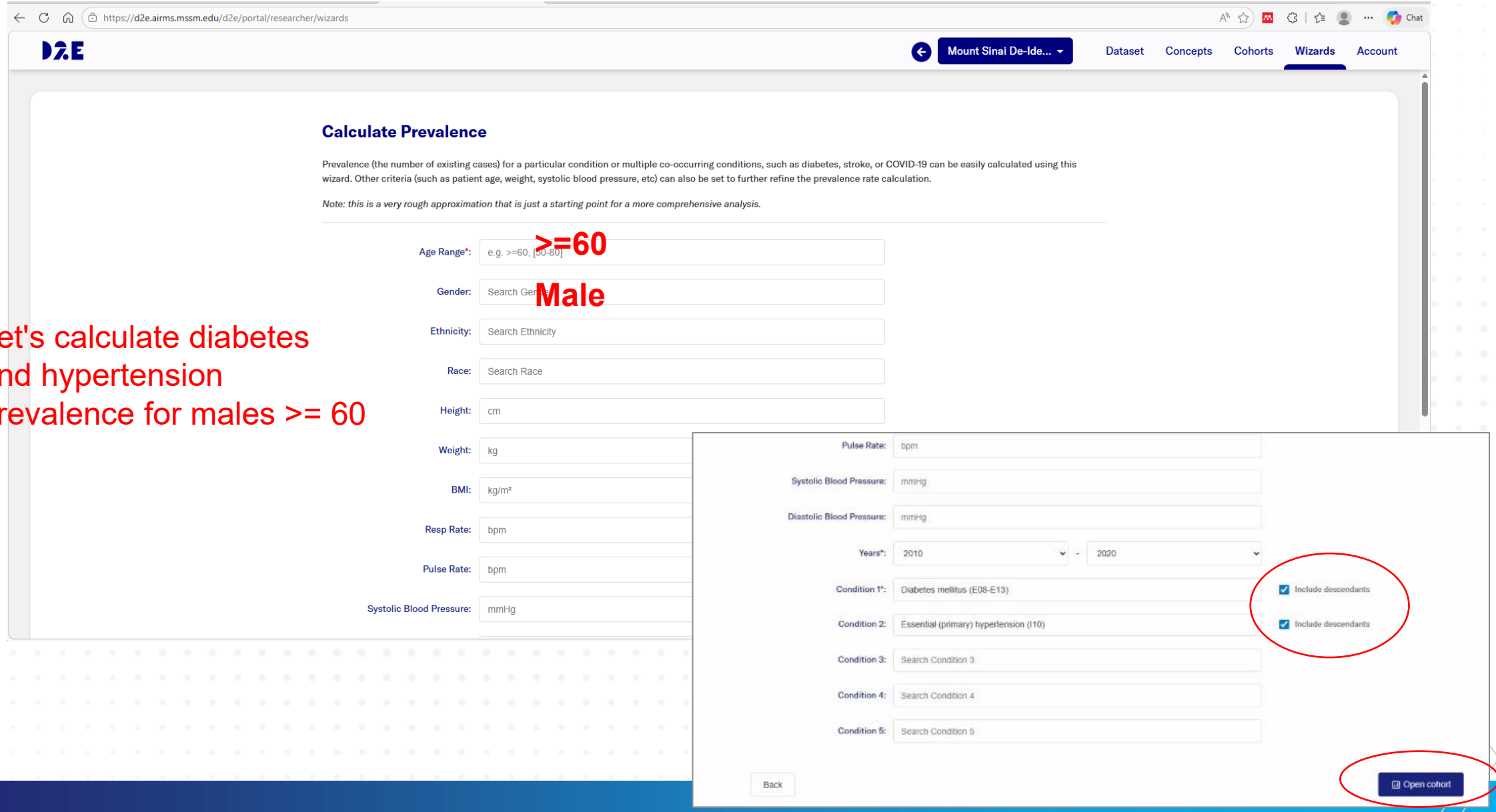
Incidence rate dashboard

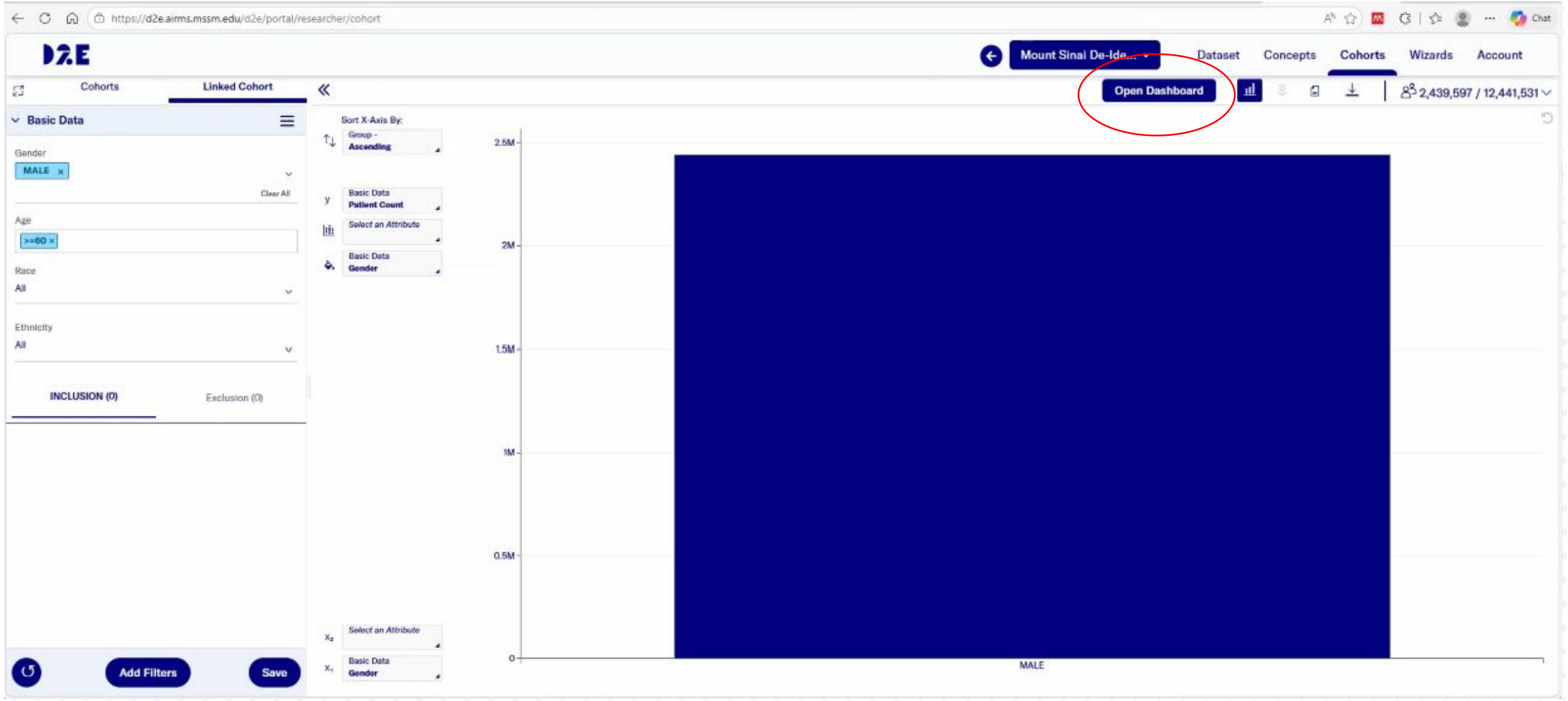
Calculated as the percentage of the cohort that had these conditions as an incidence during the years 2010 to 2020 based on condition start date, and did not have the conditions a year prior.




https://d2e.airms.mssm.edu/d2e/portal/researcher/wizards
D2E
Mount Sinai De-Ide...
Dataset
Concepts
Cohorts
Wizards
Account
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

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




https://d2e.aimms.mssm.edu/d2e/portal/researcher/cohort
D2E
Mount Sinai De-Ide...
Dataset
Concepts
Cohorts
Wizards
Account
Cohorts
Linked Cohort
Open Dashboard
2,439,597 / 12,441,531
Basic Data
Gender
MALE
Clear All
Age
>=60
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
Select an Attribute
Basic Data Gender
Add Filters
Save
2.5M
2M
1.5M
1M
0.5M
0
MALE

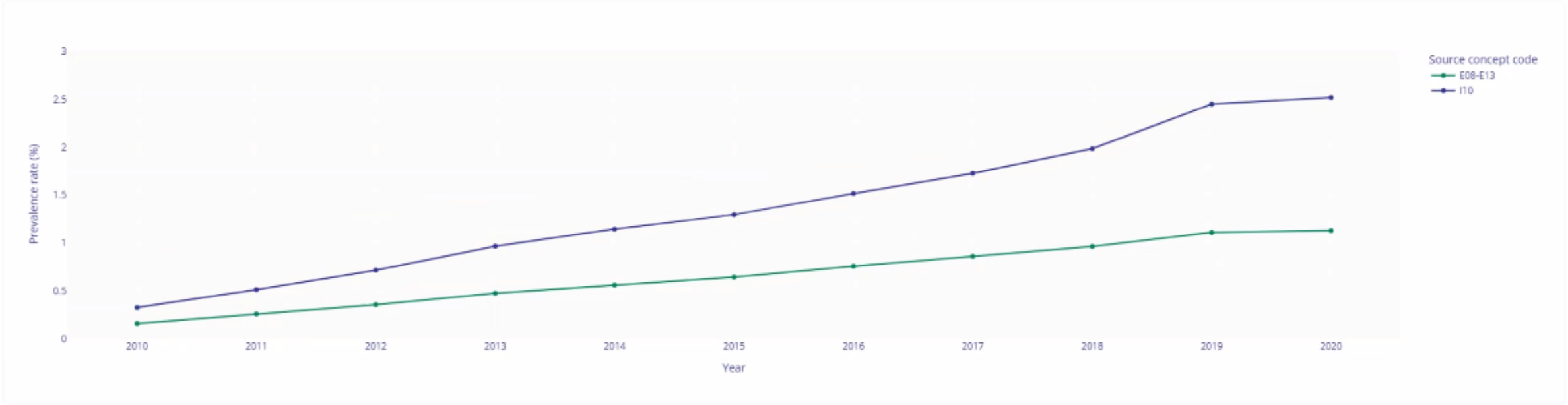
https://d2e.aiims.mssm.edu/d2e/portal/researcher/cohort

Chat

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!