Patterns, Volume *3*

Supplemental information

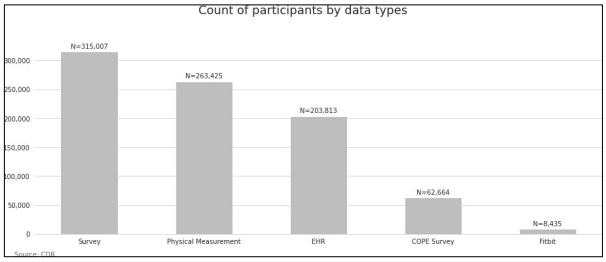
The *All of Us* Research Program: Data quality, utility, and diversity

Andrea H. Ramirez, Lina Sulieman, David J. Schlueter, Alese Halvorson, Jun Qian, Francis Ratsimbazafy, Roxana Loperena, Kelsey Mayo, Melissa Basford, Nicole Deflaux, Karthik N. Muthuraman, Karthik Natarajan, Abel Kho, Hua Xu, Consuelo Wilkins, Hoda Anton-Culver, Eric Boerwinkle, Mine Cicek, Cheryl R. Clark, Elizabeth Cohn, Lucila Ohno-Machado, Sheri D. Schully, Brian K. Ahmedani, Maria Argos, Robert M. Cronin, Christopher O'Donnell, Mona Fouad, David B. Goldstein, Philip Greenland, Scott J. Hebbring, Elizabeth W. Karlson, Parinda Khatri, Bruce Korf, Jordan W. Smoller, Stephen Sodeke, John Wilbanks, Justin Hentges, Stephen Mockrin, Christopher Lunt, Stephanie A. Devaney, Kelly Gebo, Joshua C. Denny, Robert J. Carroll, David Glazer, Paul A. Harris, George Hripcsak, Anthony Philippakis, Dan M. Roden, and the All of Us Research Program

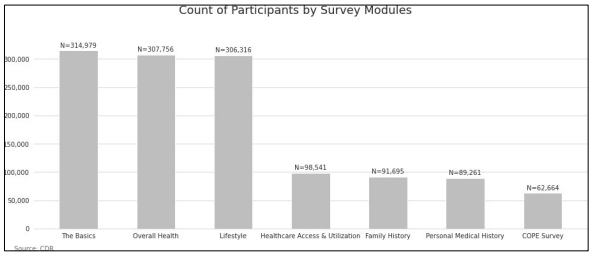
Supplemental Information

Supplemental Figure 1. Participant counts by data type.

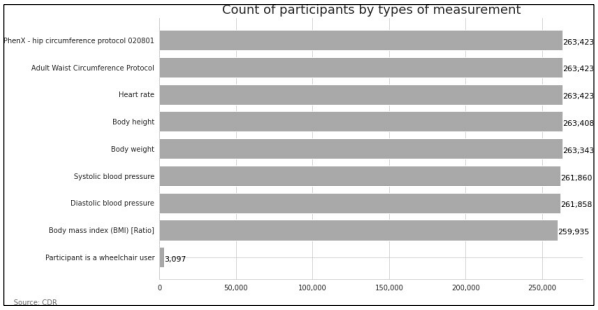
A



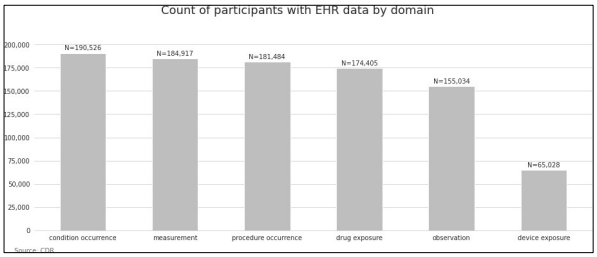
B



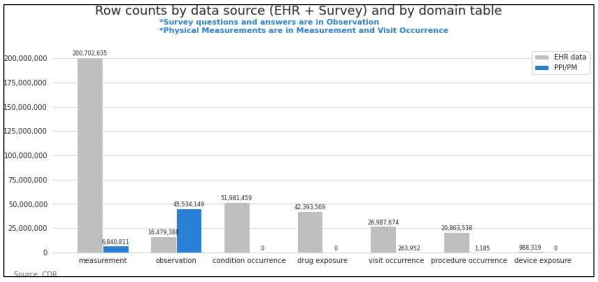
C



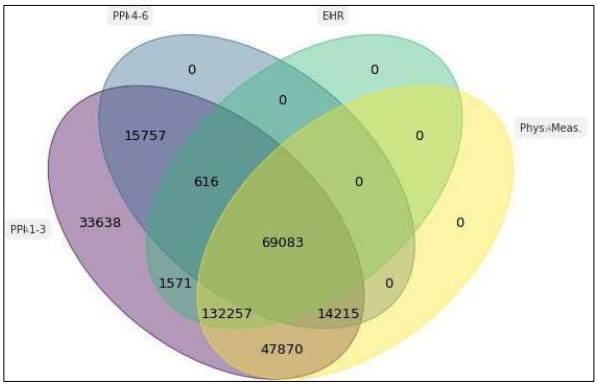
D



E



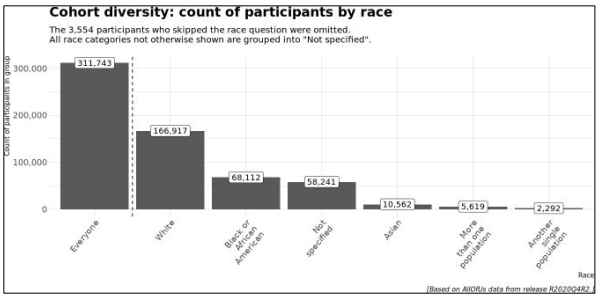
F



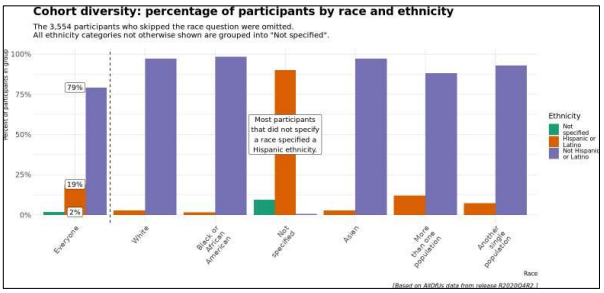
Supplemental Figure 1: Participant counts by data type. A: Participant count by data type including EHR, PPI surveys, and physical measurements. B: Participant count grouped by PPI survey module including “the Basics”, “Lifestyle”, “Overall Health”, “Personal Medical History”, “Health Care Access and 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 OMOP domain including observation, condition occurrence, measurement, drug exposure, procedure occurrence 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 included in the beta release, a participant must have answered PPI 1, therefore the zeros shown represent confirmation of this rule.

Supplemental Figure 2: Participant age, race, and ethnicity characteristics.
Supplemental Figure 2 shows responses to the question “Which categories describe you? Select all that apply”. Figure 2A describes participant responses as generalized into the permitted race categories. For example, inclusion in ‘White’ indicates no other answers in the race categories were selected. If more than one race was selected, such as ‘White’ and ‘Asian’, the participant was included in ‘More than one population’. The response ‘Hispanic’ was mapped separately to the ethnicity variable. Figure 2B shows that most participants that did not select an answer mapped as race, did select the ‘Hispanic’ response. This mapping allows use of both variables to reflect, for example, ‘Black’ and ‘Hispanic’ while other multiple responses are generalized in ‘More than one population’. Figure 2C shows the distribution of ages within each race, including a higher percentage of older participants in the ‘White’ and ‘Black’ only race. Figure 2D shows the percentage of participants by race and sex at birth demonstrating consistent higher enrollment of women in each group.

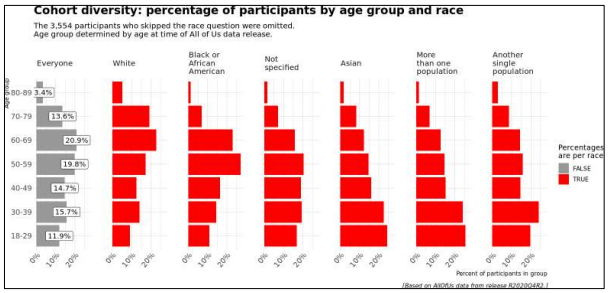
A



B



C



D

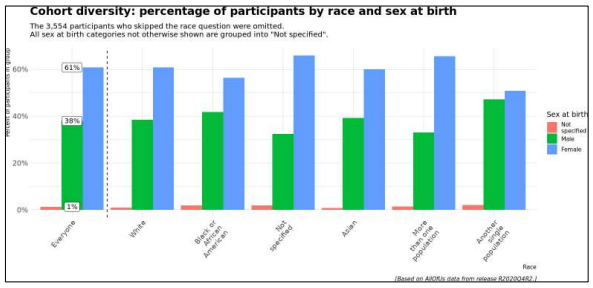


Figure 2: Participant age, race, and ethnicity characteristics. A: Count of participants by race options available in “the Basics” including White, African American or Black, Not Specified, Asian, More than one population, and Another single population. B: Percent of participants by race and ethnicity; ethnicity categories available in “the Basics” survey include “Hispanic or Latino” and “Not Hispanic or Latino”. All ethnicity categories not otherwise shown are grouped as “Not Specified”. C: Percent of participants by age group and race; for ages ranging from 18-89 years of age and race options available in “the Basics” survey D: Percent of participants by race and sex at birth; sex at birth categories available in “the Basics” survey include “Male”, “Female” and “Not specified” when category not otherwise specified.

Supplemental Table 1. Individual Site EHR Medication Sequencing participants.

EHR site	Number of participants with anti-diabetes sequences	Number of participants with antidepressant sequences
EHR site 144	205	486
EHR site 195	304	145
EHR site 199	179	243
EHR site 209	150	303
EHR site 243	29	43
EHR site 250	858	1078
EHR site 267	98	225
EHR site 305	1767	3887
EHR site 310	190	370
EHR site 314	249	134
EHR site 321	2207	1638
EHR site 392	1024	1764
EHR site 412	34	41
EHR site 433	472	1303
EHR site 481	257	80
EHR site 497	401	635
EHR site 520	321	314
EHR site 545	1027	857
EHR site 559	411	620

EHR site 567	84	246
EHR site 588	322	937
EHR site 638	613	975
EHR site 654	176	585
EHR site 657	258	324
EHR site 693	49	99
EHR site 699	2909	6631
EHR site 734	29	81
EHR site 752	962	1094
EHR site 765	892	950
EHR site 783	1093	1217
EHR site 796	20 or less	20 or less
EHR site 840	114	96
EHR site 864	251	390
EHR site 885	474	329
EHR site 925	80	20 or less
EHR site 930	334	246
EHR site 944	23	25
EHR site 990	391	1044

*EHR site: individual health care EHR provider. According to *All of Us* publishing policy, any number that is 20 or lower is suppressed and presented as “20 or less”

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

Ever Smoke PPI / Ever Smoke EHR	0.0	1.0	Unknown	All
No Codes and No EHR	67397	44860	11480	123737
No Codes and Some EHR	102235	39899	3710	145844
One Code	2848	9586	527	12961
Two or More codes	3329	28179	1247	32755
All	175809	122524	16964	315297

Supplemental Table 3. Cancer PheWAS effect sizes

PheCode	Description	EHR Ever Smoking OR (95% CI)	Survey Ever Smoking OR (95% CI)
Top 5 Increased risk effects			
149.40	Cancer of larynx	4.7 (3.25, 6.79)	NS
165.10	Cancer of the bronchus; lung	4.58 (4.01, 5.22)	3.6 (3.14, 4.13)
165.00	Cancer within the respiratory system	4.52 (3.97, 5.14)	3.49 (3.05, 3.99)
150.00	Cancer of esophagus	3.07 (2.21, 4.25)	NS
180.10	Cervical cancer	2.34 (1.9, 2.89)	1.71 (1.41, 2.08)
Top 5 Decreased risk effects			
228.10	Hemangioma of skin and subcutaneous tissue	0.35 (0.3, 0.41)	0.62 (0.56, 0.68)
217.1	Nevus, non-neoplastic	0.42 (0.37, 0.47)	0.65 (0.6, 0.71)
217.0	Vascular hamartomas and non-neoplastic nevi	0.42 (0.37, 0.48)	0.66 (0.61, 0.71)
216.0	Benign neoplasm of skin	0.43 (0.4, 0.46)	0.66 (0.63, 0.68)
213.0	Benign neoplasm of bone andarticular cartilage	0.44 (0.31, 0.63)	NS

Supplemental Table 4. 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 5. ASCVD risk score demographic distributions

	All participants (n=321,591)	Participants with any EHR (n= 200,428)	ASCVD scores within five year of enrollment (n=63,766)
Gender			
Male	122,154 (37.98%)	74,344 (37.09%)	23,912 (37.5%)
Female	195,524 (60.80%%)	123,719 (61.73%)	39,064 (61.26%)
Prefer not to answer	2,789 (1.22%)	2365 (1.18%)	790 (1.23%)
Race			
White	166,834 (51.88%)	105,280 (52.53%)	37,987 (59.57%)
Black	68,027 (21.15%)	41,049 (20.48%)	12,249 (19.21%)
Hispanic	59,247 (18.42%)	38,107 (19.01%)	9,500 (14.90%)
Asian	10,557 (3.28%)	5,577 (2.78%)	1,254 (1.97%)
More than one population	5,613 (1.75%)	3,356 (1.67%)	666 (1.04%)
Another single population	2290 (0.71%)	1,417 (0.71%)	362 (0.57%)
None of these	3338 (1.04%)	2,104 (1.05%)	609 (0.96%)
Prefer not to answer	2,176 (0.68%)	1,401 (0.70%)	365 (0.57%)
Skip	3,509 (1.09%)	2,137 (1.67%)	774 (1.21%)

Supplemental Table 6. ASCVD scores by race.
Descriptive statistics of risk factors used to calculate the ASCVD score for each race group for all participants at any time and those participants whose measurements were available within a year of enrollment in *All of Us*. SD: standard deviation, AA: African American.

	ASCVD within five years of enrollment Mean (SD), Median		
	White n=30087	AA n=9331	Other n=10564
Age	59.08 (10.01), 60	54.65 (8.7), 54	54.7 (9.57),54
Total cholesterol	190.51 (32.90), 188	186.25 (33.22), 182	189.35 (34.02), 185.5
HDL	56.26 (15.83), 54	54.15 (14.67), 52	52.00 (14.26), 50
Systolic blood pressure	126.02 (12.95), 125.5	132.20 (14.99), 131	127.07 (14.86), 126
Risk	8.47 (8.51), 5.51	9.48 (8.23), 7.31	6.37 (7.59), 3.48
Optimal Risk	4.97 (4.89), 3.201	3.33 (2.56), 2.77	3.17 (3.78), 1.68

Supplemental Table 7. 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
Medicationsequencing	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.

Supplemental File 1. Past and Present All of Us Research Program Principal Investigators

Brian Ahmedani, PhD, MSW[1]; Christine D Cole Johnson, PhD, MPH[1]; Habib Ahsan, MD, MMedSc[2]; Donna Antoine-LaVigne, PhD, MPH, MSEd[3]; Glendora Singleton[3]; Hoda Anton- Culver, PhD[4]; Eric Topol, MD[5]; Katie Baca-Motes, MBA[5]; Steven Steinhubl, MD[5]; James Wade, MD[6]; Mark Begale[6]; Praduman Jain, MSEE[6]; Scott Sutherland[6]; Beth Lewis[7]; Bruce Korf, MD, PhD[7]; Melissa Behringer, MD*[7]; Ali G Gharavi, MD[8]; David B Goldstein, PhD[8]; George Hripcsak, MD, MS[8]; Louise Bier, MS[8]; Eric Boerwinkle, PhD, MS, MA[9]; Murray H Brilliant, PhD*[10]; Narayana Murali[10]; Scott Joseph Hebbring[10]; Dorothy Farrar-Edwards, PhD[11]; Elizabeth Burnside[11]; Marc K Drezner, MD*[11]; Amy Taylor[12]; Veena Channamsetty, MD*[12]; Wanda Montalvo, PhD, RN*[12]; Yashoda Sharma, PhD[12]; Carmen Chinea, MD, MPH[13]; Nancy Jenks, MS, CFNP, FAANP[13]; Mine Cicek, PhD[14]; Steve Thibodeau[14]; Beverly Wilson Holmes, MSW[15]; Eric Schlueter, MD[15]; Ever Collier, NP*[15]; Joyce Winkler, MPH*[15]; John Corcoran, MD[16]; Nick D'Addezio[17]; Martha Daviglus, MD, PhD[18]; Robert Winn, MD*[18]; Consuelo Wilkins, MD, MSCI[19]; Dan Roden, MD, CM[19]; Joshua Denny, MD, MS*[19]; Kim Doheny[20]; Debbie Nickerson, PhD[21]; Evan Eichler[21]; Gail Jarvik, MD, PhD[21]; Gretchen Funk[22]; Anthony Philippakis, MD, PhD[23]; Heidi Rehm, PhD, MMSc, FACMG[23]; Niall Lennon[23]; Sekar Kathiresan, MD*[23]; Stacey Gabriel, PhD[23]; Richard Gibbs[24]; Edgar M Gil Rico, MBA, MSc[25]; David Glazer[26]; Joannie Grand, MSN, RN*[27]; Philip Greenland, MD[28]; Paul Harris, PhD+[29]; Elizabeth Shenkman, PhD[30]; William R Hogan, MD, MS[30]; Priscilla Igho-Pemu, MD, MSCR, FACP[31]; Cliff Pollan[32]; Milena Jorge*[32]; Sally Okun, MMHS, RN*[32]; Elizabeth W Karlson, MD[33]; Jordan Smoller, MD, ScD[33]; Shawn N Murphy, MD, PhD[33]; Margaret Elizabeth Ross, MD, PhD[34]; Rainu Kaushal, MD, MPH[34]; Eboni Winford, PhD[35]; Febe Wallace, MD[35]; Parinda Khatri, PhD[35]; Vik Kheterpal[36]; Akinlolu Ojo, MD, PhD, MPH, MBA*[37]; Francisco A Moreno, MD[37]; Irving Kron[37]; Rachele Peterson, MS*[37]; Usha Menon, PhD, RN, FAAN*[37]; Patricia Watkins Lattimore[38]; Noga Leviner[39]; Juno Obedin-Maliver[40]; Mitchell Lunn, MD, MAS, FASN[40]; Lynda Malik-Gagnon[41]; Lara Mangravite, PhD[42]; Adria Marallo[43]; Oscar Marroquin, MD44; Shyam Visweswaran, MD, PhD44; Steven Reis, MD44; Gailen Marshall, Jr., MD, PhD*[45]; Patrick McGovern[46]; Deb Mignucci[47]; John Moore[48]; Fatima Munoz, MD, MPH[49]; Gregory Talavera, MD, MPH[49]; George T O'Connor, MD, MS[50]; Christopher O'Donnell, MD, MPH[51]; Lucila Ohno-Machado, MD, PhD[52]; Greg Orr[53]; Fornessa Randal, MCRP[54]; Andreas A Theodorou, MD[55]; Eric Reiman, MD[55]; Mercedita Roxas-Murray[56]; Louisa Stark[57]; Ronnie Tepp, MPP[58]; Alicia Zhou, PhD[59]; Scott Topper, PhD, FACMG[59]; Rhonda Trousdale, MD[60]; Phil Tsao, PhD[61]; Lisa Weidman[62]; Scott T Weiss, MD, MS[63]; David Wellis, PhD[64]; Jeffrey Whittle, MD, MPH[65]; Amanda Wilson, MS[66]; Stephan Zuchner, MD, PhD[67]; Michael E Zwick, PhD[68]

Legend

*Past Principal Investigator
+ Principal Investigator/Lead Author for the *All of Us* Research Program protocol
(paul.a.harris@vumc.org)

Affiliations

1. Henry Ford Health System
2. University of Chicago Medical Center
3. Jackson-Hinds Comprehensive Health Center
4. University of California, Irvine
5. Scripps Research Translational Institute
6. Vibrent Health
7. University of Alabama at Birmingham
8. Columbia University
9. University of Texas Health Science Center at Houston
10. Marshfield Clinic Research Institute
11. University of Wisconsin at Madison
12. Community Health Center, Inc.
14. Mayo Clinic and Foundation, Rochester
15. Cooperative Health
16. EMSI
17. Blue Cross Blue Shield Association
18. University of Illinois at Chicago
19. Vanderbilt University Medical Center
20. Johns Hopkins University School of Medicine
21. University of Washington
22. FiftyForward
23. Broad Institute
24. Baylor University
25. National Alliance for Hispanic Health
26. Verily Life Sciences
27. Mitre Corporation
28. Northwestern University
29. Vanderbilt University Medical Center
30. University of Florida
31. Morehouse School of Medicine, Atlanta
32. PatientsLikeMe
33. Partners Health Care
34. Cornell University, Weill Medical College
35. Cherokee Health Systems
36. CareEvolution, Inc.
37. University of Arizona, Tucson
38. Delta Research and Educational Foundation
39. PicnicHealth
40. Stanford University
41. DXC Technology
42. Sage Bionetworks
43. Quest Diagnostics Incorporated
44. University of Pittsburgh
45. University of Mississippi Medical Center
46. Wondros

47. WebMD Health Corporation
48. Fitbit, Inc.
49. San Ysidro Health Center
50. Boston Medical Center
51. VA AoU Coordinating Center - Boston
52. University of California, San Diego
53. Walgreen Co.
54. Asian Health Coalition
55. Banner Health
56. Montage Marketing Group
57. University of Utah
58. HCM Strategists
59. Color Genomics, Inc.
60. NYC Health + Hospitals
61. VA AoU Coordinating Center - Palo Alto
62. QTC
63. Brigham and Women's Hospital
64. San Diego Blood Bank
65. Medical College of Wisconsin
66. National Library of Medicine (NLM)
67. University of Miami
68. Emory University