Windreich Department of AI & Human Health

Introduction to Data2Evidence

Ashwin Sawant, MD

May 12, 2026

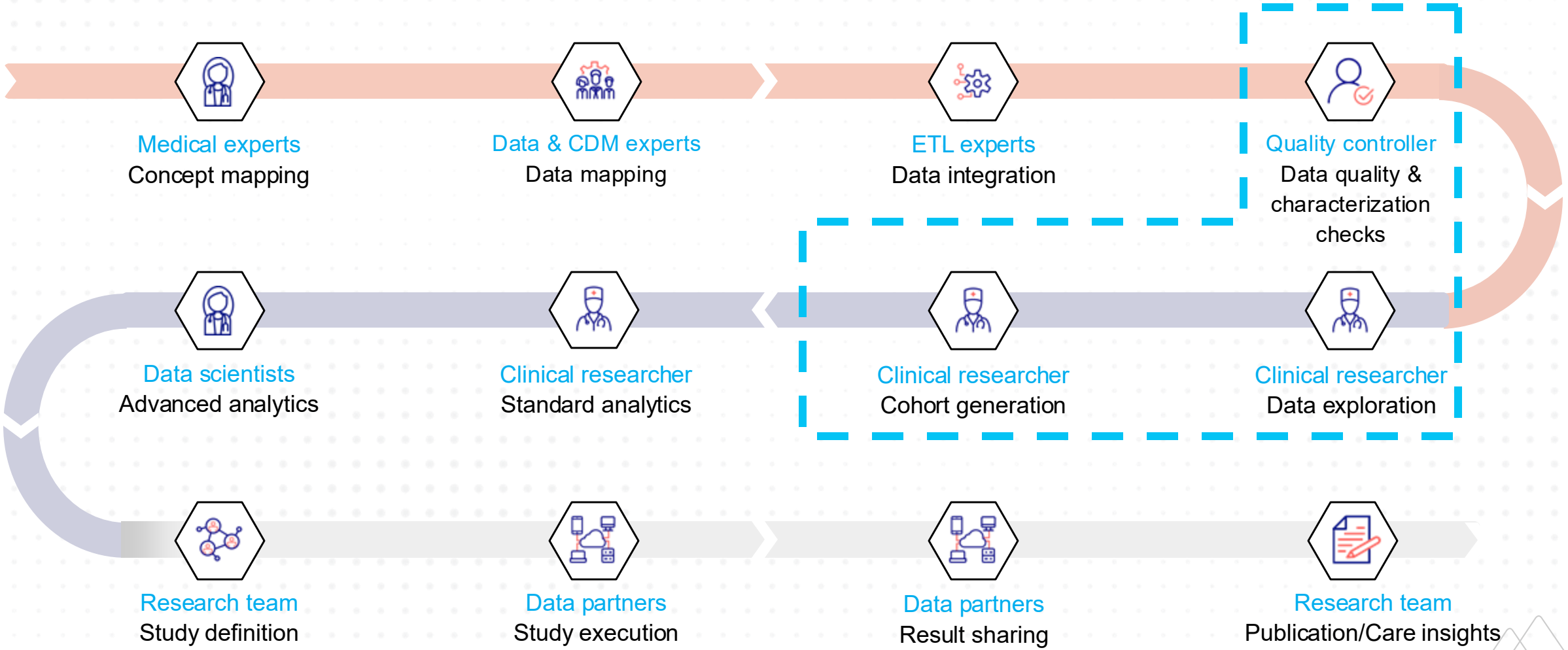
HPI·MS

Mount Sinai

Hasso Plattner Institute for Digital Health at Mount Sinai

DATA2EVIDENCE

Utilization along the research lifecycle



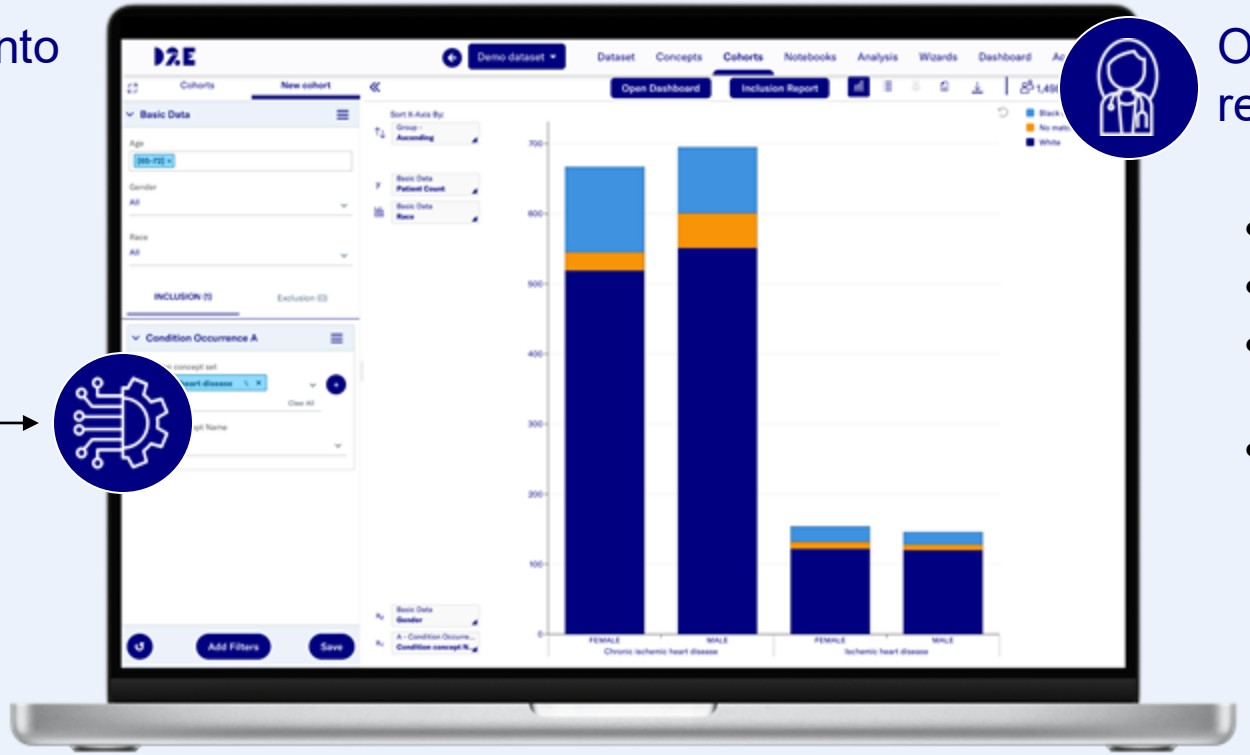
Slide credit: Pablo Guerrero, Data4Life

DATA2EVIDENCE

Making health data readily accessible for researchers

Extract and combine data into interoperable dataset

- Electronic health records
- Lab records
- Genomics data
- Image (meta-)data
- Unstructured data (notes)
- Wearable data (PROs)

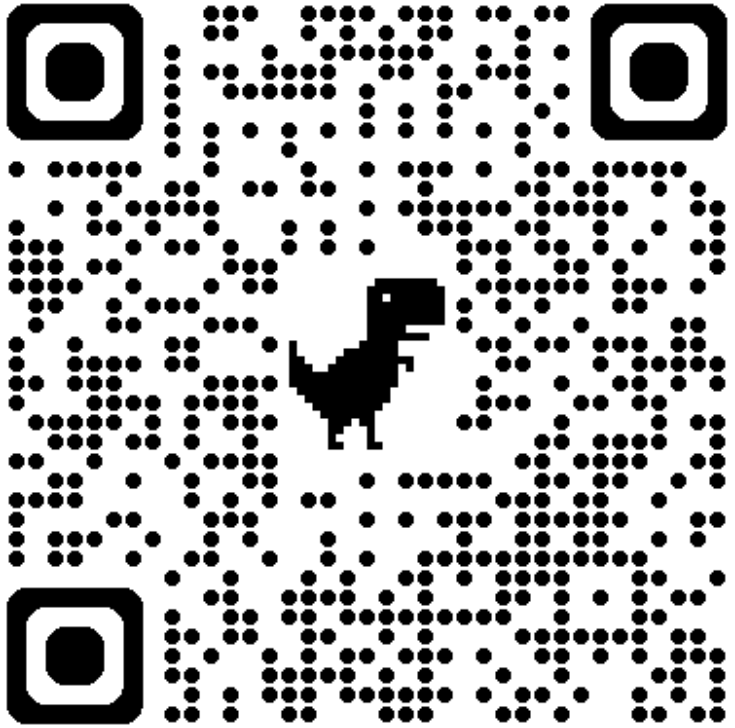


Offer a one-stop platform to researchers

- Existing dataset exploration
- Customized cohort creation
- Data quality checks for datasets and cohorts
- Collaborative analysis with other researchers

Slide credit: Pablo Guerrero, Data4Life

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



d2e.airms.mssm.edu

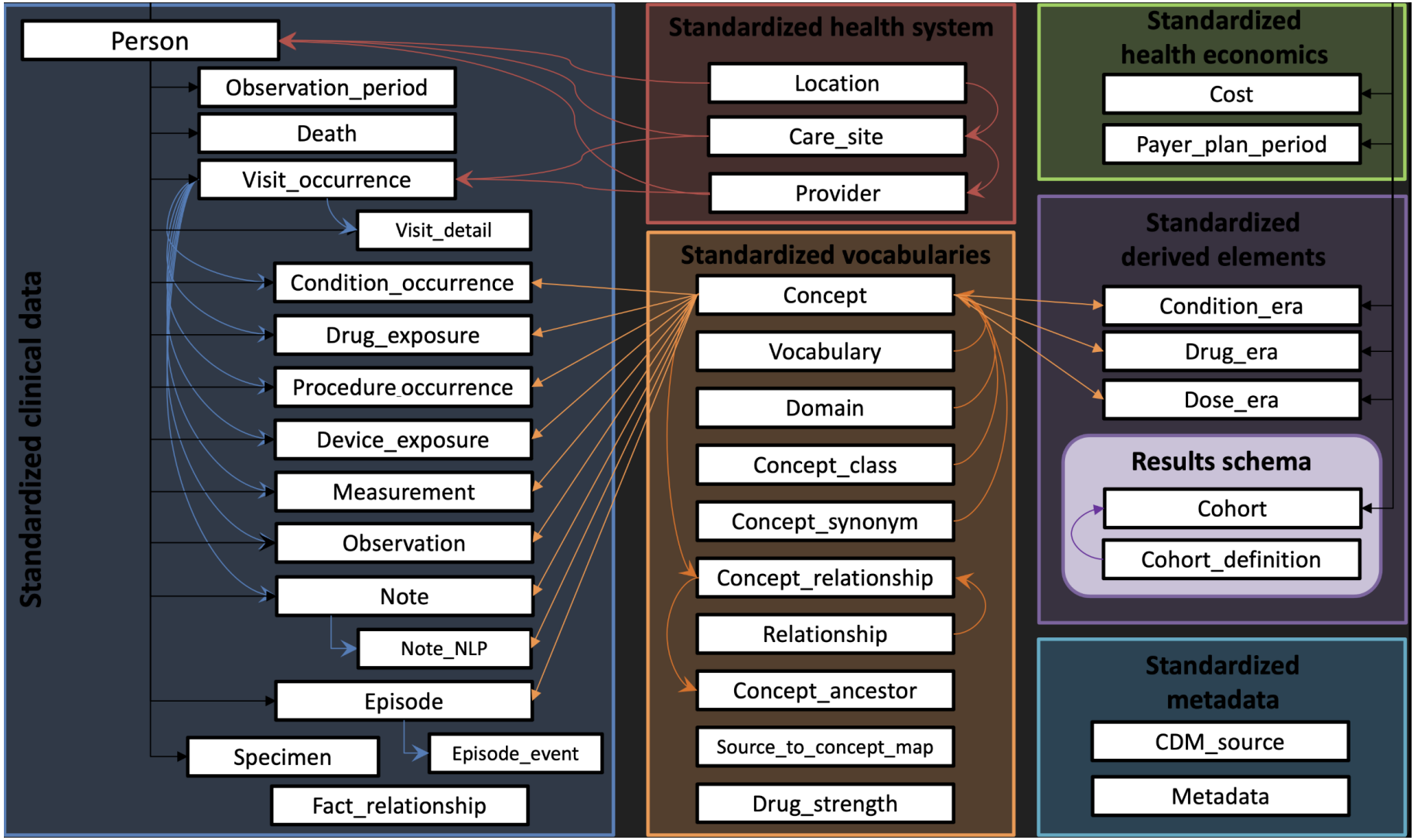
The Data Model

Data elements

Constraints on elements – standard vocabularies

Relationship between elements

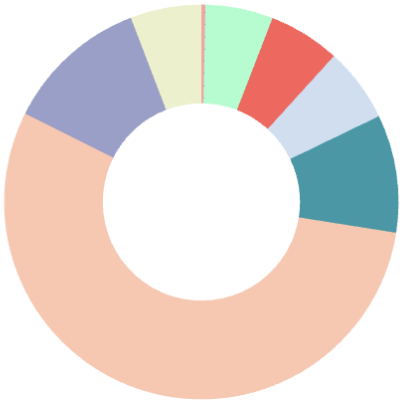
The Data Model



The Data Model – Domains

Domain	Standard vocabulary	Source vocabularies
Condition	SNOMED	ICD-9, ICD-10
Drug	RxNorm	NDC
Procedure	SNOMED	CPT
Measurement	SNOMED	LOINC
Observation	SNOMED	Various

Datasets



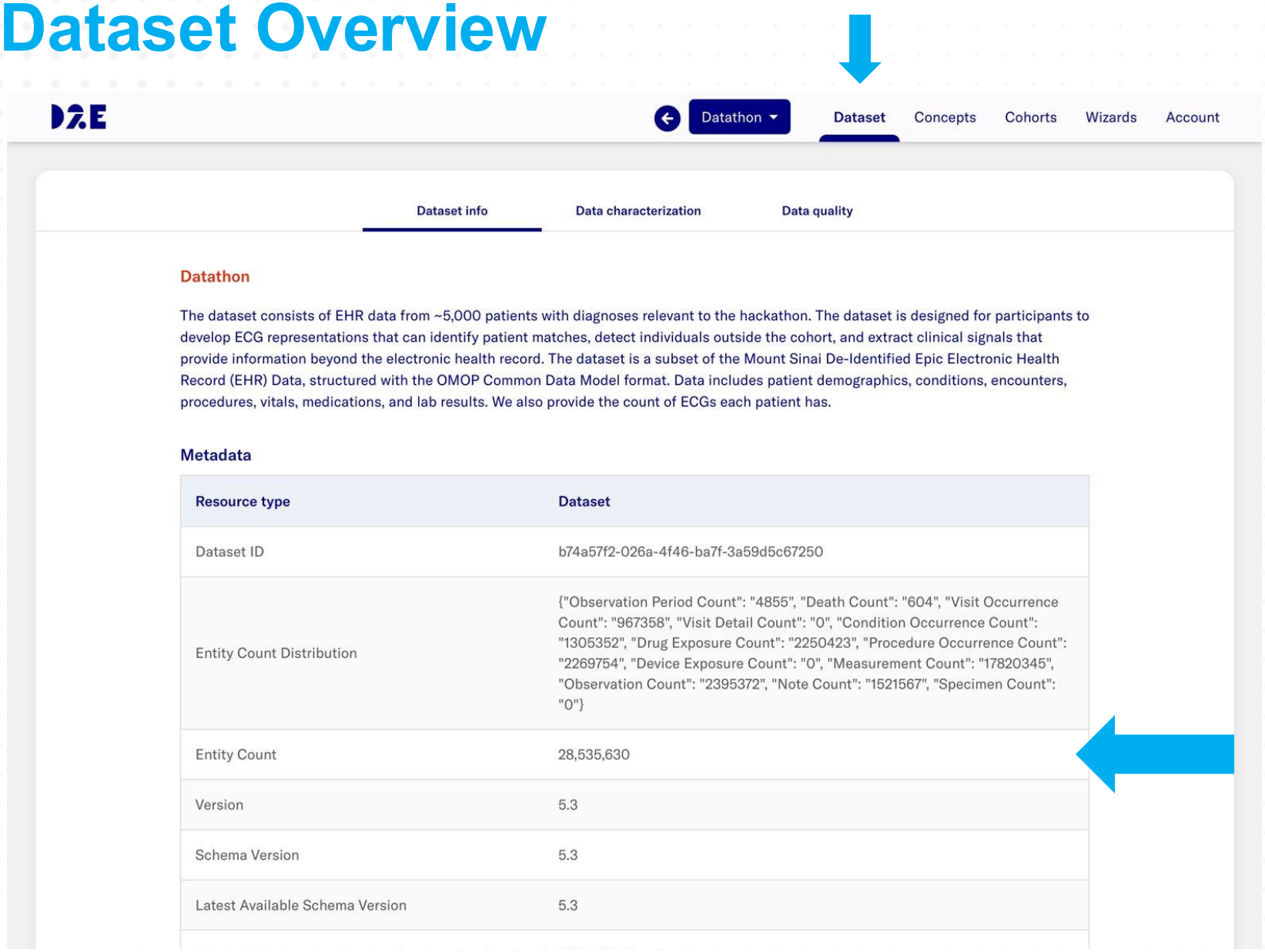
Mount Sinai De-Identified Epic Electronic Health Record (EHR) Data

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, Data4Life

Dataset Characterization

Dataset info | Data characterization | Data quality

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

Select data characterization report

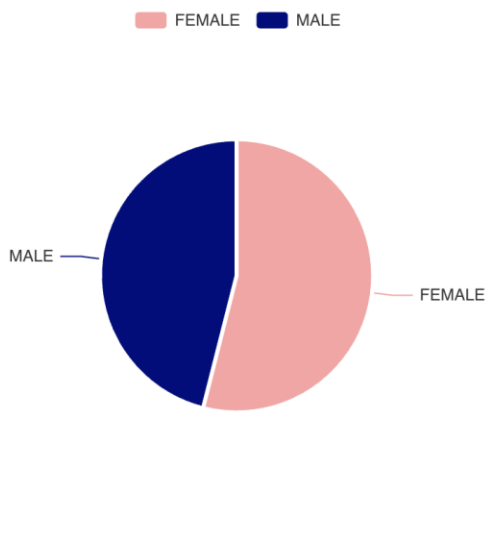
Show all reports

CDM Summary

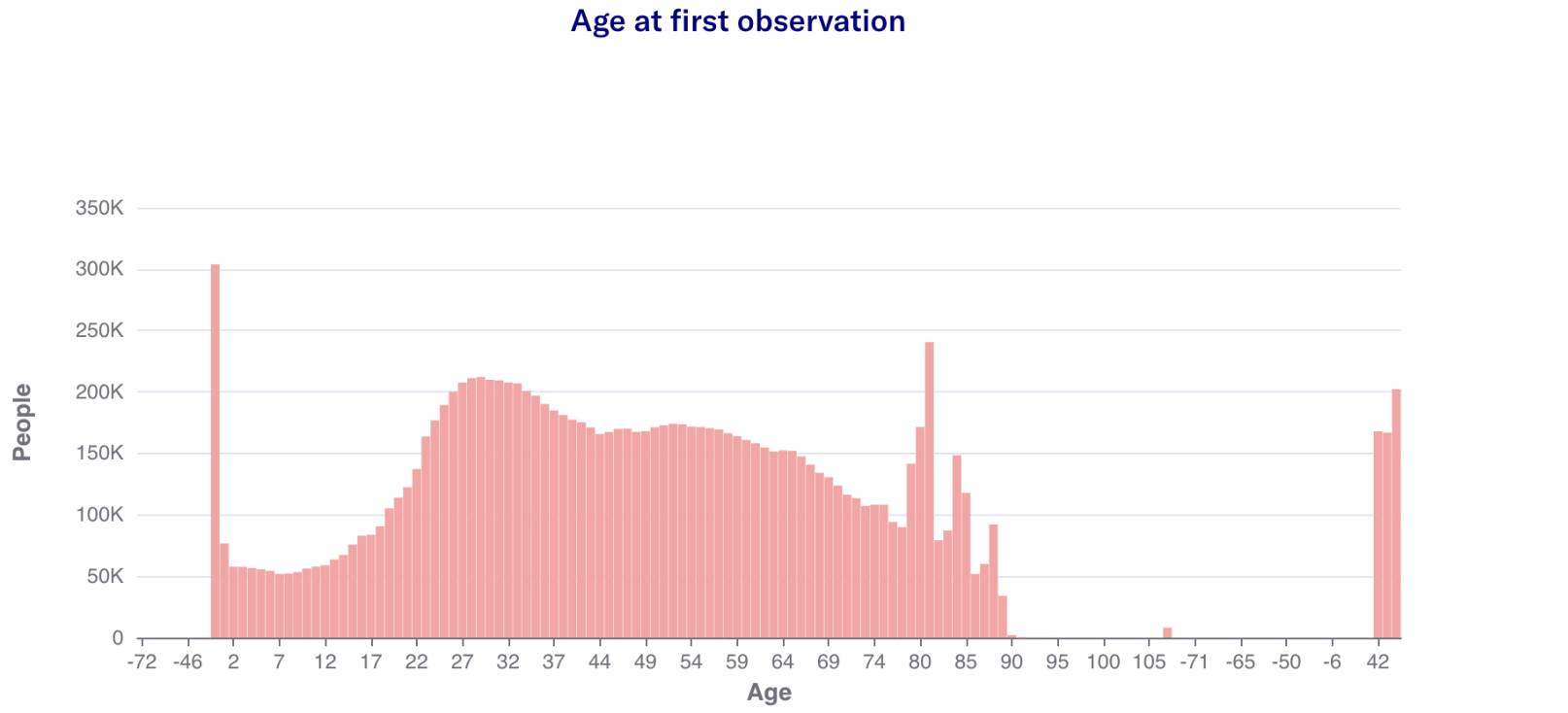
Source name: CDMDEID

Number of persons: 12,441,531

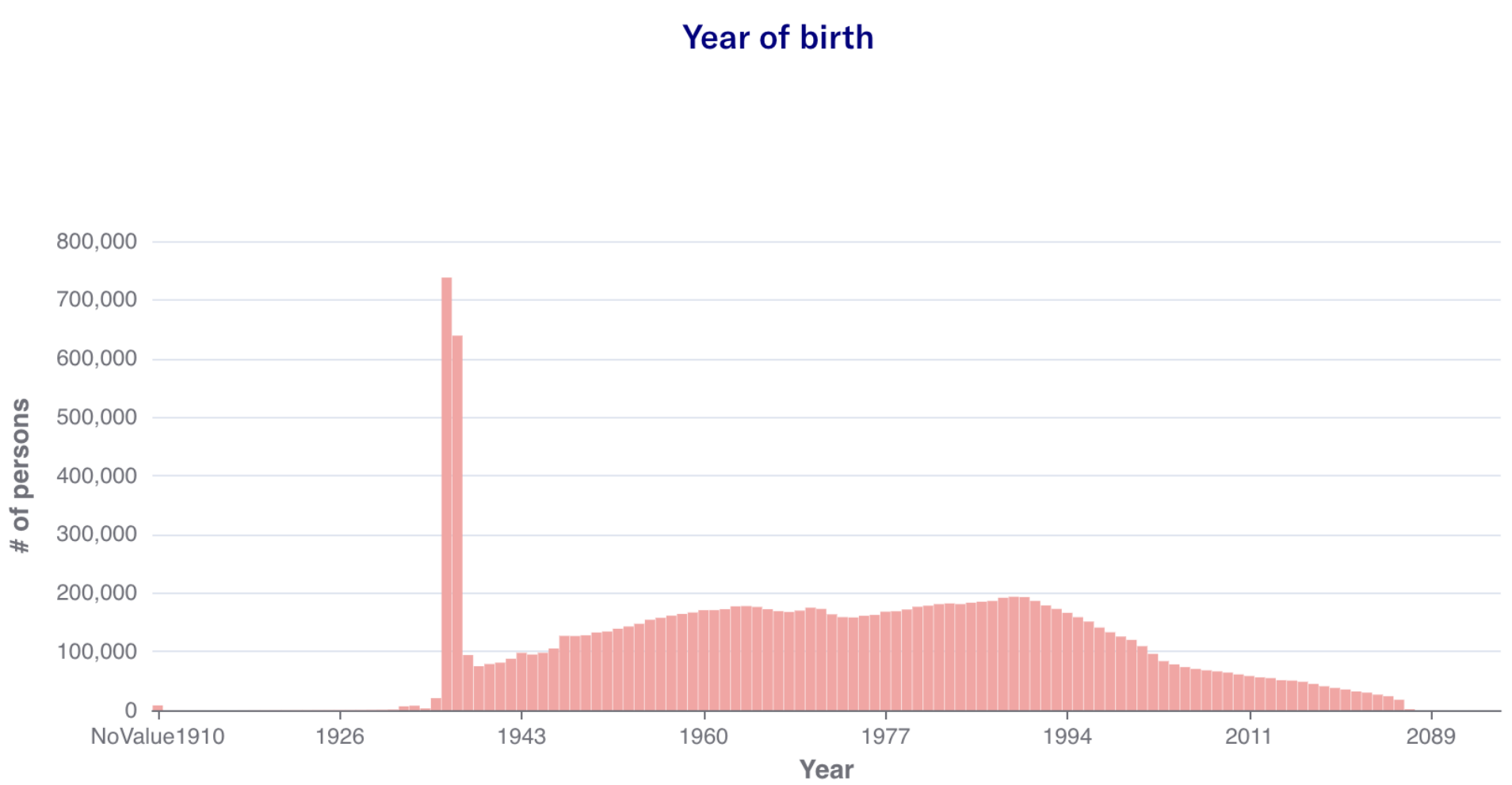
Gender



Dataset Characterization



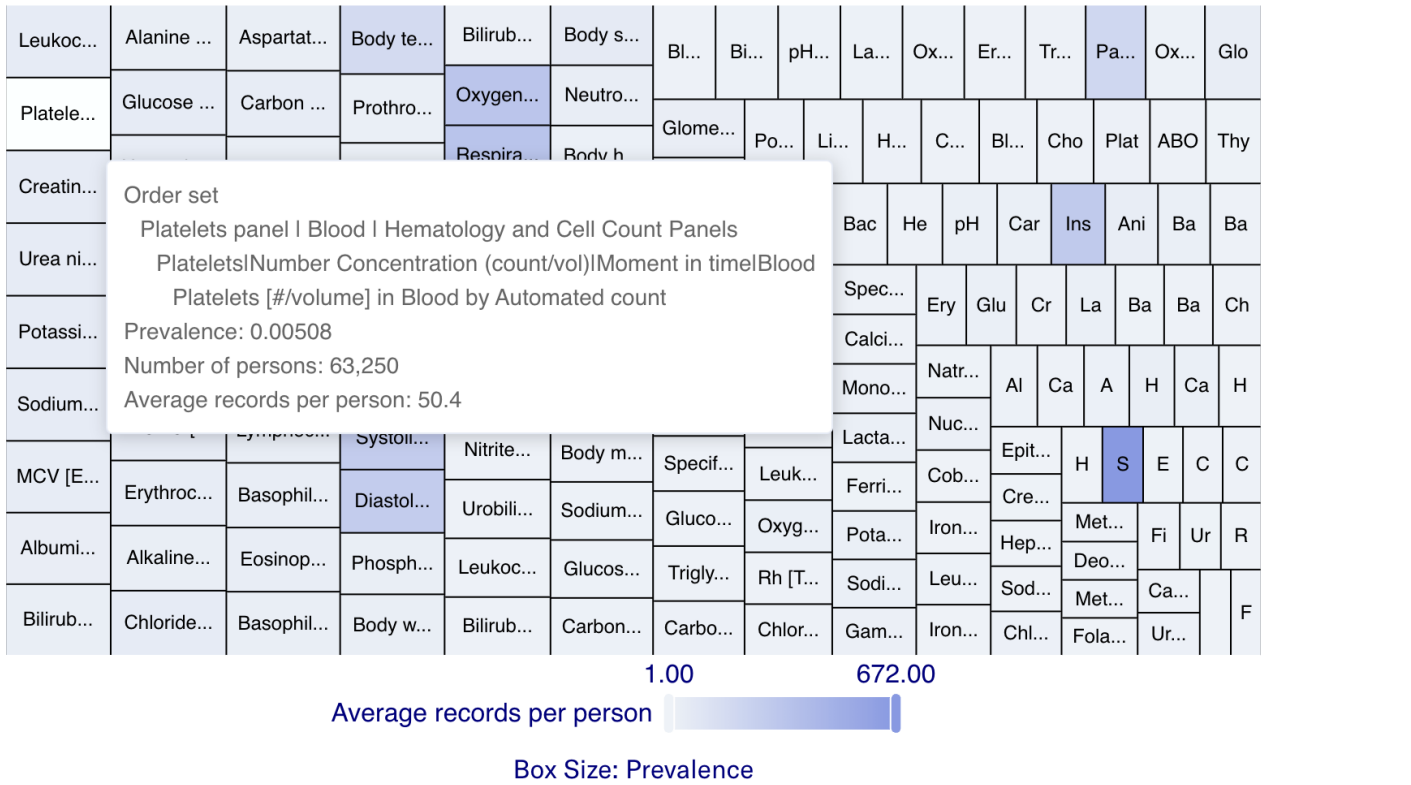
Dataset Characterization



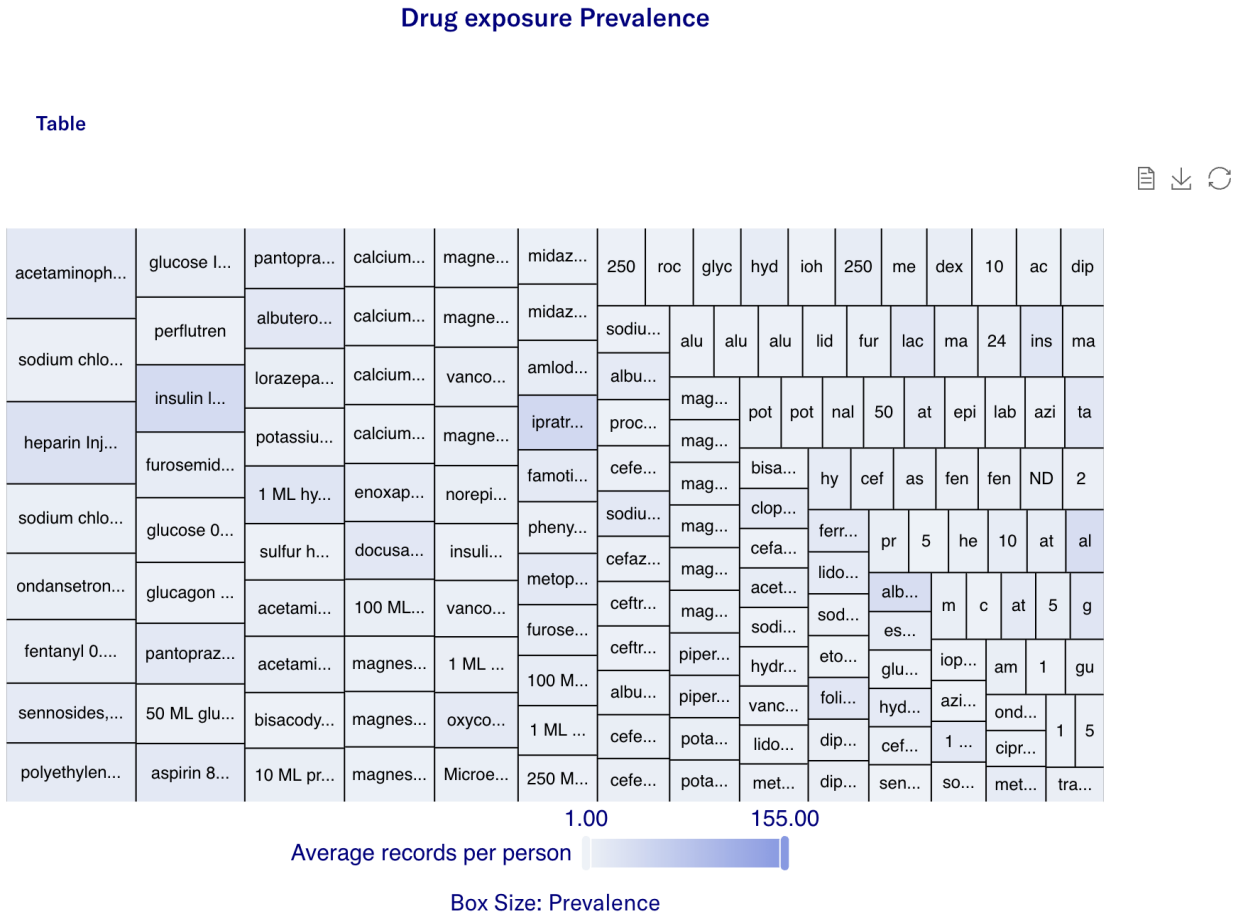
Dataset Characterization

Measurement Prevalence

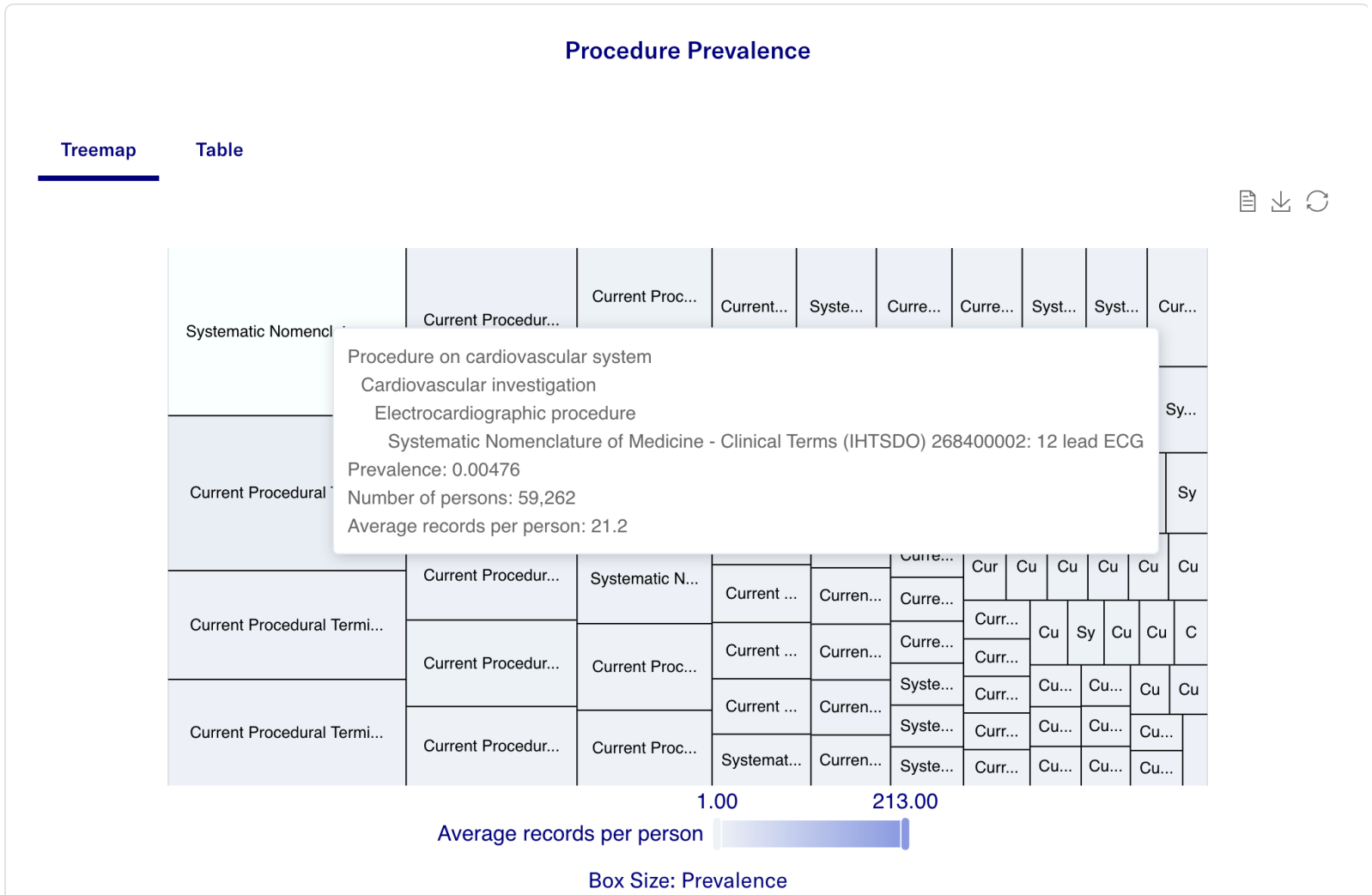
Treemap | Table



Dataset Characterization



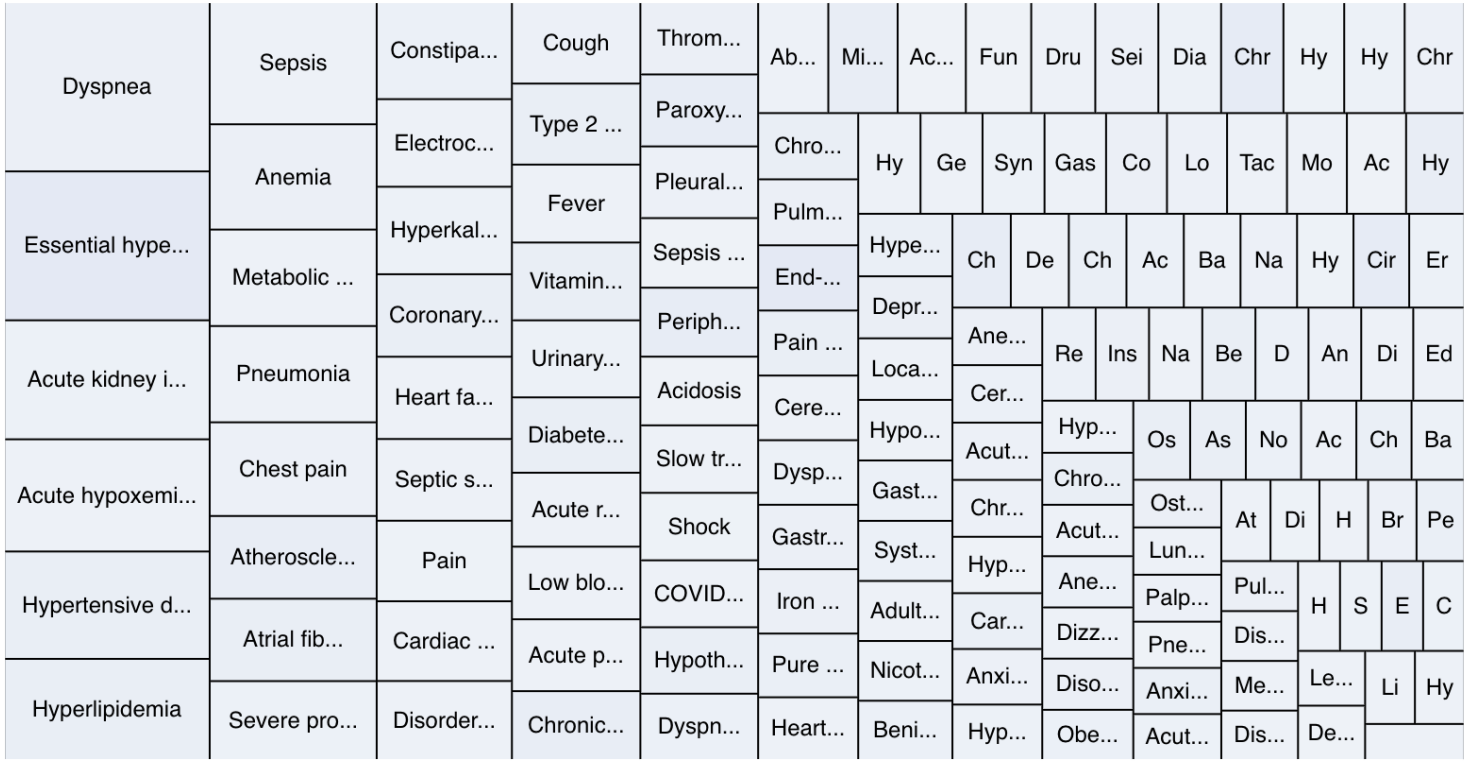
Dataset Characterization



Dataset Characterization

Condition occurrence 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%

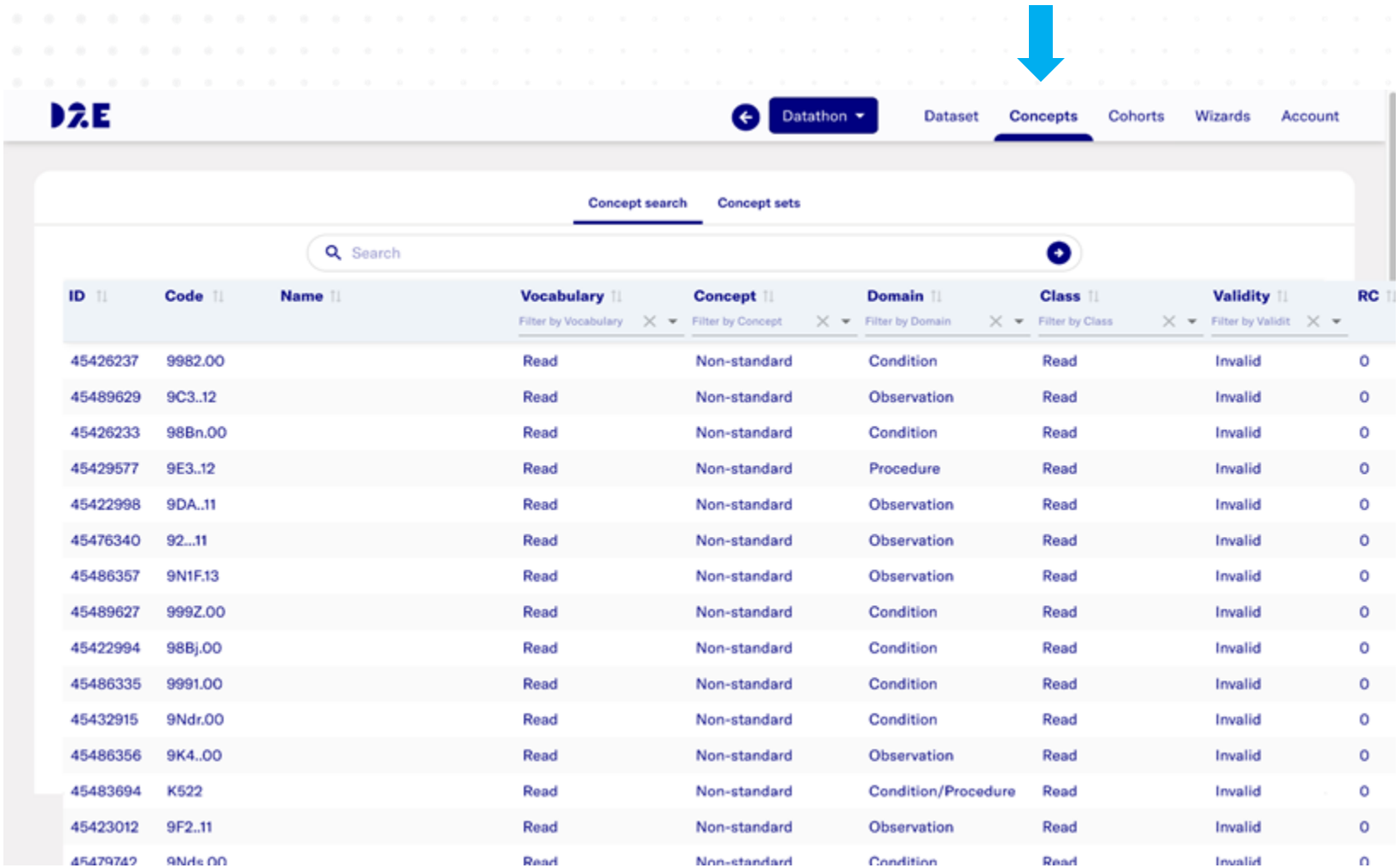
570 out of 1,739 passed checks are not applicable, due to empty tables or fields.

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

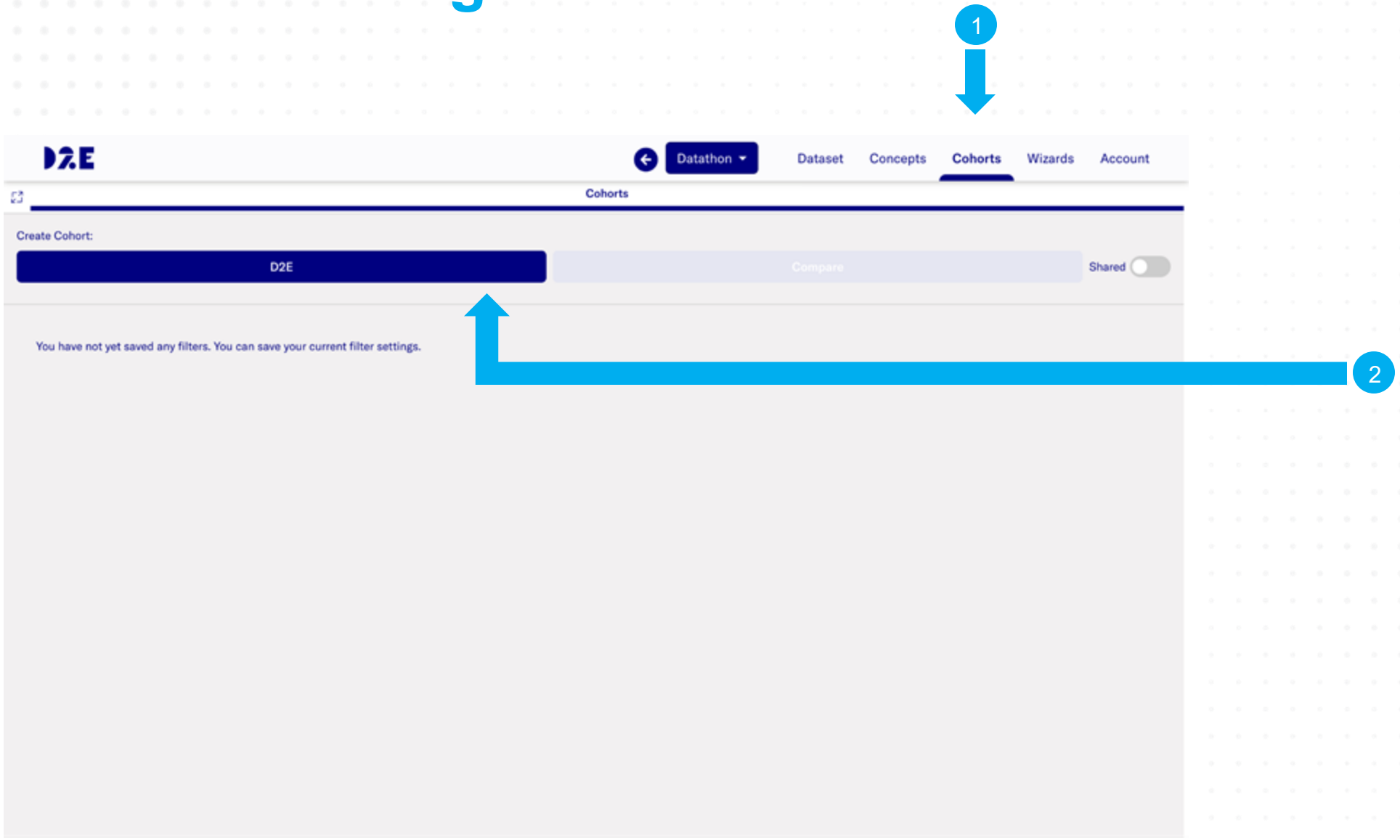
Concept Management



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

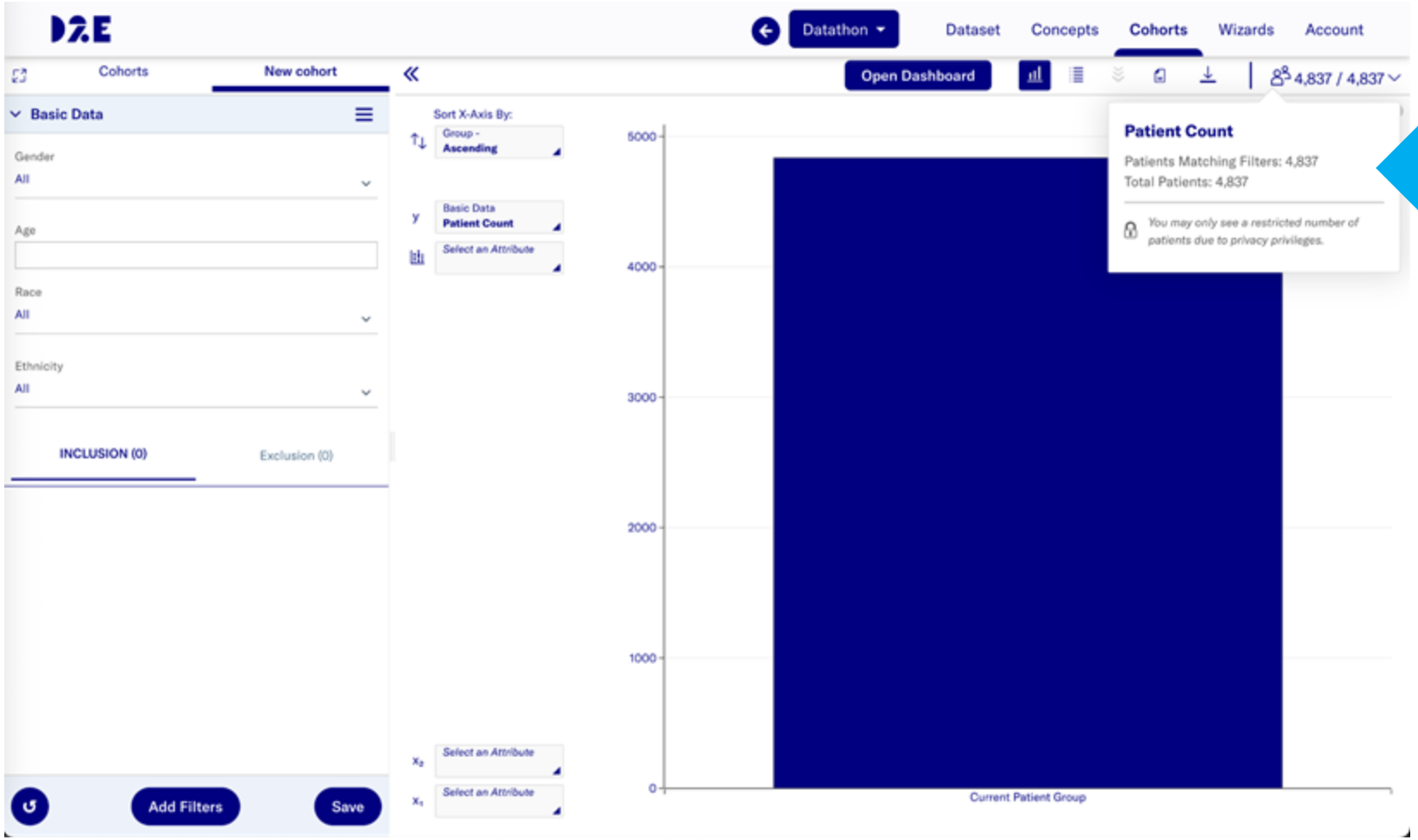
Slide credit: Pablo Guerrero, Data4Life

Cohort Building



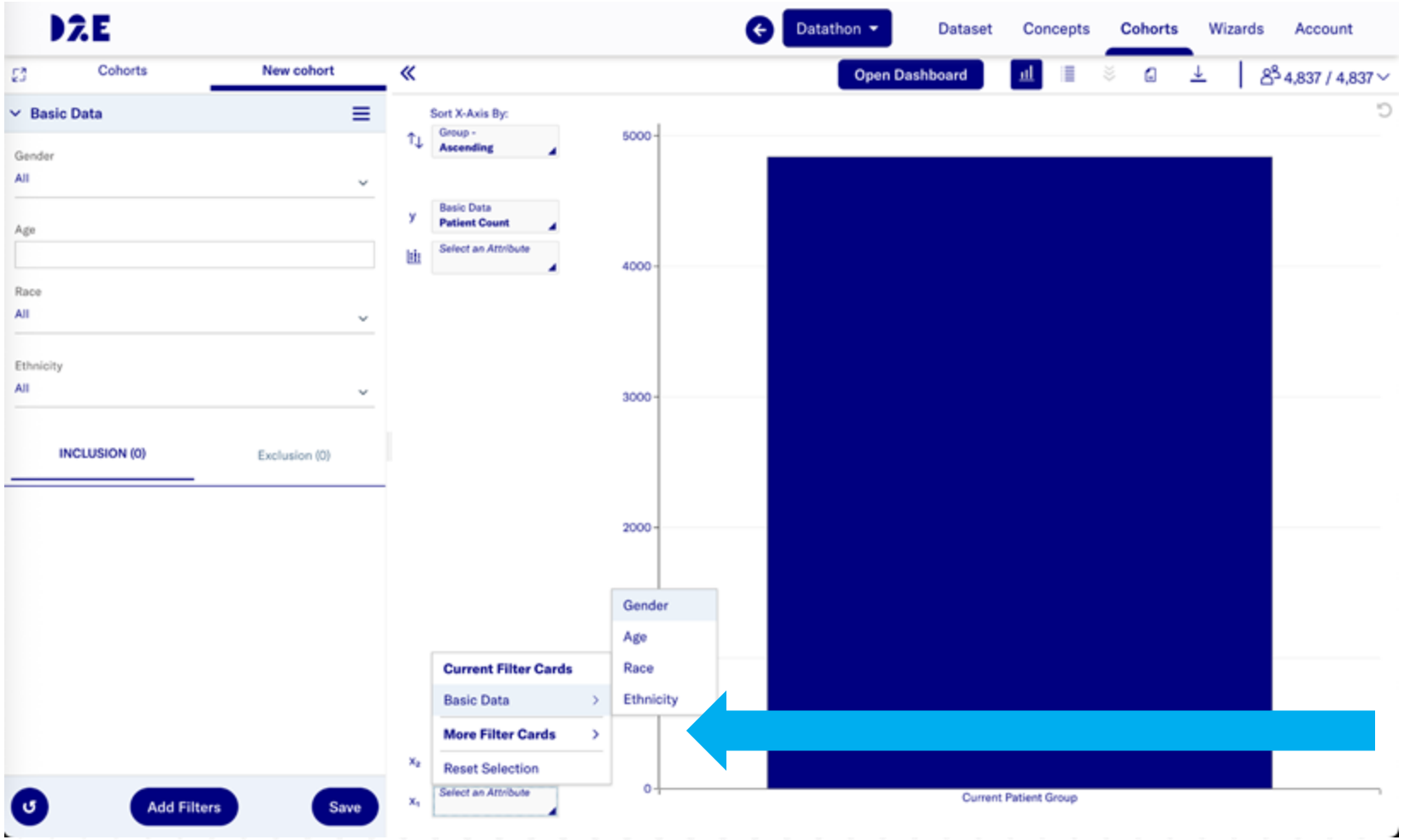
Slide credit: Pablo Guerrero, Data4Life

Cohort Building



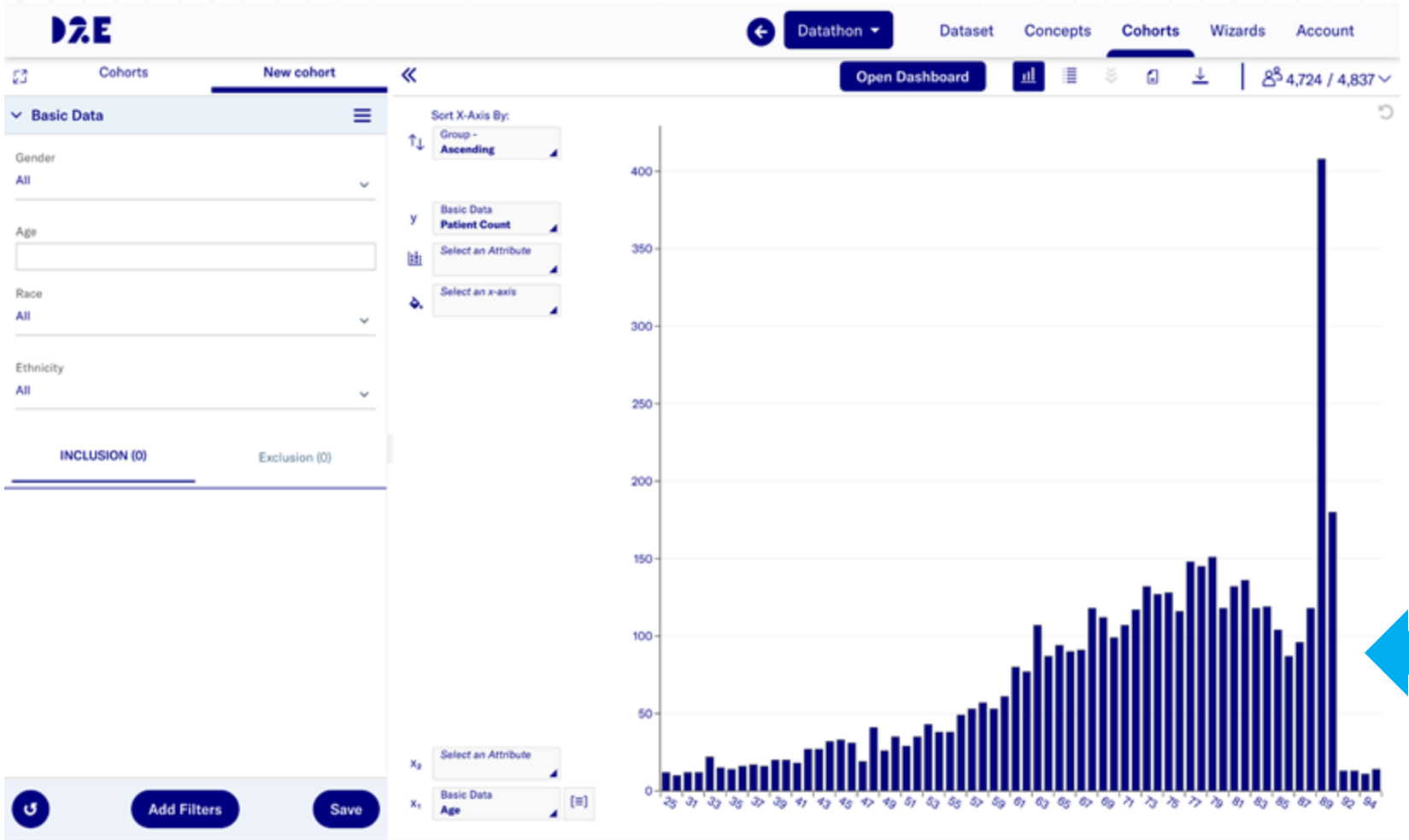
Total & matching patients

Cohort Building



Charting options

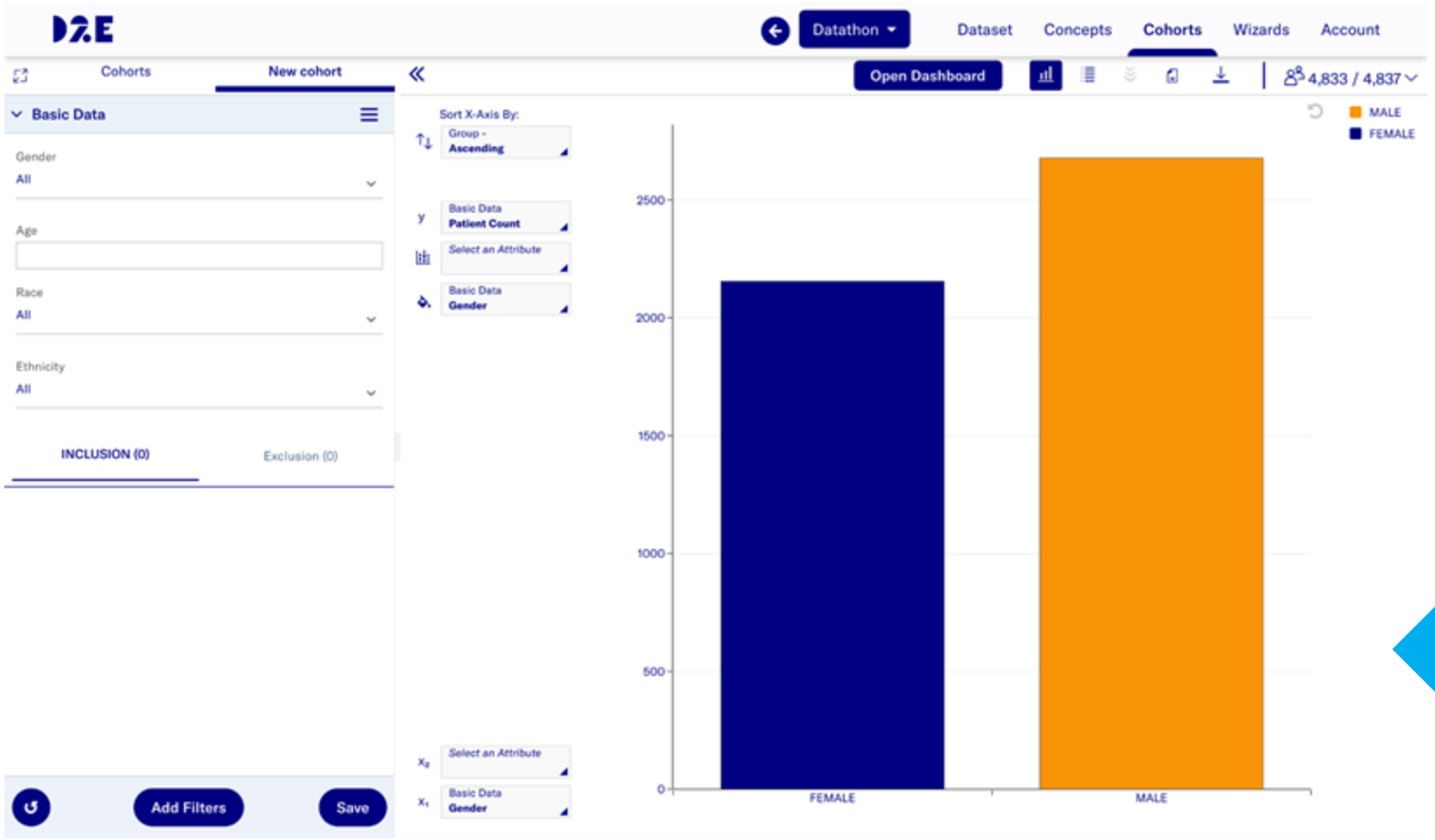
Cohort Building



Classify by age

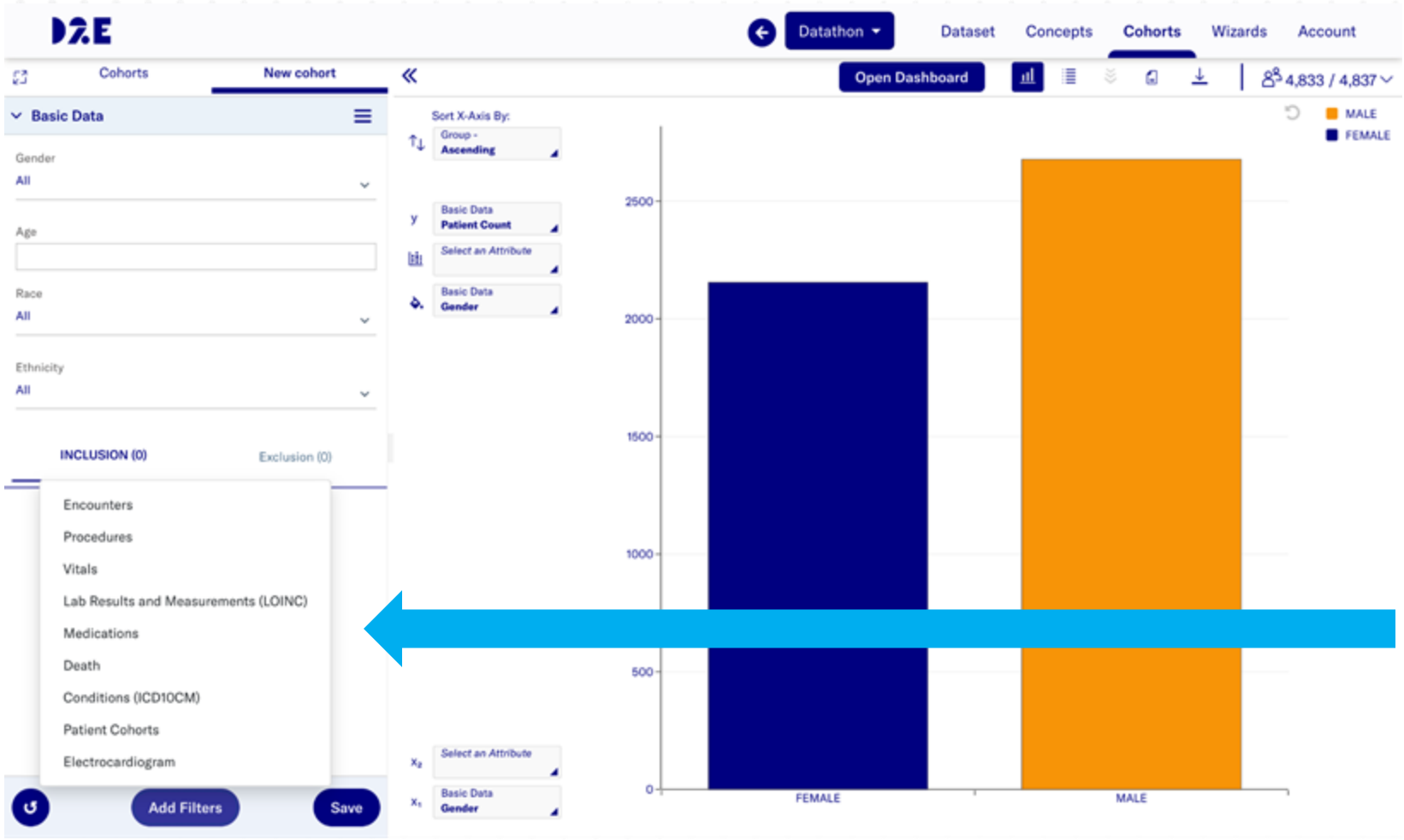
Slide credit: Pablo Guerrero, Data4Life

Cohort Building



Classify by gender

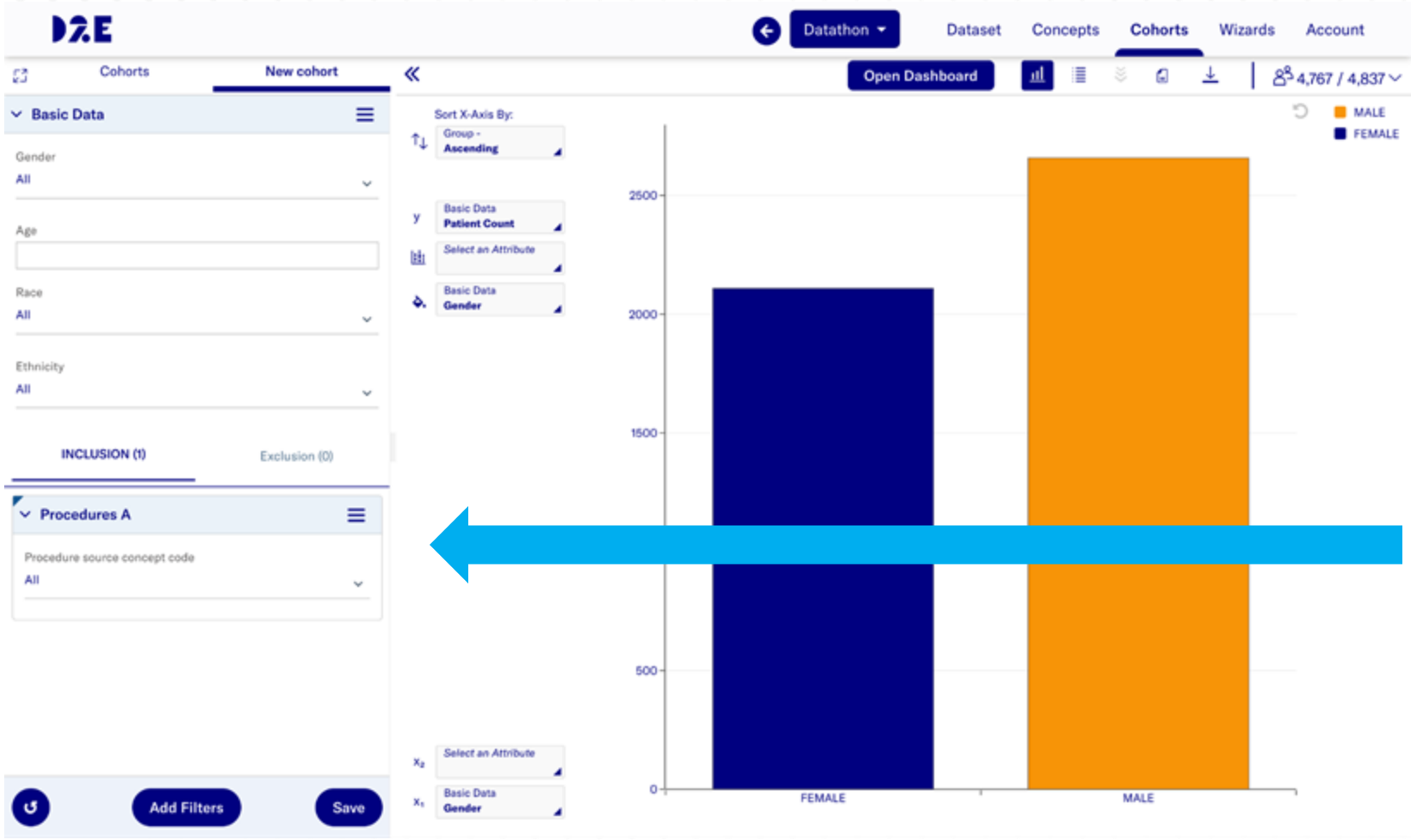
Cohort Building



Inclusion criteria (filters)

Slide credit: Pablo Guerrero, Data4Life

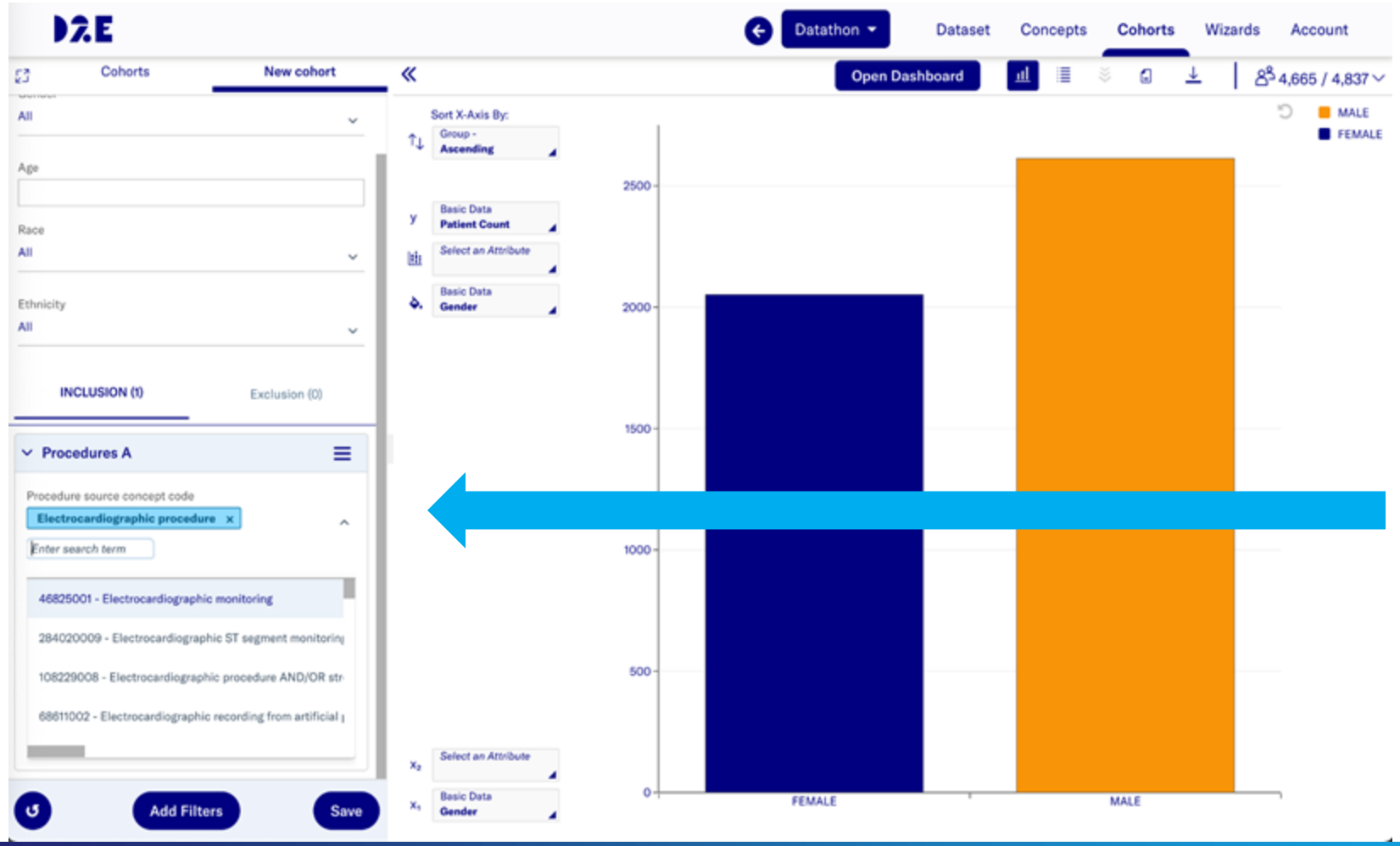
Cohort Building



Show only patients who have a procedure

Slide credit: Pablo Guerrero, Data4Life

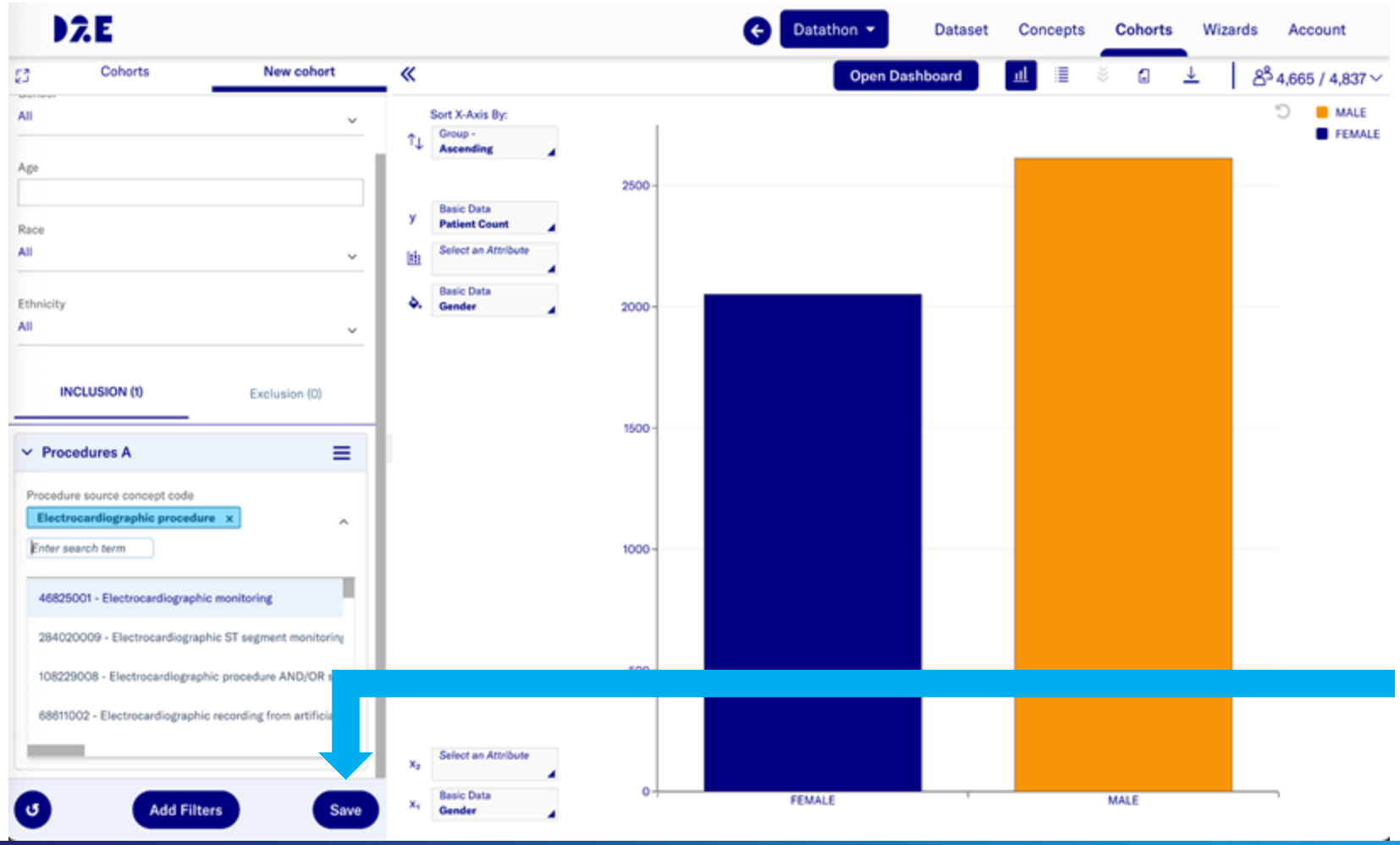
Cohort Building



Show only patients who have a procedure with *source concept code* "Electrocardiographic procedure"

Slide credit: Pablo Guerrero, Data4Life

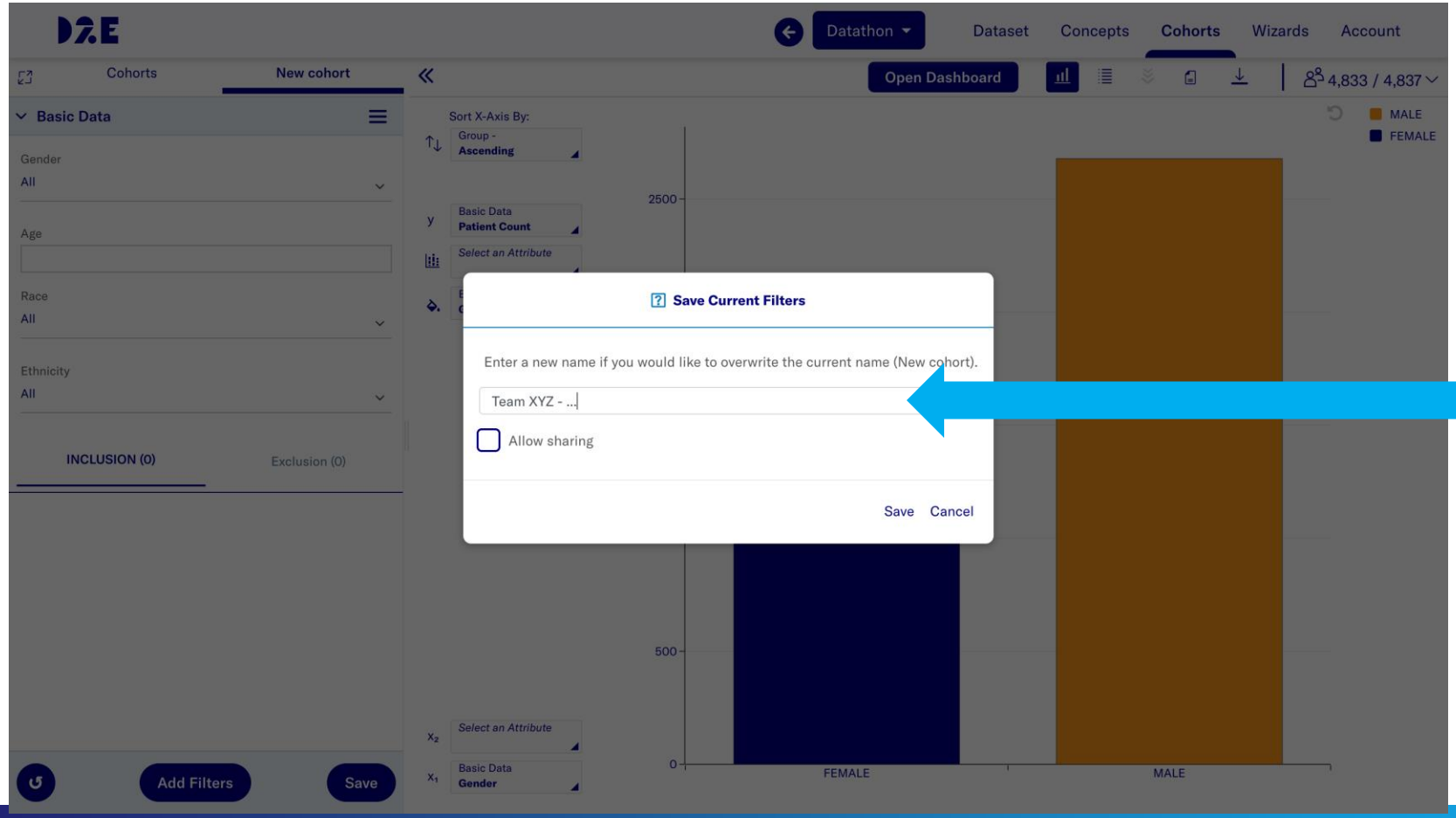
Cohort Building



Save your cohort

Slide credit: Pablo Guerrero, Data4Life

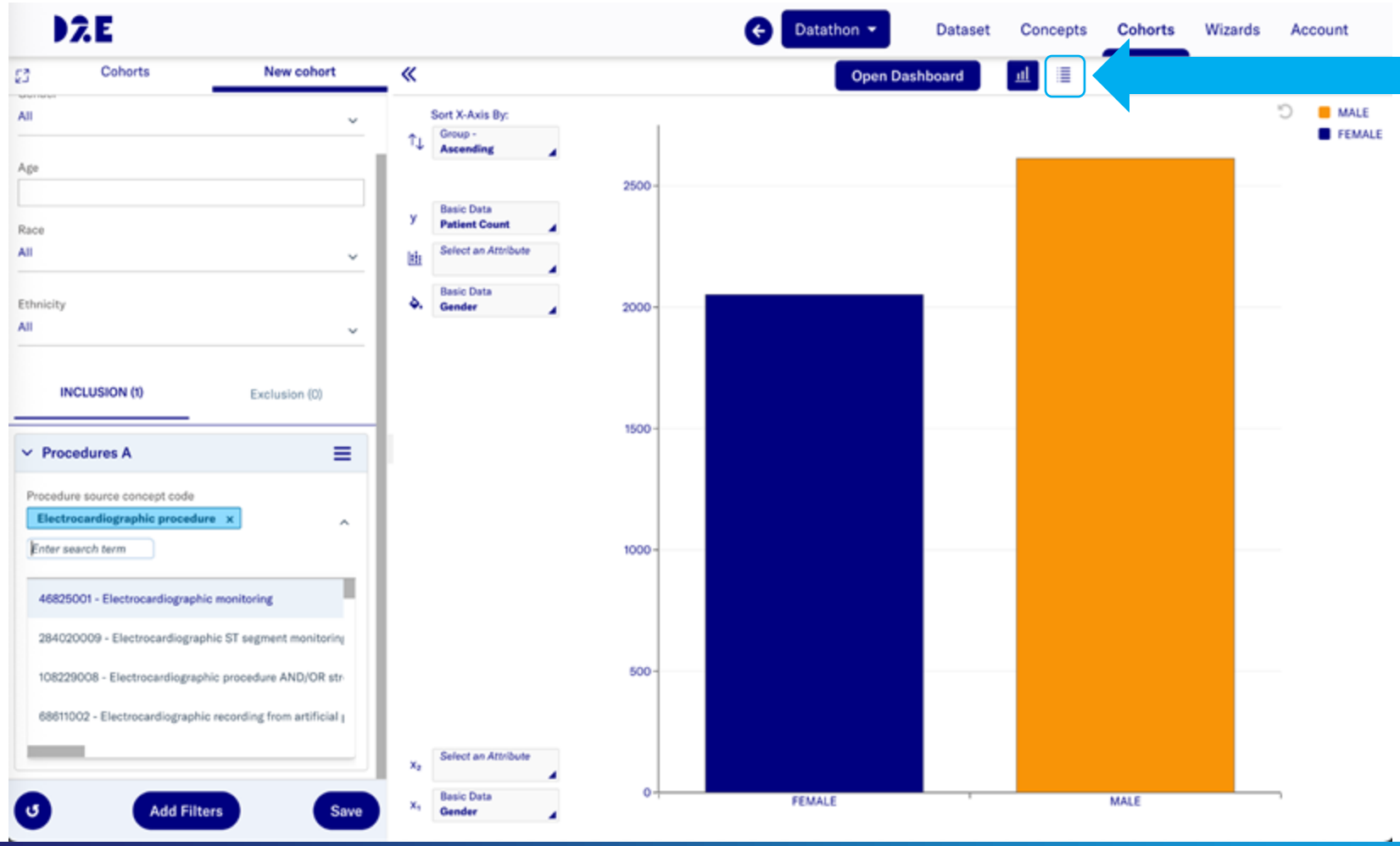
Cohort Building



Define cohort name

Slide credit: Pablo Guerrero, Data4Life

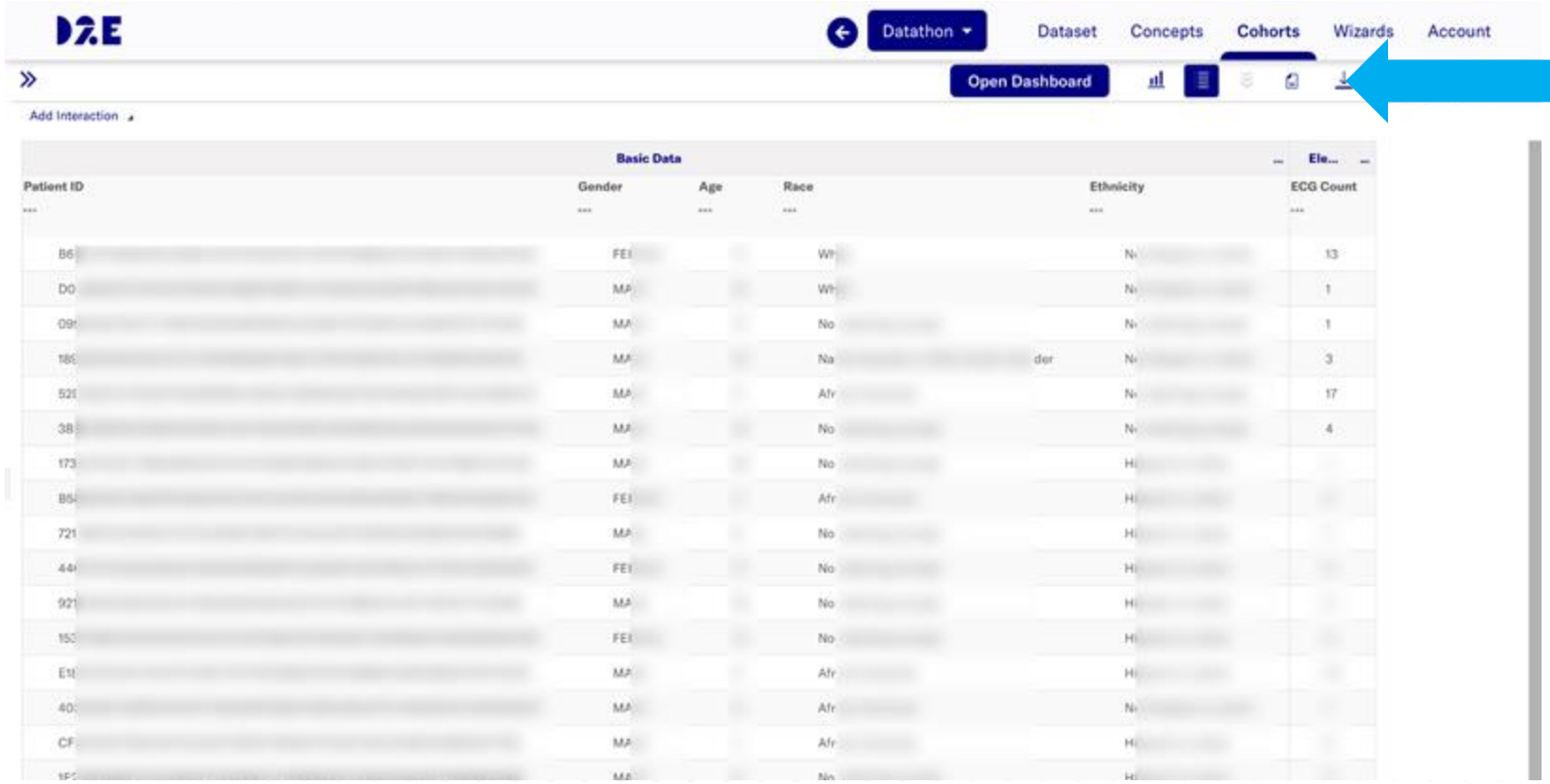
Cohort Building



Access patient list

Slide credit: Pablo Guerrero, Data4Life

Cohort Building



Download selected cohort

Slide credit: Pablo Guerrero, Data4Life

Thank you!