The *All of Us* Research Program: data quality, utility, and diversity

Andrea H. Ramirez*[1,2], Lina Sulieman[2], David J. Schlueter[2], Alese Halvorson[2], Jun Qian[2], Francis Ratsimbazafy[3], Roxana Loperena[3], Kelsey Mayo[3], Melissa Basford[3], Nicole Deflaux[4], Karthik N. Muthuraman[4], Karthik Natarajan[5], Abel Kho[6], Hua Xu[7], Consuelo Wilkins[1], Hoda Anton-Culver[8], Eric Boerwinkle[9], Mine Cicek[10], Cheryl R.Clark[11], Elizabeth Cohn[12], Lucila Ohno-Machado[13], Sheri Schully[14], Brian K. Ahmedani[15], Maria Argos[16], Robert M. Cronin[2], Christopher O'Donnell[17], Mona Fouad[18], David B.Goldstein[19], Philip Greenland[20], Scott J. Hebbring[21], Elizabeth W. Karlson[11], Parinda Khatri[22], Bruce Korf[23], Jordan W. Smoller[24], Stephen Sodeke[25], John Wilbanks[26], Justin Hentges[14], Christopher Lunt[14], Stephanie A. Devaney[14], Kelly Gebo[14], Joshua C Denny[14], Robert J. Carroll[2], David Glazer[4], Paul A. Harris[2], George Hripcsak[5], Anthony Philippakis[27], Dan M. Roden[1,2,28], and On behalf of the *All of Us* Research Program

[1]Departments of Medicine, Vanderbilt University Medical Center; [2] Department of Biomedical Informatics, Vanderbilt University Medical Center; [3]Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center; [4]Verily Life Sciences; [5]Department of Biomedical Informatics, Columbia University Medical Center; [6]Center for Health Information Partnerships, Northwestern University; [7]School of Biomedical Informatics, The University of Texas Health Science Center at Houston; [8]Department of Medicine, University of California Irvine; [9]School of Public Health, The University of Texas Health Science Center at Houston; [10]Department of Laboratory Medicine and Pathology, Mayo Clinic; [11]Department of Medicine, Brigham and Women's Hospital; [12]Hunter-Bellevue School of Nursing, Hunter College City University of New York; [13]Department of Biomedical Informatics, UCSD Health; [14]All of Us Research Program, National Institutes of Health; [15]Center for Health Policy & Health Services Research, Henry Ford Health System; [16]School of Public Health, University of Illinois at Chicago; [17]Cardiology Section, Department of Medicine, Veterans Administration Boston Healthcare System and Harvard Medical School; [18]Division of Preventive Medicine, University of Alabama at Birmingham; [19]Institute of Genomic Medicine, Columbia University Medical Center; [20]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University; [21]Center for Precision Medicine, Marshfield Clinic; [22]Cherokee Health Systems; [23]Department of Genetics, University of Alabama at Birmingham; [24]Department of Psychiatry and Center for Genomic Medicine, Massachusetts General Hospital; [25]Center for Biomedical Research,Tuskegee University; [26]Sage Bionetworks; [27]Broad Institute; Department of Pharmacology, Vanderbilt University Medical Center

***Address for correspondence:**

Andrea* H. Ramirez, MD, MS

1215 21st Avenue South

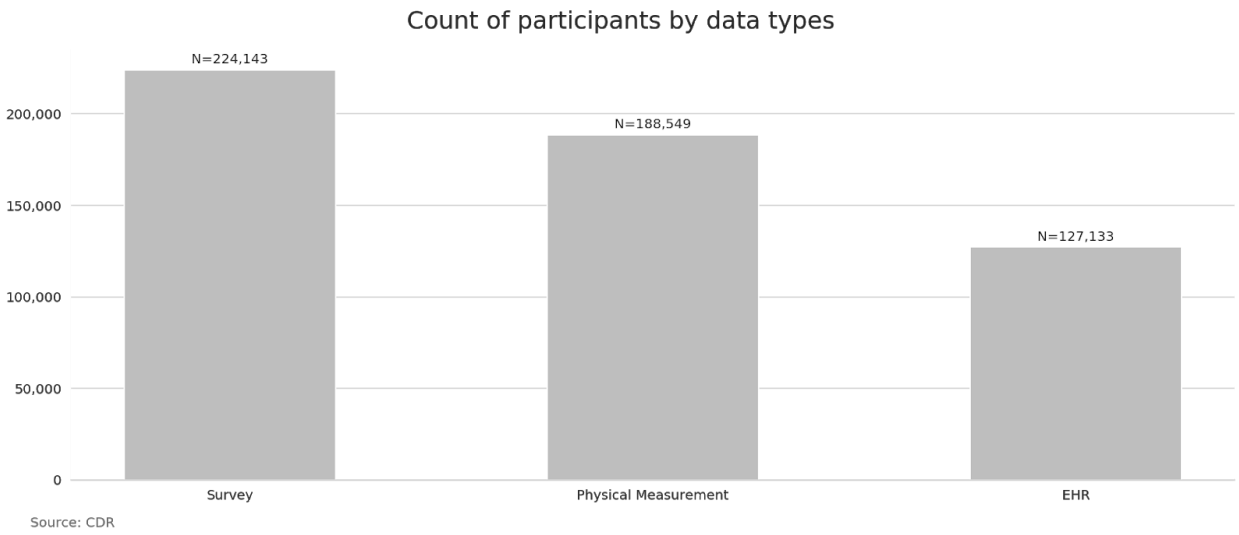
Medical Center East, South Tower

Room / Suite 8210

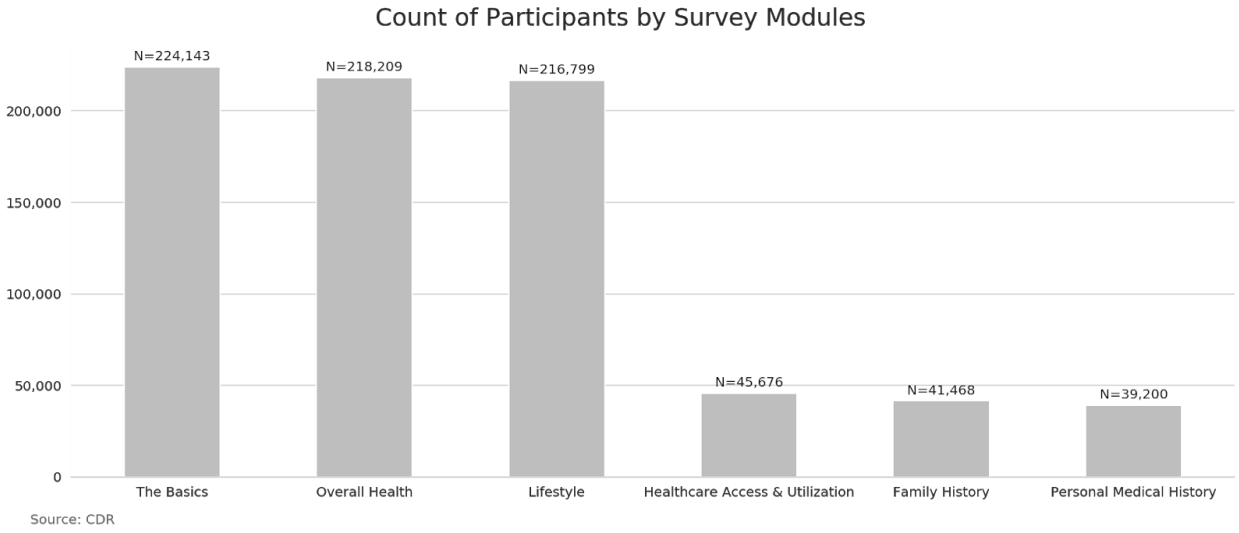
Nashville, TN 37232-8148

Supplemental Figure 1. Participant counts by data type.

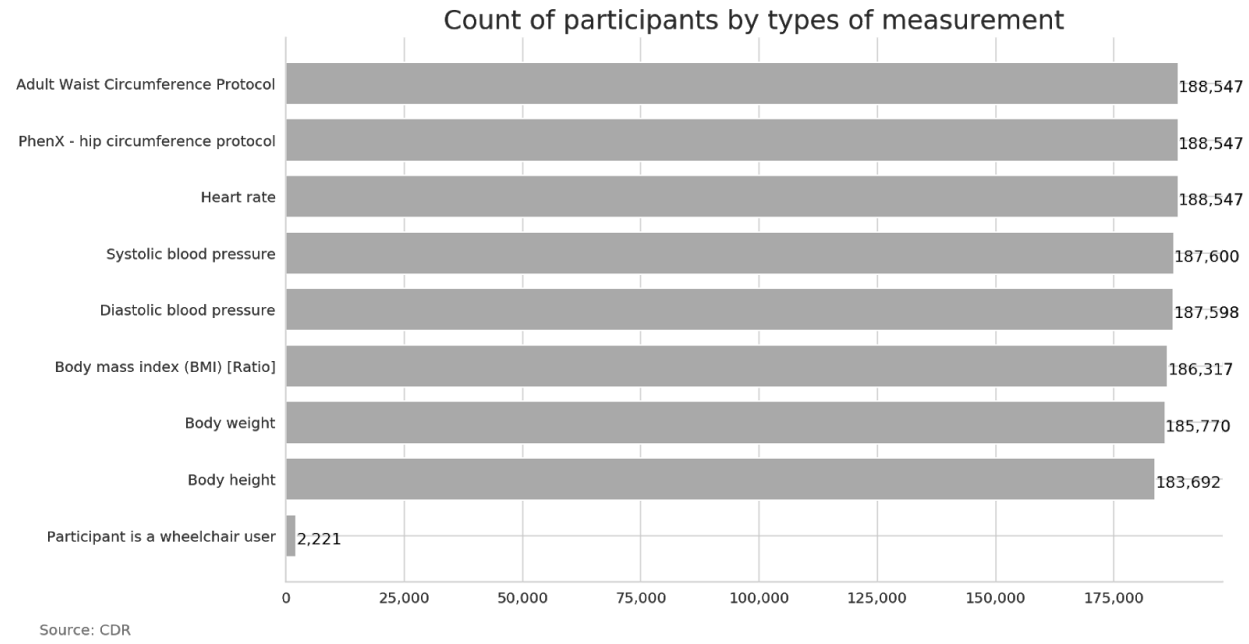
A



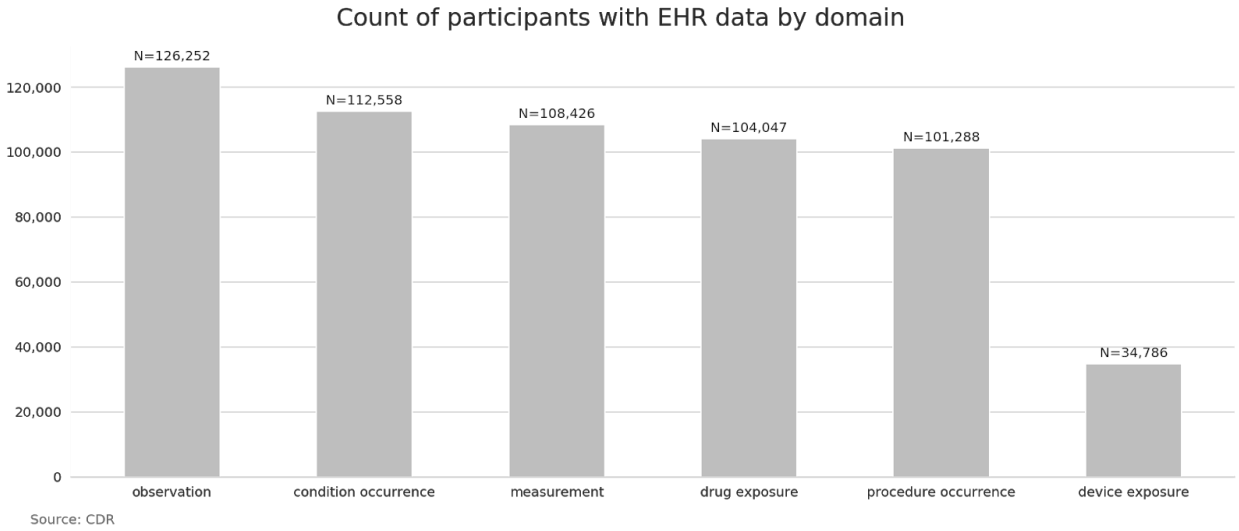
B



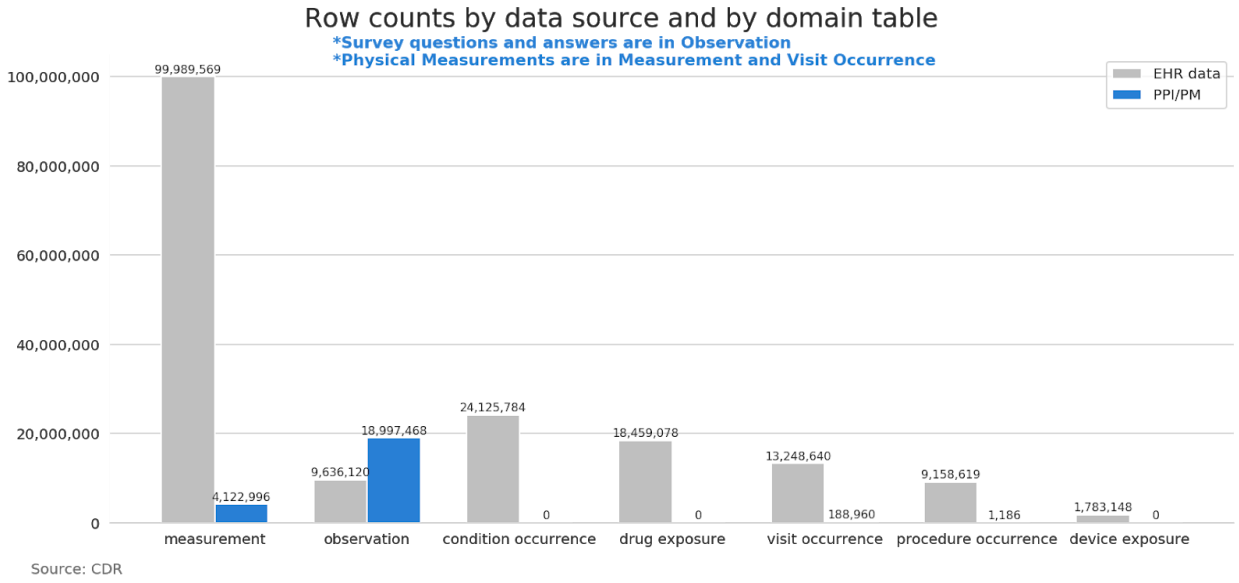
C


Count of participants by types of measurement
Adult Waist Circumference Protocol
PhenX - hip circumference protocol
Heart rate
Systolic blood pressure
Diastolic blood pressure
Body mass index (BMI) [Ratio]
Body weight
Body height
Participant is a wheelchair user
188,547
188,547
188,547
187,600
187,598
186,317
185,770
183,692
2,221
0
25,000
50,000
75,000
100,000
125,000
150,000
175,000
Source: CDR

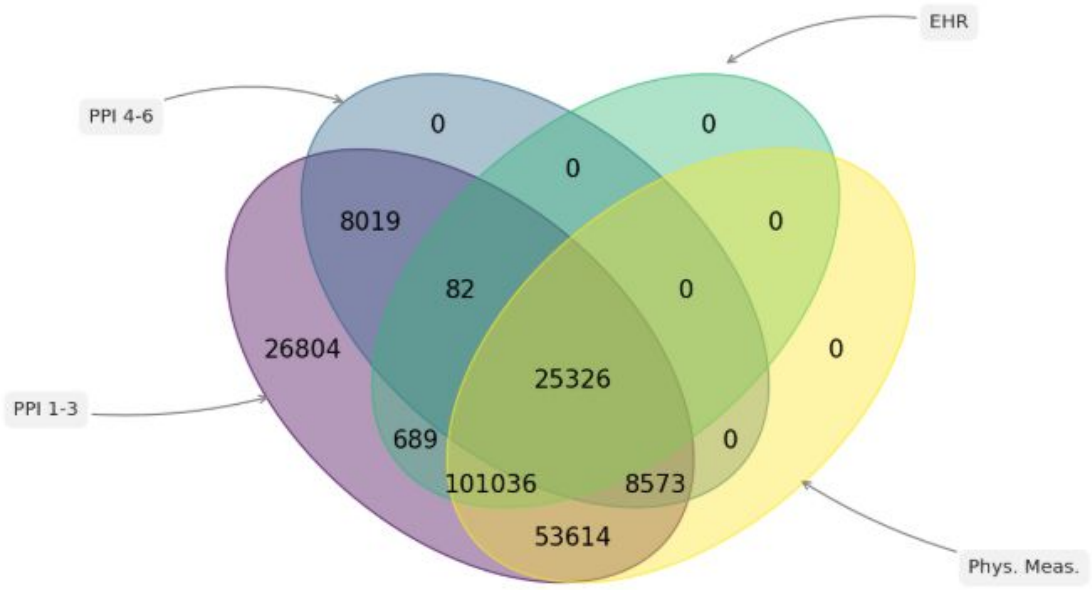
D


Count of participants with EHR data by domain
N=126,252
N=112,558
N=108,426
N=104,047
N=101,288
N=34,786
120,000
100,000
80,000
60,000
40,000
20,000
0
observation
condition occurrence
measurement
drug exposure
procedure occurrence
device exposure
Source: CDR

E


Row counts by data source and by domain table
*Survey questions and answers are in Observation
*Physical Measurements are in Measurement and Visit Occurrence
EHR data
PPI/PM
100,000,000
80,000,000
60,000,000
40,000,000
20,000,000
0
99,989,569
4,122,996
9,636,120
18,997,468
24,125,784
0
18,459,078
0
13,248,640
188,960
9,158,619
1,186
1,783,148
0
measurement
observation
condition occurrence
drug exposure
visit occurrence
procedure occurrence
device exposure
Source: CDR

F



Supplemental Figure 1: Participant counts by data type. A: Participant count by data type including EHR, PPI surveys, and physical measurements. B: Participant count grouped by PPI survey module including “The Basics”, “Lifestyle”, “Overall Health”, “Personal Medical History”, “Health Care Access and Utilization”, and “Family Medical History”. C: Participant count by different physical measurements including waist circumference, hip circumference, heart rate, systolic and diastolic blood pressures, body mass index, body weight, body height, and wheelchair use. D: Counts of participants with EHR data by domain including observation, condition occurrence, measurement, drug exposure, procedure occurence and device exposure. E: Row counts by data source and domain table grouped by EHR data (gray) and PPI and PM (blue). F: Venn diagram showing overlap of participants with data from EHR, PPI Surveys Part 1 (“The Basics”, “Lifestyle”, “Overall Health”), PPI Surveys Part 2 (“Personal Medical History”, “Health Care Access and Utilization”), and Program Physical Measurement data. By definition to be a participant included in the beta release, you must have answered PPI 1, therefore the zeros shown represent confirmation of this rule.

Supplemental Table 1. Codes used in ASCVD risk score calculation.

ischemic_heart_codes=['I20','I21','I22','I23','I24','I25','410','411','412','413','414']

stroke_codes=["430","431","432","434","434","435","I63.0","I63.1","I63.2","I63.3",

"I63.4","I63.5","I63.6","I63.8","I63.9","G45.9","I69.33","I69.34","I69.35"]

Diabetes ancestor codes: Type 2 diabetes mellitus 201826

Hypertension disease ancestor code: Hypertensive disorder 316866

HTN med treatment: Antihypertensive, Diuretics, Peripheral vasodilators, beta-blocking agents, calcium channel blockers, agents acting on the renin-angiotensin

21600381,21601461,21601560,21601664,21601744,21601782

Supplemental Table 2. Individual Site EHR Medication Sequencing participants.

EHR site*	Number of participants with anti-diabetes sequences	Number of participants with antidepressant sequences
EHR site 111	74	61
EHR site 146	110	422
EHR site 199	545	580
EHR site 211	50	120
EHR site 234	200	411
EHR site 298	222	211
EHR site 323	35	<20

EHR site 352	354	245
EHR site 387	180	307
EHR site 394	110	371
EHR site 395	20	28
EHR site 473	206	467
EHR site 499	539	1028
EHR site 601	244	676
EHR site 609	35	62
EHR site 611	72	235
EHR site 638	103	48
EHR site 662	22	37
EHR site 671	96	152
EHR site 705	765	526
EHR site 712	726	670
EHR site 721	156	172

EHR site 772	99	344
EHR site 788	160	86
EHR site 803	212	63
EHR site 824	2032	4867
EHR site 842	31	86
EHR site 863	<20	<20
EHR site 868	179	250
EHR site 906	541	637
EHR site 916	<20	53
EHR site 950	688	963
EHR site 974	623	1591

*EHR site: individual health care EHR provider

Supplemental Table 2. Overlap of participants in Smoking Exposure PheWAS.

ever_smoking_ppi ehr_smoking	Ever Smoker	Never Smoker	Unknown	All
No Codes and No EHR	34164	53833	9013	97010
No Codes and some EHR	29853	68025	2649	100527
One Code	6157	1798	316	8271
Two Codes	15675	2020	640	18335
All	85849	125676	12618	224143

Supplemental Table 3. Expanded PheWAS results for Ever smoking EHR and survey data.

The “top” phenotypes are defined as the top 10 phenome-wide significant phenotypes sorted by effect size.

EHR Ever Smoker Oncologic PheWAS

Nonprotective

phecode	description	cases	control	p_value	OR
165.10	Cancer of bronchus; lung	547	95108	2.276786e-64	4.94 (4.11, 5.95)
165.00	Cancer within the respiratory system	562	95108	5.500971e-66	4.94 (4.12, 5.92)
189.21	Malignant neoplasm of bladder	339	94889	4.220036e-13	2.36 (1.87, 2.98)
149.00	Cancer of larynx, pharynx, nasal cavities	242	95145	7.431440e-09	2.26 (1.71, 2.97)
189.20	Cancer of bladder	364	94889	1.486141e-12	2.25 (1.8, 2.82)
153.30	Malignant neoplasm of rectum, rectosigmoid jun...	466	81540	2.245329e-14	2.17 (1.78, 2.65)
155.10	Malignant neoplasm of liver, primary	189	90747	1.392056e-06	2.16 (1.58, 2.96)
155.00	Cancer of liver and intrahepatic bile duct	229	90747	7.841085e-07	2.06 (1.54, 2.74)
227.10	Benign neoplasm of adrenal gland	264	93746	2.975358e-06	1.87 (1.44, 2.43)
184.00	Cancer of other female genital organs	500	53869	3.457014e-09	1.83 (1.5, 2.23)

Protective

phecode	description	cases	control	p_value	OR
217.0	Vascular hamartomas and non-neoplastic nevi	869	84356	1.227894e-11	0.51 (0.42, 0.62)
217.1	Nevus, non-neoplastic	821	84356	1.175278e-10	0.52 (0.43, 0.64)
216.0	Benign neoplasm of skin	5803	84356	6.001888e-51	0.53 (0.49, 0.58)
228.1	Hemangioma of skin and subcutaneous tissue	943	92037	9.085644e-10	0.57 (0.47, 0.68)
228.0	Hemangioma and lymphangioma, any site	1536	92037	5.674045e-07	0.71 (0.62, 0.81)

Overall

phecode	description	cases	control	p_value	OR
165.10	Cancer of bronchus; lung	547	95108	2.276786e-64	4.94 (4.11, 5.95)
165.00	Cancer within the respiratory system	562	95108	5.500971e-66	4.94 (4.12, 5.92)
189.21	Malignant neoplasm of bladder	339	94889	4.220036e-13	2.36 (1.87, 2.98)
149.00	Cancer of larynx, pharynx, nasal cavities	242	95145	7.431440e-09	2.26 (1.71, 2.97)
189.20	Cancer of bladder	364	94889	1.486141e-12	2.25 (1.8, 2.82)
153.30	Malignant neoplasm of rectum, rectosigmoid jun...	466	81540	2.245329e-14	2.17 (1.78, 2.65)
155.10	Malignant neoplasm of liver, primary	189	90747	1.392056e-06	2.16 (1.58, 2.96)
155.00	Cancer of liver and intrahepatic bile duct	229	90747	7.841085e-07	2.06 (1.54, 2.74)
217.00	Vascular hamartomas and non-neoplastic nevi	869	84356	1.227894e-11	0.51 (0.42, 0.62)
217.10	Nevus, non-neoplastic	821	84356	1.175278e-10	0.52 (0.43, 0.64)

Survey Current Smoker Oncologic PheWAS

Nonprotective

phecode	description	cases	control	p_value	OR
165.10	Cancer of bronchus; lung	268	77672	1.968977e-19	3.52 (2.68, 4.62)
165.00	Cancer within the respiratory system	275	77672	9.589970e-19	3.38 (2.58, 4.44)
189.21	Malignant neoplasm of bladder	189	77447	4.125790e-07	2.52 (1.76, 3.6)

Protective

phecode	description	cases	control	p_value	OR
217.10	Nevus, non-neoplastic	582	69636	6.238387e-14	0.24 (0.17, 0.35)
217.00	Vascular hamartomas and non-neoplastic nevi	623	69636	1.154638e-14	0.25 (0.18, 0.36)
228.10	Hemangioma of skin and subcutaneous tissue	654	75337	3.781155e-12	0.26 (0.17, 0.38)
216.00	Benign neoplasm of skin	4237	69636	5.821750e-87	0.27 (0.23, 0.3)
173.00	Neoplasm of uncertain behavior of skin	1377	72648	1.109613e-24	0.34 (0.27, 0.42)
204.40	Multiple myeloma	204	76060	2.475480e-05	0.35 (0.22, 0.57)
172.21	Basal cell carcinoma	738	72648	1.053635e-10	0.36 (0.26, 0.49)
204.00	Leukemia	498	76060	8.905774e-11	0.36 (0.27, 0.49)
228.00	Hemangioma and lymphangioma, any site	1077	75337	2.464672e-16	0.36 (0.28, 0.46)
172.10	Melanomas of skin, dx or hx	485	72648	3.876019e-08	0.37 (0.26, 0.53)

Overall

phecode	description	cases	control	p_value	OR
217.10	Nevus, non-neoplastic	582	69636	6.238387e-14	0.24 (0.17, 0.35)
217.00	Vascular hamartomas and non-neoplastic nevi	623	69636	1.154638e-14	0.25 (0.18, 0.36)
228.10	Hemangioma of skin and subcutaneous tissue	654	75337	3.781155e-12	0.26 (0.17, 0.38)
216.00	Benign neoplasm of skin	4237	69636	5.821750e-87	0.27 (0.23, 0.3)
165.10	Cancer of bronchus; lung	268	77672	1.968977e-19	3.52 (2.68, 4.62)
165.00	Cancer within the respiratory system	275	77672	9.589970e-19	3.38 (2.58, 4.44)
173.00	Neoplasm of uncertain behavior of skin	1377	72648	1.109613e-24	0.34 (0.27, 0.42)
204.40	Multiple myeloma	204	76060	2.475480e-05	0.35 (0.22, 0.57)
172.21	Basal cell carcinoma	738	72648	1.053635e-10	0.36 (0.26, 0.49)
204.00	Leukemia	498	76060	8.905774e-11	0.36 (0.27, 0.49)

Survey Ever Smoker Oncologic PheWAS

Nonprotective

phecode	description	cases	control	p_value	OR
165.10	Cancer of bronchus; lung	592	100501	3.550585e-34	3.19 (2.65, 3.84)
165.00	Cancer within the respiratory system	607	100501	1.977280e-34	3.15 (2.62, 3.78)
180.10	Cervical cancer	247	50045	2.105232e-07	2.0 (1.54, 2.59)
189.21	Malignant neoplasm of bladder	372	100275	2.547259e-07	1.76 (1.42, 2.18)
189.20	Cancer of bladder	402	100275	6.152779e-07	1.69 (1.37, 2.08)

Protective

phecode	description	cases	control	p_value	OR
217.0	Vascular hamartomas and non-neoplastic nevi	895	89274	2.323566e-15	0.55 (0.48, 0.64)
217.1	Nevus, non-neoplastic	847	89274	6.477657e-14	0.57 (0.49, 0.66)
216.0	Benign neoplasm of skin	6116	89274	1.483150e-51	0.62 (0.58, 0.66)
228.1	Hemangioma of skin and subcutaneous tissue	1003	97279	2.766770e-10	0.65 (0.57, 0.74)
185.0	Cancer of prostate	1530	27165	1.380273e-09	0.69 (0.61, 0.78)
228.0	Hemangioma and lymphangioma, any site	1626	97279	1.232647e-11	0.69 (0.62, 0.77)
196.0	Radiotherapy	4466	84432	4.800319e-23	0.7 (0.65, 0.75)
197.0	Chemotherapy	6114	84432	2.095024e-24	0.73 (0.68, 0.77)
173.0	Neoplasm of uncertain behavior of skin	2101	93050	4.468407e-11	0.73 (0.66, 0.8)
218.1	Uterine leiomyoma	3080	53707	4.420974e-12	0.74 (0.68, 0.81)

Overall

phecode	description	cases	control	p_value	OR
165.10	Cancer of bronchus; lung	592	100501	3.550585e-34	3.19 (2.65, 3.84)
165.00	Cancer within the respiratory system	607	100501	1.977280e-34	3.15 (2.62, 3.78)
180.10	Cervical cancer	247	50045	2.105232e-07	2.0 (1.54, 2.59)
217.00	Vascular hamartomas and non-neoplastic nevi	895	89274	2.323566e-15	0.55 (0.48, 0.64)
217.10	Nevus, non-neoplastic	847	89274	6.477657e-14	0.57 (0.49, 0.66)
189.21	Malignant neoplasm of bladder	372	100275	2.547259e-07	1.76 (1.42, 2.18)
189.20	Cancer of bladder	402	100275	6.152779e-07	1.69 (1.37, 2.08)
216.00	Benign neoplasm of skin	6116	89274	1.483150e-51	0.62 (0.58, 0.66)
228.10	Hemangioma of skin and subcutaneous tissue	1003	97279	2.766770e-10	0.65 (0.57, 0.74)
185.00	Cancer of prostate	1530	27165	1.380273e-09	0.69 (0.61, 0.78)

Supplemental Table 4. ASCVD risk score demographic distributions

	All participants (n=229255)	Participants with any EHR (n= 118332)	ASCVD scores (n=26007)	ASCVD scores within a year of enrollment (n=6225)
Gender				
Male	85,645 (37.36%)	42,867 (63.27%)	8,655 (33.28%)	2,127 (34.17%)

Female	140,818 (61.42%)	74,867 (35.52%)	17,352 (66.72%)	4,098 (65.83%)
prefer not to answer	2,789 (1.21%)	1,435 (1.21%)	0 (0%)	0 (0%)
Race				
White	120,760 (52.67%)	61,348 (51.84%)	15,743 (60.53%)	3,344 (53.71%)
Black	46,679 (20.36%)	24,450 (20.66%)	5,491 (21.13%)	1,576 (25.31%)
Hispanic	42,736 (18.64%)	23,355 (19.74%)	3,198 (12.30%)	885 (14.22%)
Asian	7,602 (3.32%)	3,309 (2.80%)	546 (2.10%)	144 (2.31%)
More than one population	4,134 (1.80%)	2,023 (1.70%)	304 (1.17%)	87 (1.40%)
Another single population	1,705 (0.74%)	894 (0.76%)	139 (0.53%)	46 (0.74%)
None of these	2,401 (1.04%)	1,254 (1.06%)	250 (0.96%)	54 (0.87%)

Prefer not to answer	1,654 (0.72%)	873 (0.74%)	155 (0.53%)	47 (0.76%)
Skip	1,584 (0.70%)	826 (0.70%)	181 (0.7%)	42 (0.67%)

Supplemental Table 5. Approximate compute cost on Researcher Workbench for individual analyses, in dollars.

Analysis	Development cost	Single run cost	Total cost
Descriptive metrics	28.35	3.48	31.83
Medication sequencing	28.15	6.39	34.54
Smoking exposure PheWAS	7.00	4.20	11.20
ASCVD score calculation	11.71	7.20	18.91
Total	75.21	21.27	96.48

* Access to the Researcher Workbench and data are free. Compute and storage accrue usage cost. The Researcher Workbench uses Google Compute Engine (GCE) for computational resources in the cloud and Google Cloud Storage (GCS) for storage in the cloud. Jupyter Notebooks are loaded in a GCE Virtual machine, which is ephemeral. During this project, the

hourly rate of the default n1-highmem-4 machine was $0.27 per hour, including the cost for dataproc clusters. Notebooks not in active use are “paused” accruing cost of under $0.10 per day and shut down to zero cost after two weeks of inactivity. The storage “bucket” associated with Notebooks within a workspace costs $0.026 per GB per month.

Data Availability Links: The *All of Us* Researcher Workbench

Medication sequences for Type 2 diabetes and depression comparison by race:
https://workbench.researchallofus.org/workspaces/aou-rw-dd7cff0e/medicationspathwaysequencesbyracephase1/notebooks

Cardiovascular risk score:
https://workbench.researchallofus.org/workspaces/aou-rw-a8fc912d/duplicateofframinghamaharisksscore/notebooks

PheWAS Smoking:
https://workbench.researchallofus.org/workspaces/aou-rw-d59956e4/jamaphewasfinalreview05212020/data