

Windreich Department of AI & Human Health

Introduction to Data2Evidence

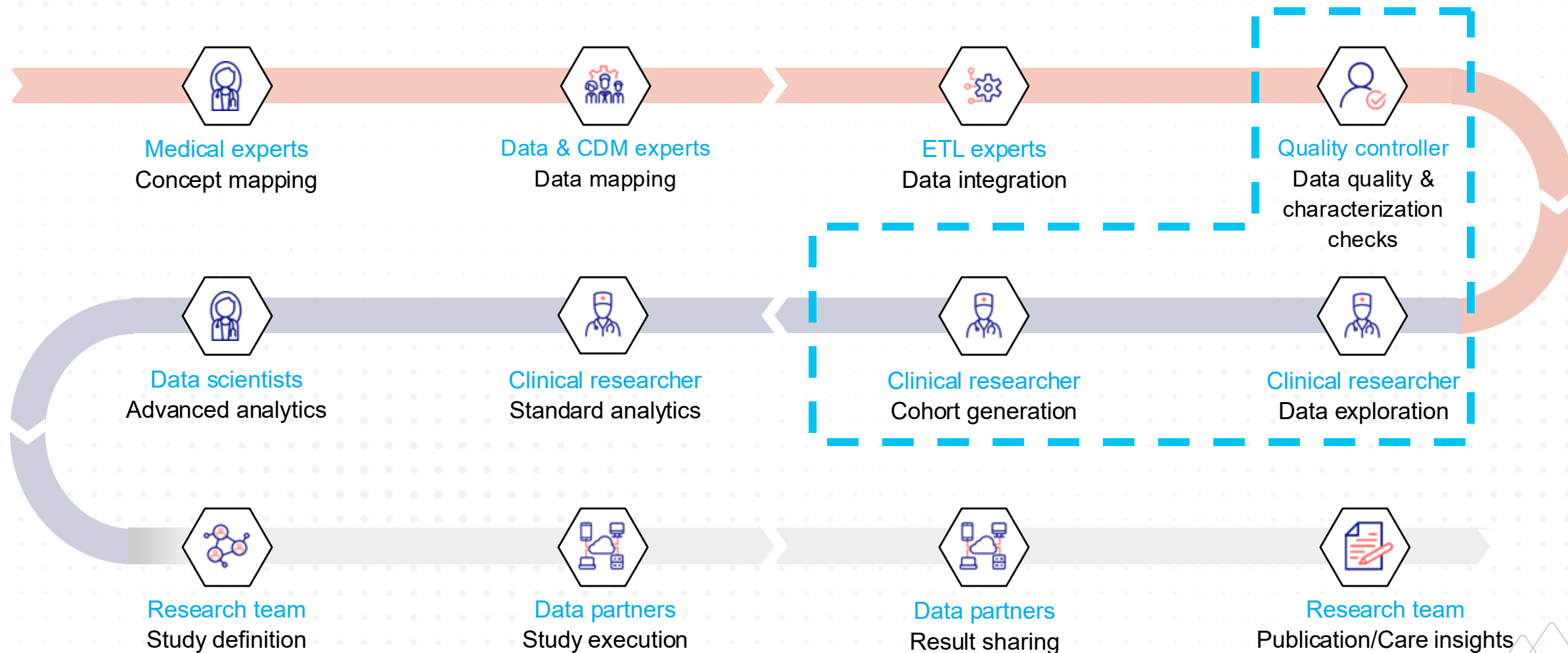
Ashwin Sawant, MD

May 12, 2026



Hasso Plattner Institute for Digital Health at Mount Sinai

Utilization along the research lifecycle



Making health data readily accessible for researchers

Extract and combine data into interoperable dataset

Electronic health records

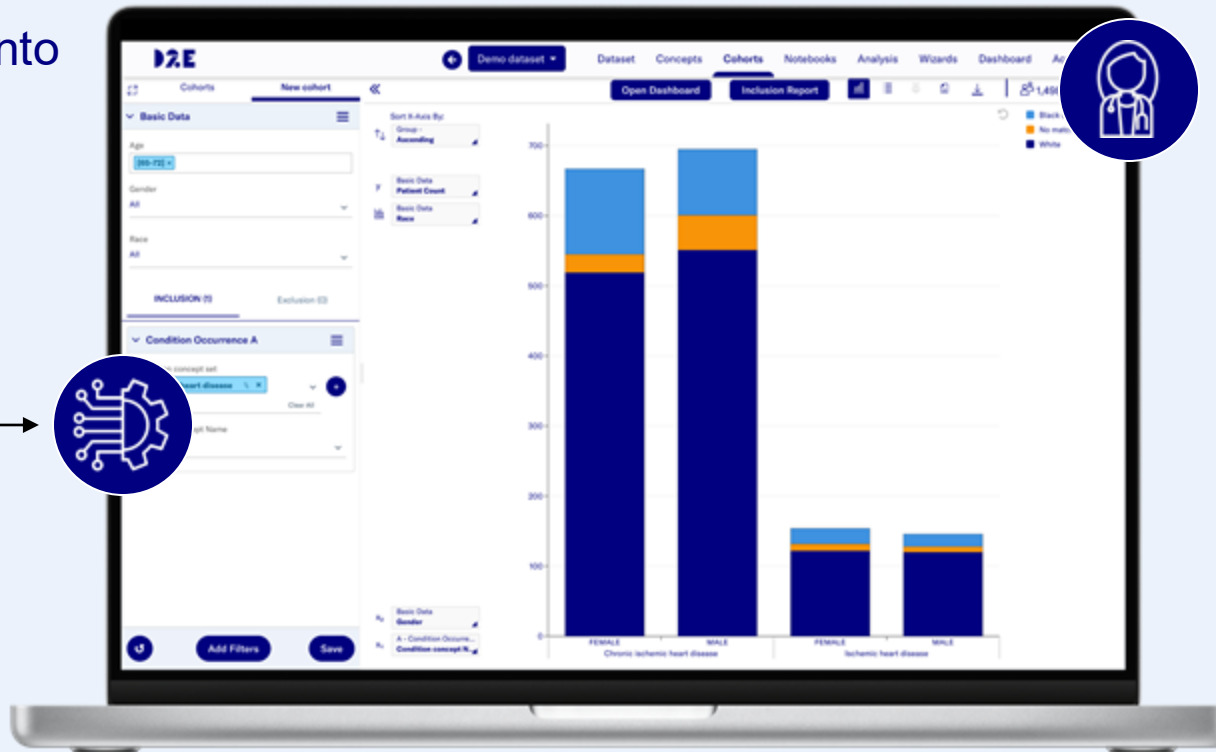
Lab records

Genomics data

Image (meta-)data

Unstructured data (notes)

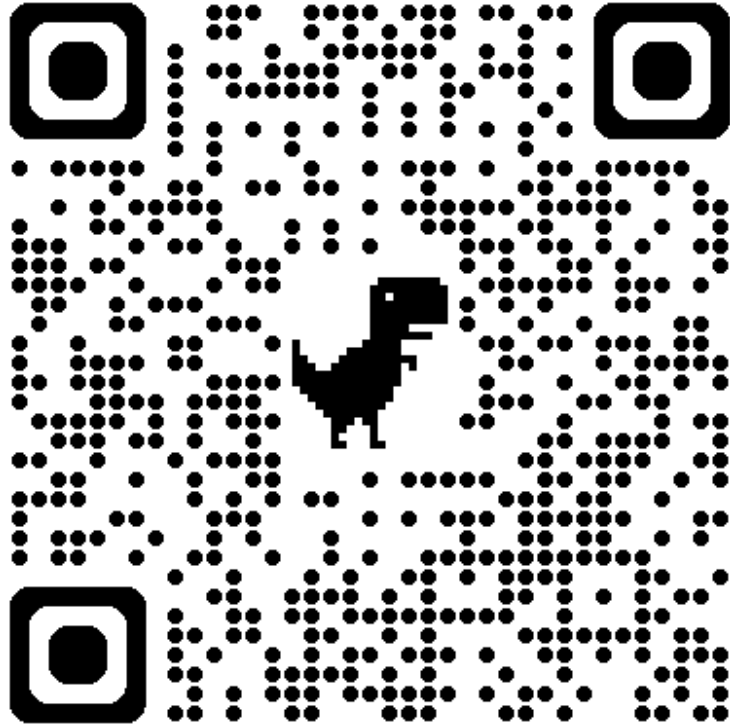
Wearable data (PROs)



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

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



d2e.airms.mssm.edu

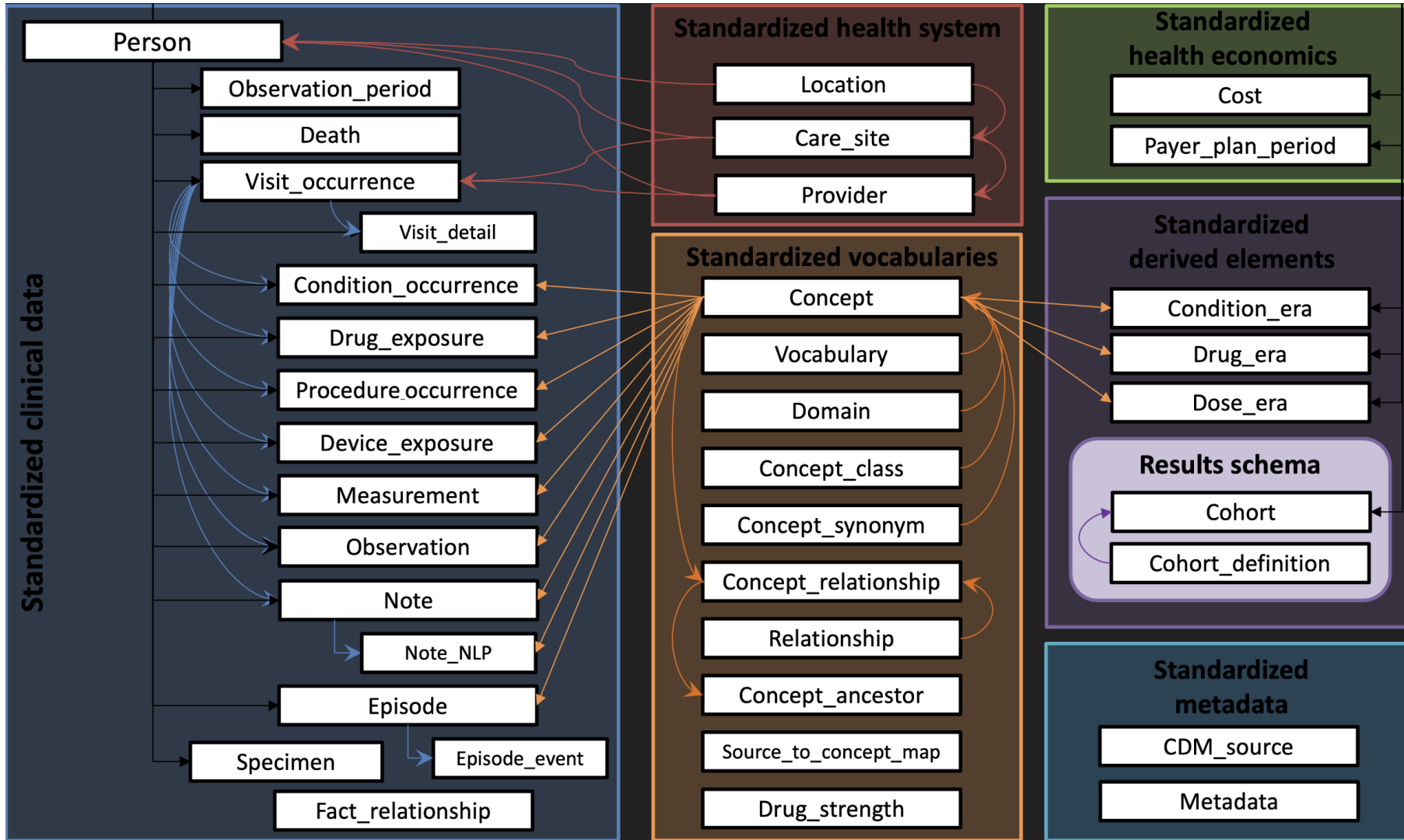
The Data Model

Data elements

Constraints on elements – standard vocabularies

Relationship between elements

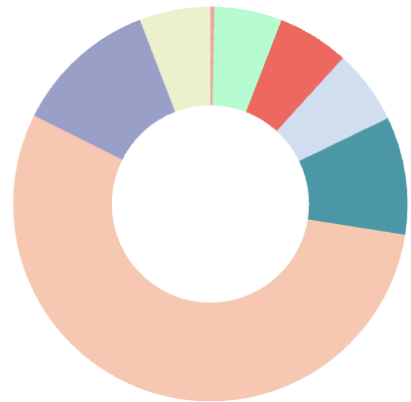
The Data Model



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



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

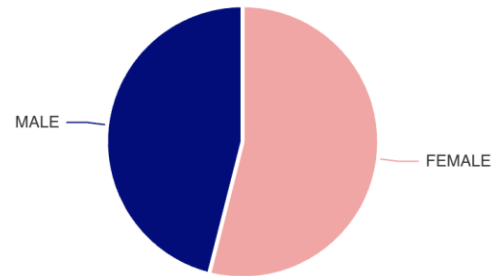
CDM Summary

Source name: CDMDEID

Number of persons:
12,441,531

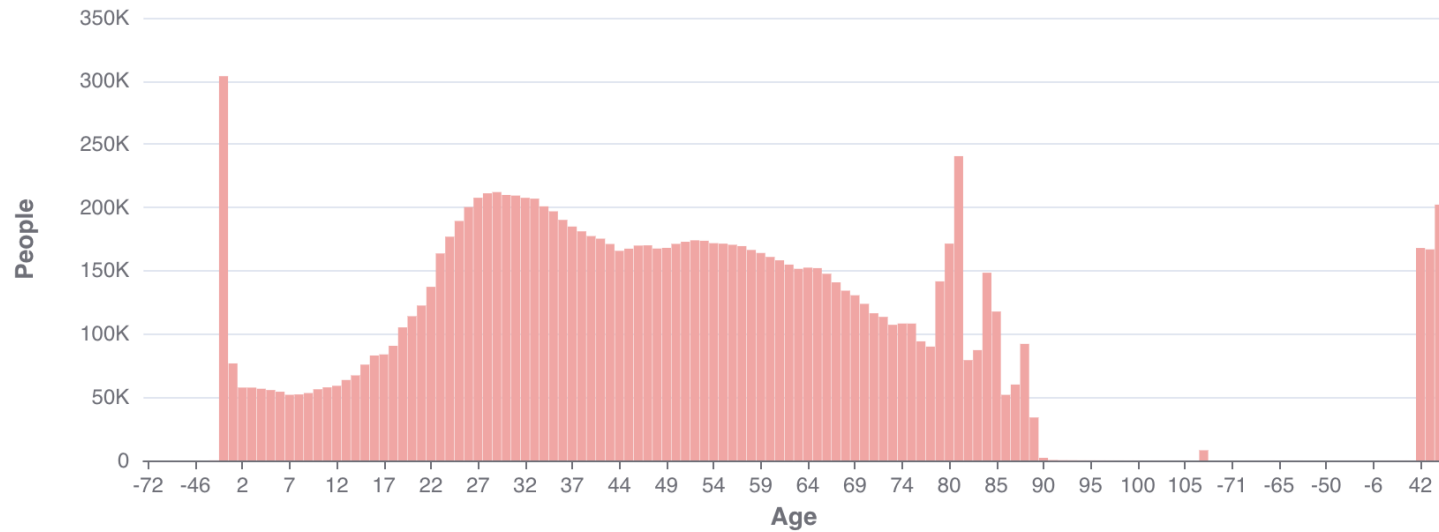
Gender

FEMALE MALE

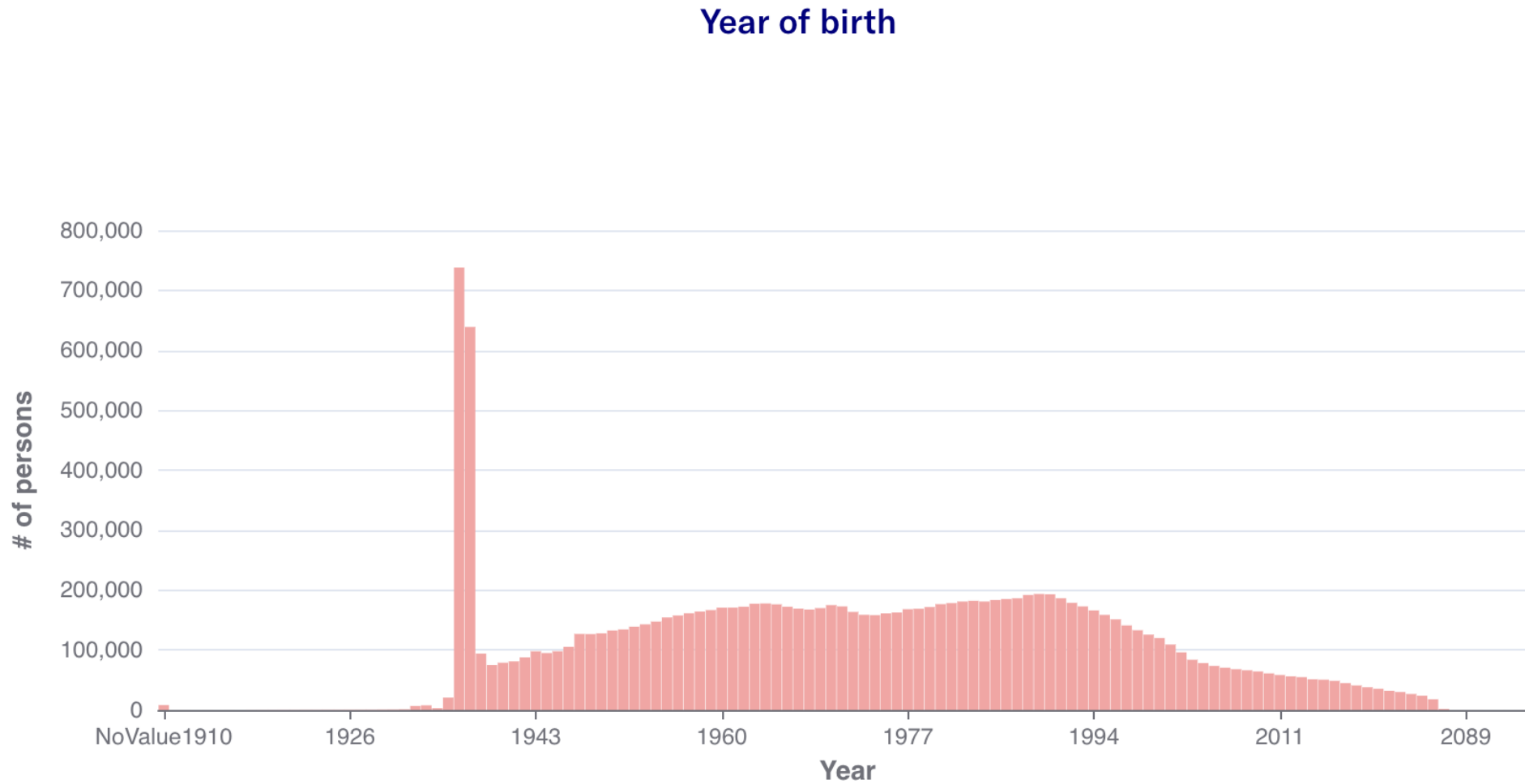


Dataset Characterization

Age at first observation



Dataset Characterization



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...								



Box Size: Prevalence



Dataset Characterization

Drug exposure Prevalence

Treemap Table



acetaminoph...	glucose l...	pantopra...	calcium...	magne...	midaz...	250	roc	glyc	hyd	ioh	250	me	dex	10	ac	dip
sodium chlo...	perflutren	albutero...	calcium...	magne...	midaz...	sodiu...	alu	alu	alu	lid	fur	lac	ma	24	ins	ma
	insulin l...	lorazepa...	calcium...	vanco...	amlod...											
heparin Inj...	furosemid...	potassiu...	calcium...	magne...	ipratr...	proc...	mag...	pot	pot	nal	50	at	epi	lab	azi	ta
	1 ML hy...	enoxap...	norepi...	famoti...	cefe...	mag...										
sodium chlo...	glucose O...	sulfur h...	docusa...	insuli...	pheny...	sodiu...	mag...	clop...	ferr...	pr	5	he	10	at	al	
ondansetron...	glucagon ...	acetami...	100 ML...	vanco...	metop...	cefaz...	mag...	lido...	alb...	m	c	at	5	g		
fentanyl 0....	pantopraz...	acetami...	magnes...	1 ML ...	furose...	ceftr...	mag...	sodi...	sod...							
					100 M...	albu...	piper...	hydr...	eto...	glu...	iop...	am	1	gu		
sennosides,...	50 ML glu...	bisacody...	magnes...	oxycy...	1 ML ...	cefe...	pota...	lido...	dip...	cef...	1 ...	ond...	1	5		
polyethylen...	aspirin 8...	10 ML pr...	magnes...	Microe...	250 M...	cefe...	pota...	met...	dip...	sen...	so...	met...	tra...			

1.00 155.00

Average records per person



Box Size: Prevalence



Dataset Characterization

Procedure Prevalence

Treemap Table



Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 268400002: 12 lead ECG	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 268400002: 12 lead ECG	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 268400002: 12 lead ECG	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 268400002: 12 lead ECG	Current Procedural Terminology (CPT) 93.05
Procedure on cardiovascular system Cardiovascular investigation Electrocardiographic procedure Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 268400002: 12 lead ECG Prevalence: 0.00476 Number of persons: 59,262 Average records per person: 21.2									
Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) 268400002: 12 lead ECG	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05
Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05
Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05	Current Procedural Terminology (CPT) 93.05

1.00 213.00

Average records per person

Box Size: Prevalence



Dataset Characterization

Condition occurrence Prevalence

Treemap **Table**



Dyspnea	Sepsis	Constipa...	Cough	Throm...	Ab...	Mi...	Ac...	Fun	Dru	Sei	Dia	Chr	Hy	Hy	Chr
		Electroc...	Type 2 ...	Paroxy...	Chro...	Hy	Ge	Syn	Gas	Co	Lo	Tac	Mo	Ac	Hy
Essential hype...	Anemia	Hyperkal...	Fever	Pleural...	Pulm...	Hype...	Ch	De	Ch	Ac	Ba	Na	Hy	Cir	Er
	Metabolic ...	Coronary...	Vitamin...	Sepsis ...	End-...	Depr...	Ane...	Re	Ins	Na	Be	D	An	Di	Ed
Acute kidney i...	Pneumonia	Urinary...	Acidosis	Cere...	Pain ...	Loca...	Cer...	Hyp...	Os	As	No	Ac	Ch	Ba	
		Heart fa...	Diabete...	Slow tr...	Dysp...	Gast...	Chro...	Chro...	Ost...	At	Di	H	Br	Pe	
Acute hypoxemi...	Chest pain	Septic s...	Acute r...	Shock	Gastr...	Syst...	Hyp...	Ane...	Lun...	Palp...	Pul...	H	S	E	C
		Atheroscle...	Pain	Low blo...	COVID...	Iron ...	Adult...	Car...	Dizz...	Pne...	Dis...	Le...	Li	Hy	
Hypertensive d...	Atrial fib...	Cardiac ...	Acute p...	Hypoth...	Pure ...	Nicot...	Anxi...	Diso...	Anxi...	Me...	De...				
	Severe pro...	Disorder...	Chronic...	Dyspn...	Heart...	Beni...	Hyp...	Obe...	Acut...	Dis...	De...				

1.00 152.00

Average records per person

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 Management



D2E

Datathon Dataset **Concepts** Cohorts Wizards Account

Concept search Concept sets

Search

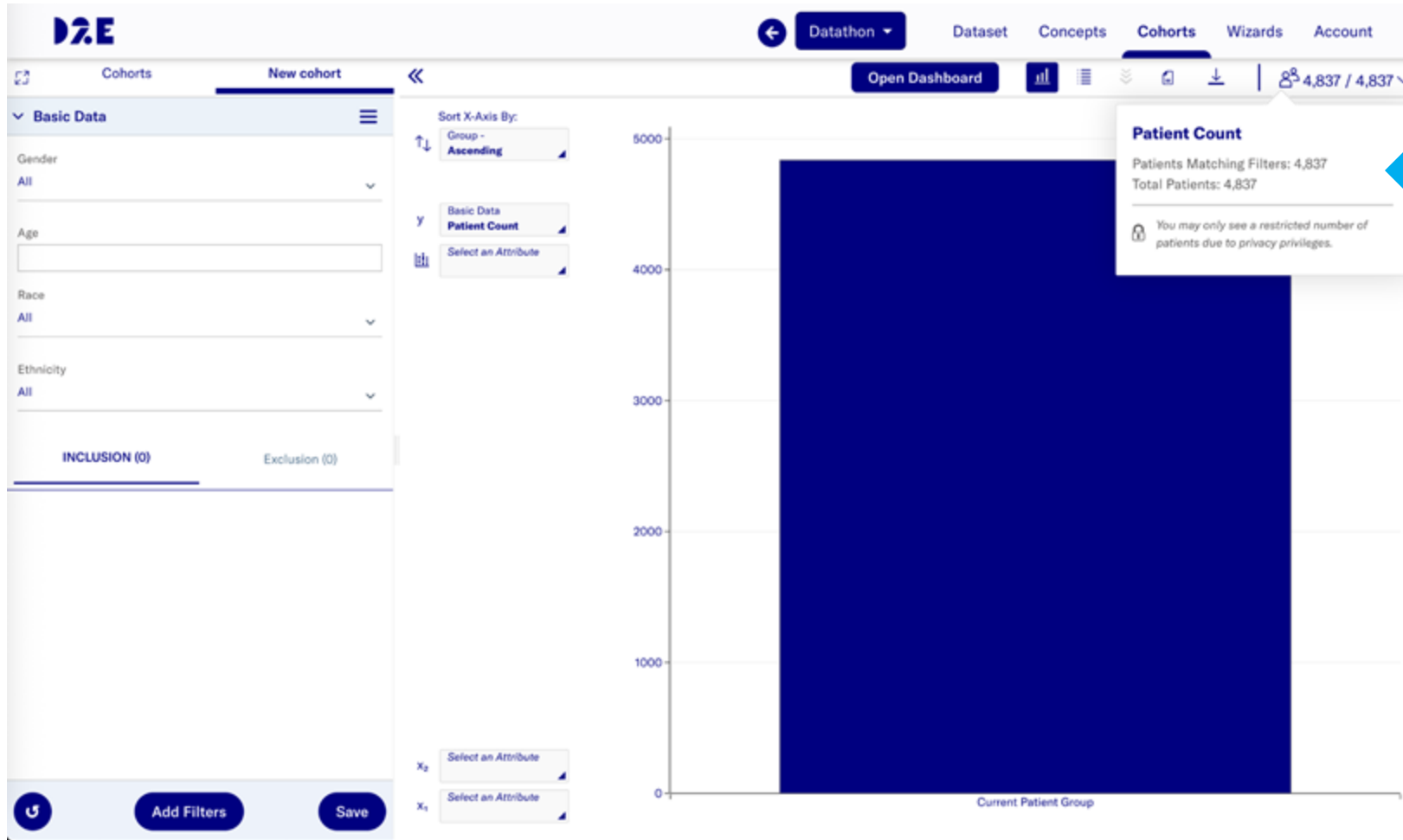
ID	Code	Name	Vocabulary	Concept	Domain	Class	Validity	RC
			Filter by Vocabulary X	Filter by Concept X	Filter by Domain X	Filter by Class X	Filter by Validity X	
45426237	9982.00		Read	Non-standard	Condition	Read	Invalid	0
45489629	9C3..12		Read	Non-standard	Observation	Read	Invalid	0
45426233	98Bn.00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

Cohort Building

The screenshot shows the 'Cohorts' page in the D2E application. At the top, a navigation bar includes 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 page title 'Cohorts' is displayed. The main content area is titled 'Create Cohort:' and features a dark blue button labeled 'D2E', a light blue button labeled 'Compare', and a 'Shared' toggle switch. A message below the buttons reads: 'You have not yet saved any filters. You can save your current filter settings.' Two blue callout arrows are present: arrow '1' points to the 'Cohorts' tab in the navigation bar, and arrow '2' points to the 'D2E' button.

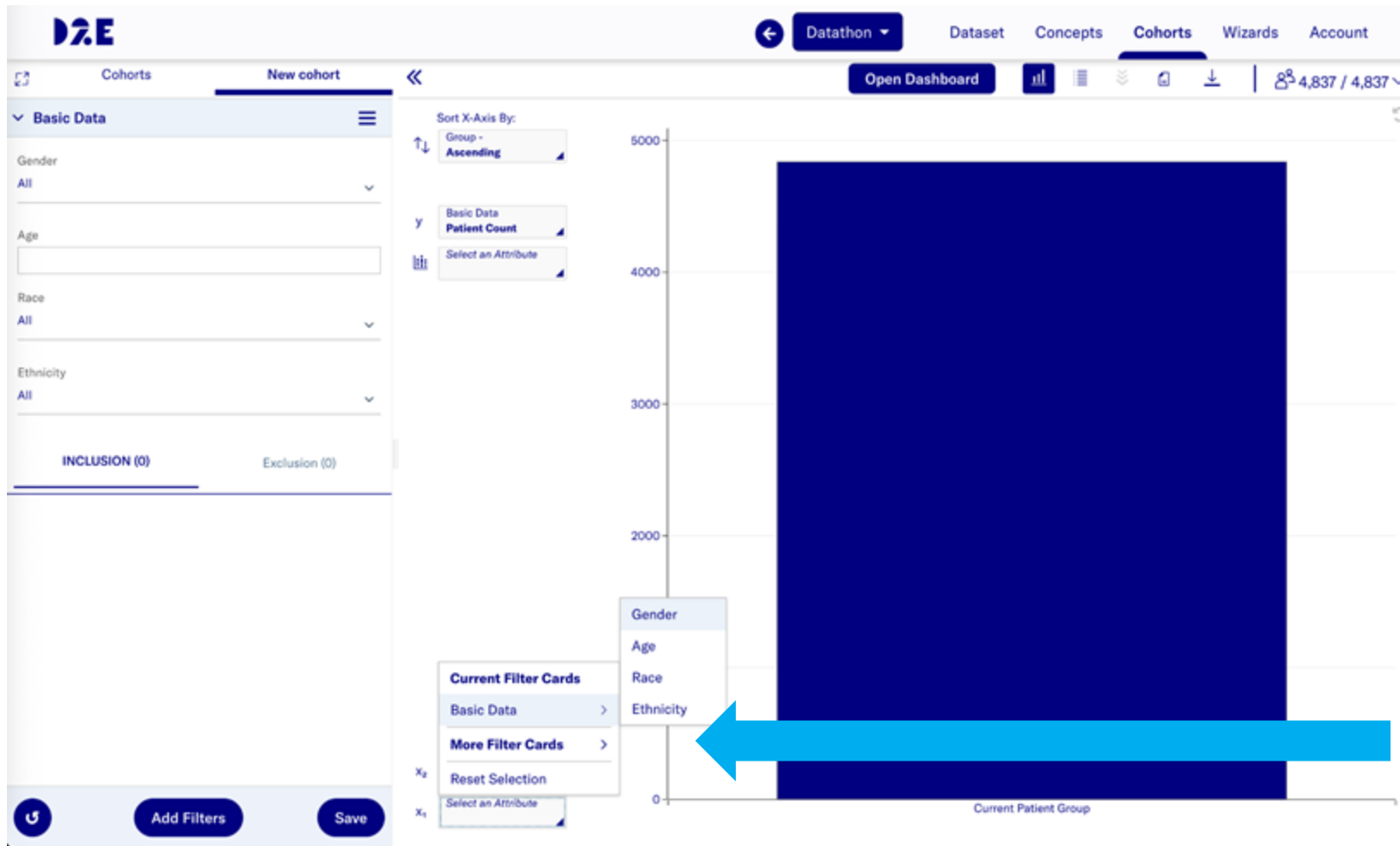


Cohort Building



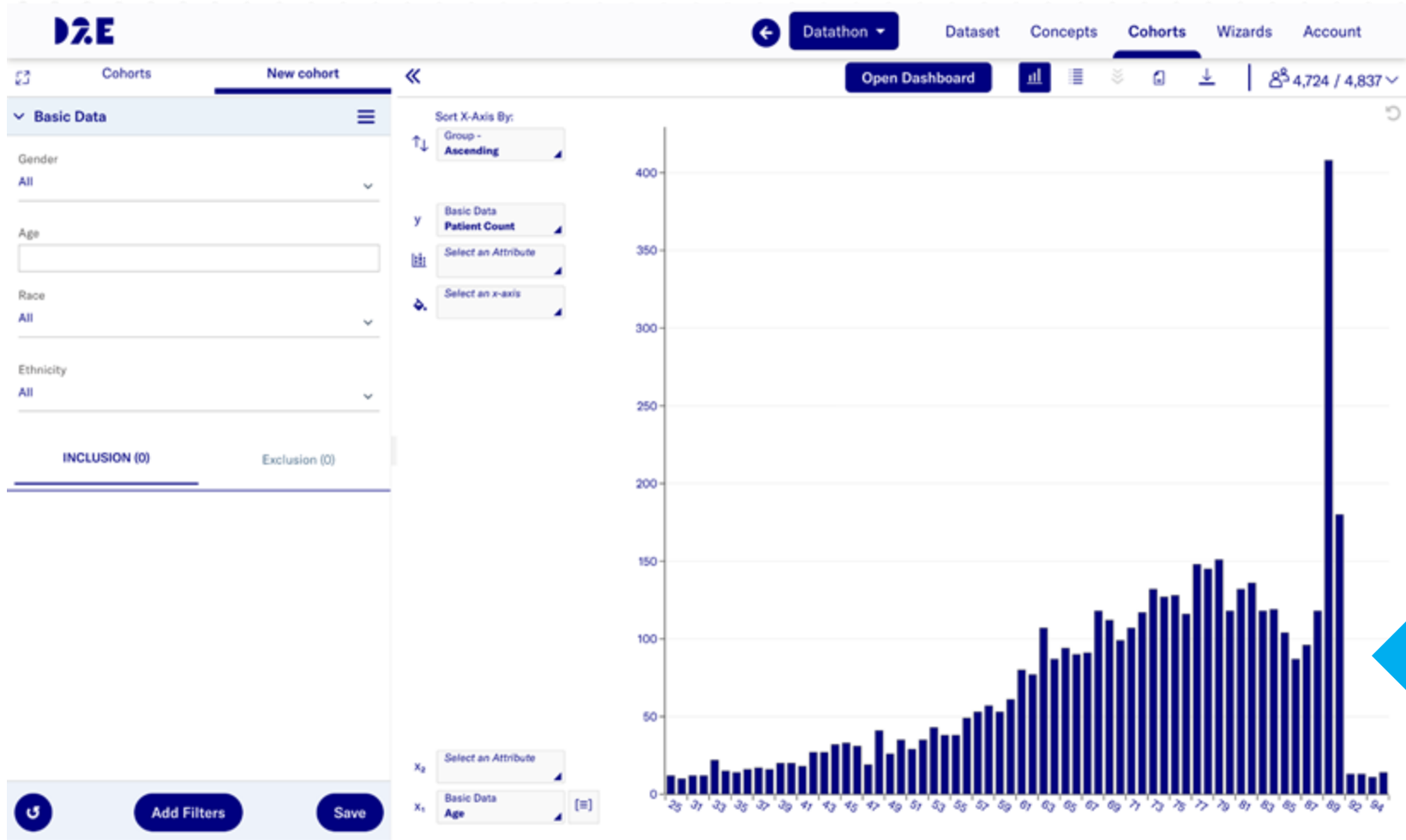
Total & matching patients

Cohort Building



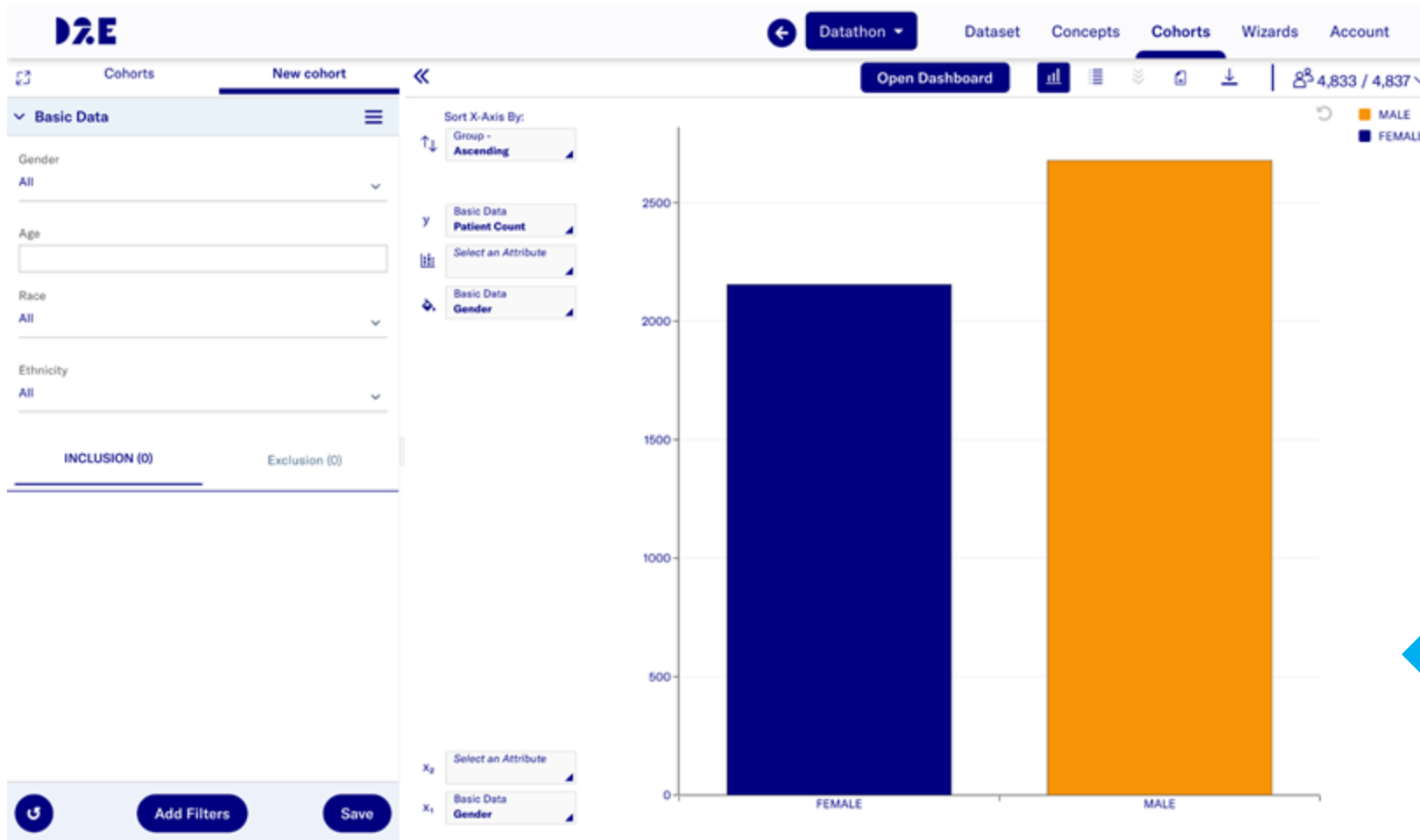
Charting options

Cohort Building



Classify by age

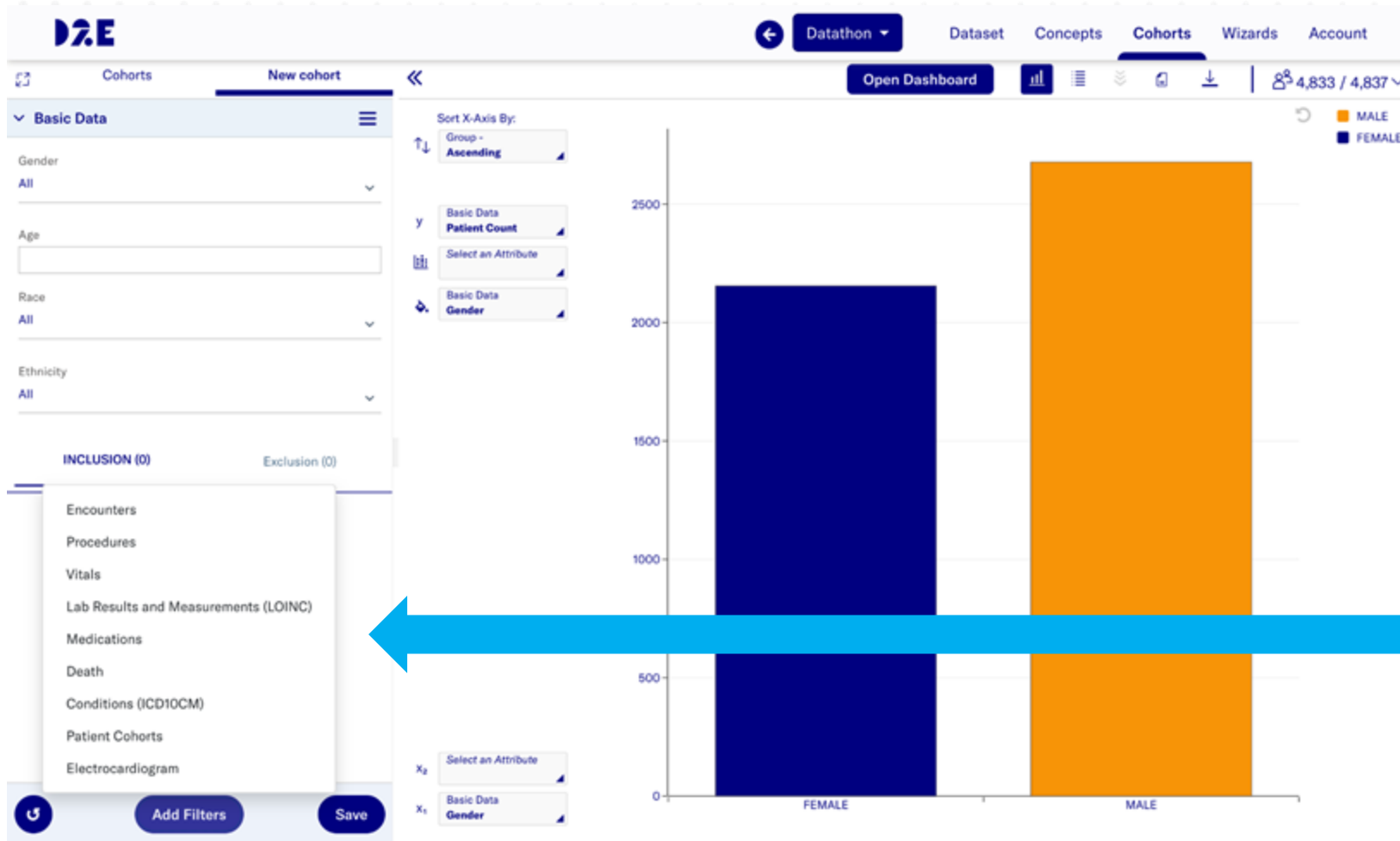
Cohort Building



Classify by gender

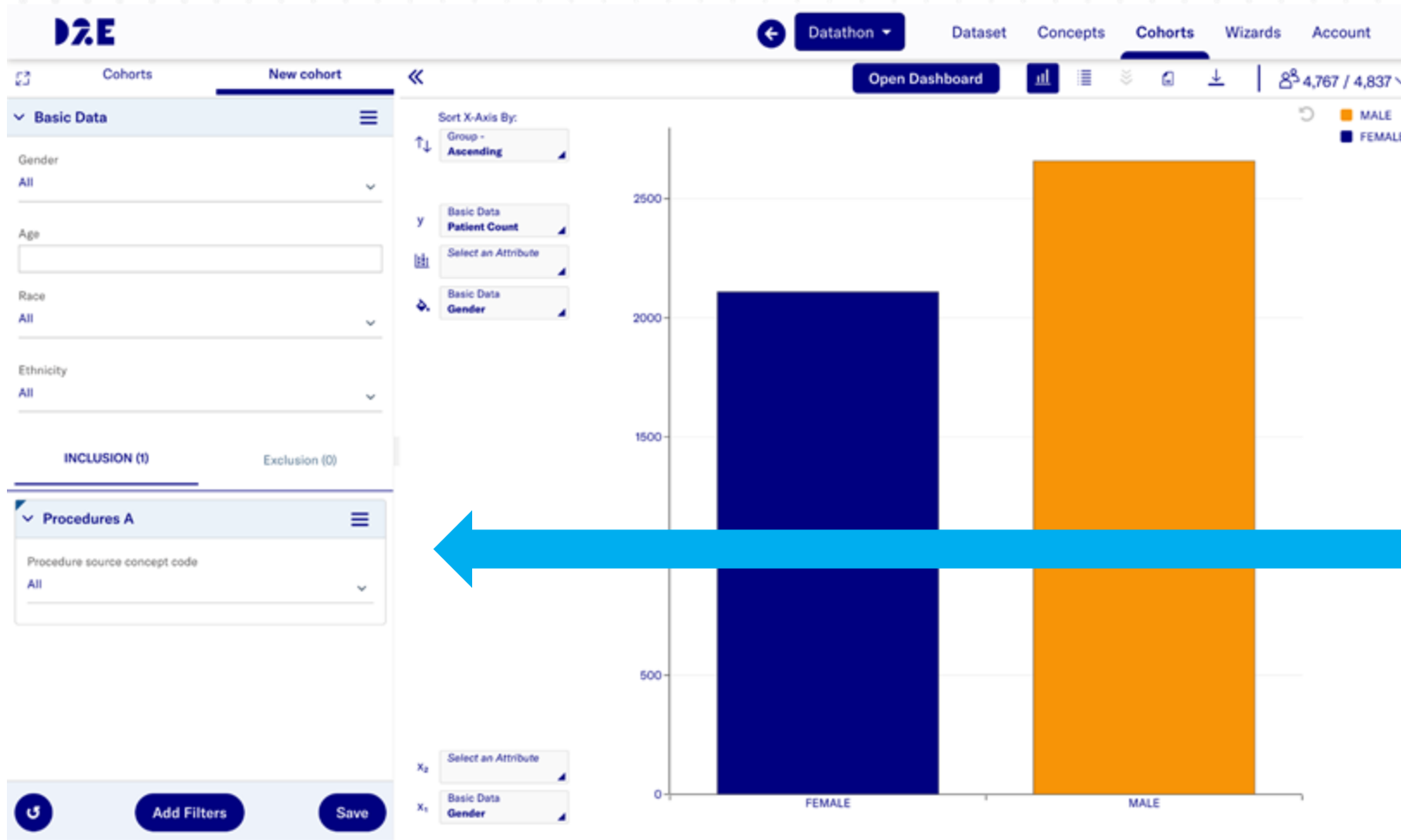


Cohort Building



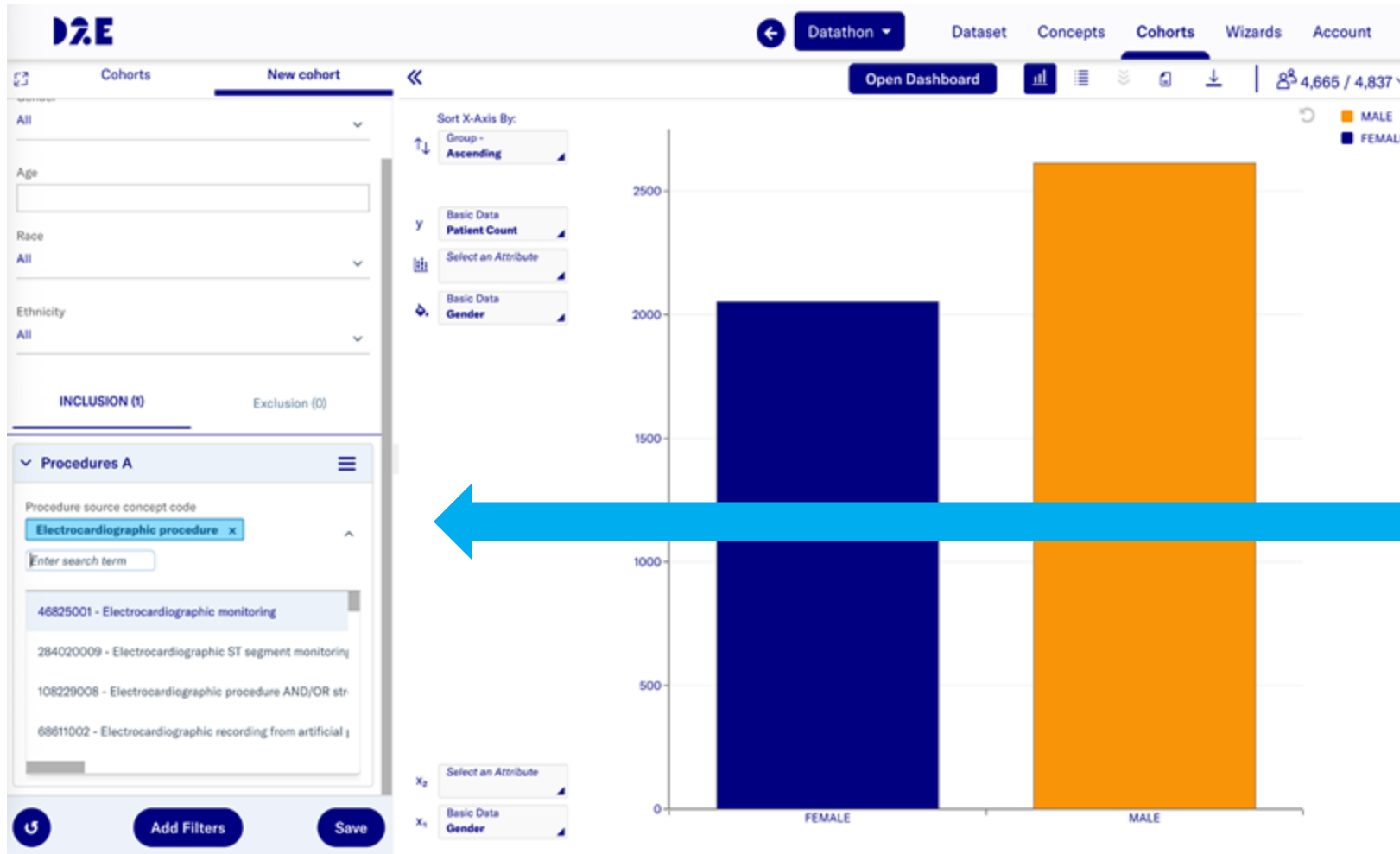
Inclusion criteria (filters)

Cohort Building



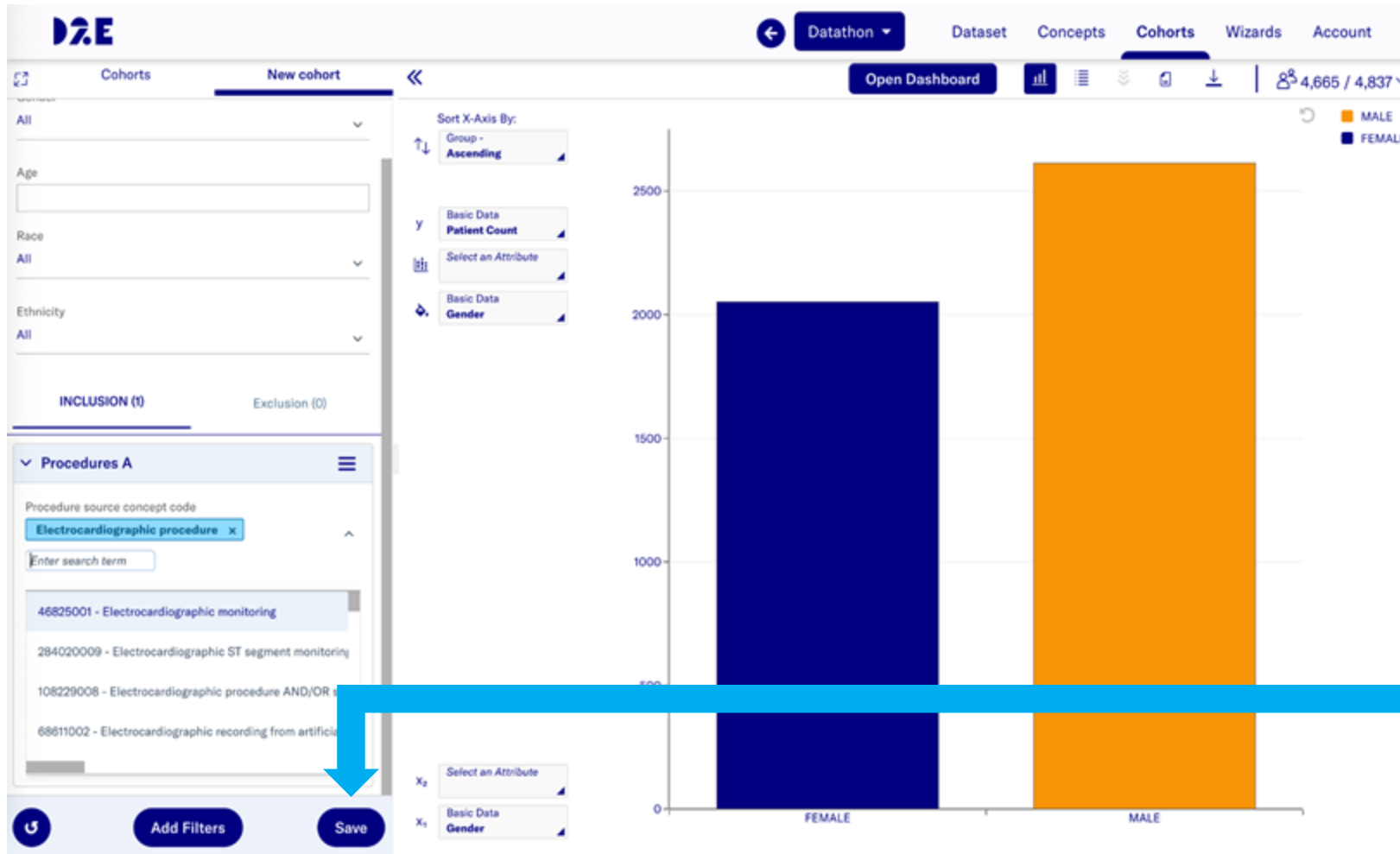
Show only patients who have a procedure

Cohort Building



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

Cohort Building



Save your cohort

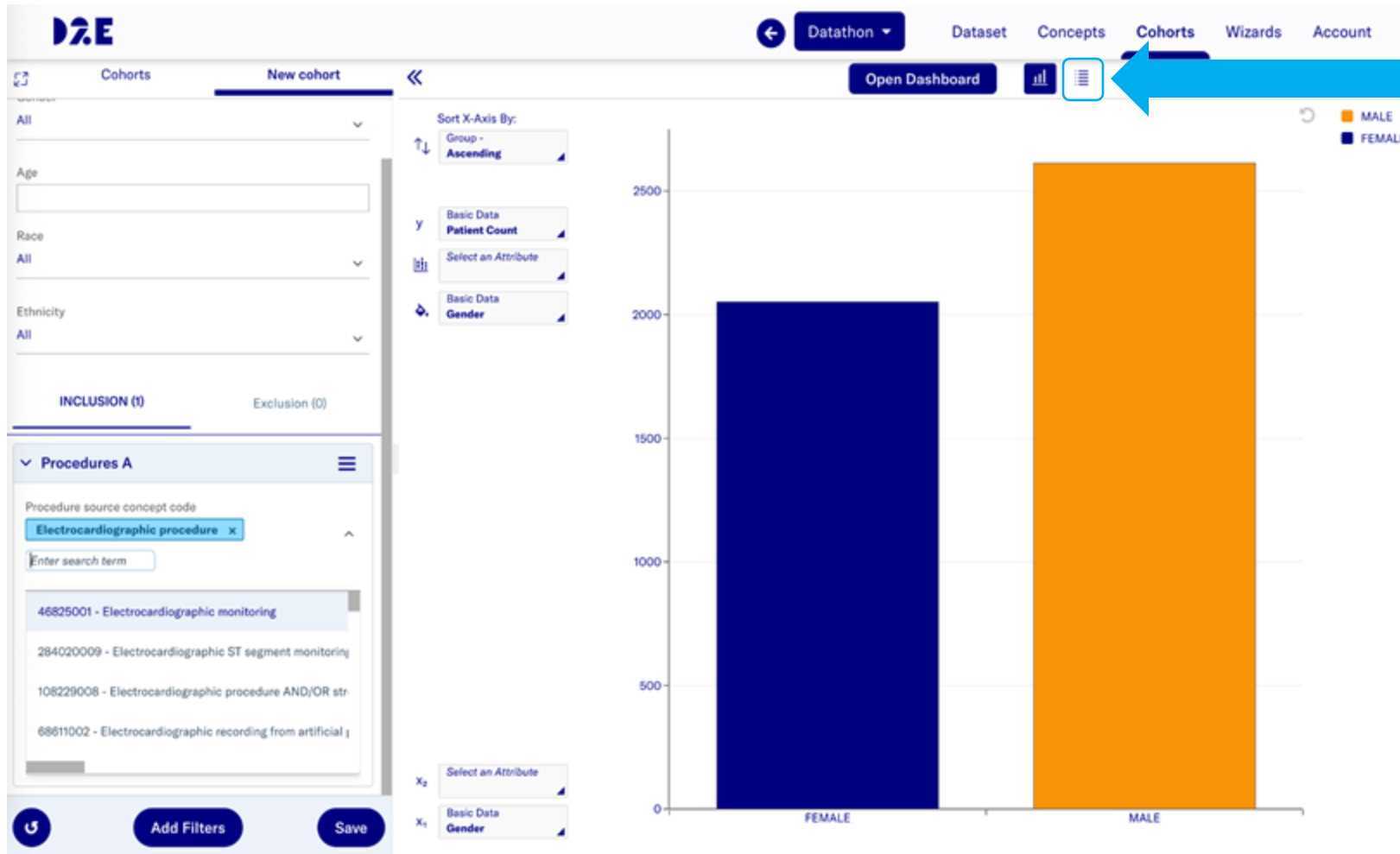
Cohort Building

The screenshot displays the D7E Cohort Building interface. On the left, a sidebar lists 'Basic Data' filters for Gender, Age, Race, and Ethnicity, all set to 'All'. Below this are 'INCLUSION (0)' and 'Exclusion (0)' sections. The main area features a bar chart with 'Patient Count' on the y-axis (0 to 2500) and 'Gender' on the x-axis (FEMALE, MALE). A legend indicates MALE is represented by a brown bar and FEMALE by a dark blue bar. A 'Save Current Filters' dialog box is overlaid on the chart, containing a text input field with 'Team XYZ - ...', an 'Allow sharing' checkbox, and 'Save' and 'Cancel' buttons. A blue arrow points from the text 'Define cohort name' to the input field.

Define cohort name

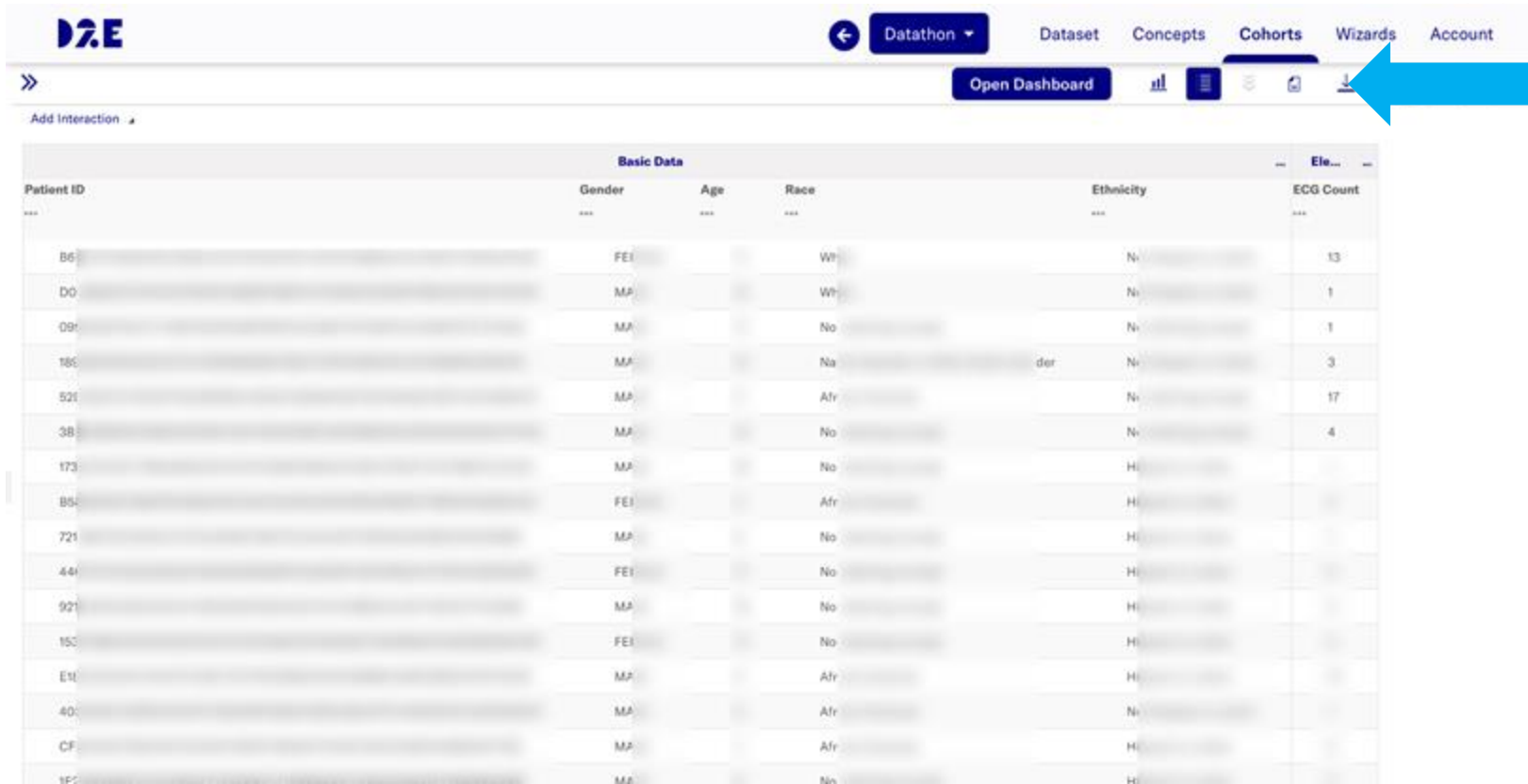


Cohort Building



Access patient list

Cohort Building



The screenshot shows a web interface for cohort building. At the top, there is a navigation bar with the logo 'D2E' on the left and a menu on the right containing 'Datathon', 'Dataset', 'Concepts', 'Cohorts', 'Wizards', and 'Account'. Below the navigation bar, there is a toolbar with 'Open Dashboard' and several icons. A blue arrow points to a download icon in the toolbar. Below the toolbar is a table with the following columns: Patient ID, Gender, Age, Race, Ethnicity, and ECG Count. The table contains 15 rows of patient data.

Patient ID	Gender	Age	Race	Ethnicity	ECG Count
B6	FEI		Wh	Non-Hispanic	13
DO	MA		Wh	Non-Hispanic	1
09	MA		Non	Non-Hispanic	1
186	MA		Non	Hispanic	3
521	MA		Afr	Non-Hispanic	17
38	MA		Non	Non-Hispanic	4
173	MA		Non	Hispanic	
85	FEI		Afr	Hispanic	
721	MA		Non	Hispanic	
44	FEI		Non	Hispanic	
921	MA		Non	Hispanic	
150	FEI		Non	Hispanic	
E11	MA		Afr	Hispanic	
40	MA		Afr	Non-Hispanic	
CF	MA		Afr	Hispanic	
182	MA		Non	Hispanic	

Download selected cohort



Thank you!