2024Q4R5 v8 Data Characterization Report: Overall *All of Us* Cohort Demographics

Authors

Hiral Master, Aymone Kouame, Hunter Hollis, Kayla Marginean, Justin Cook

On behalf of the Data & Research Center and the National Institutes of Health

Summary

Background and Purpose

The National Institutes of Health's (NIH) *All of Us* Research Program is a historic effort to collect and study data from a million or more people living in the United States and its territories. The goal of *All of Us* is to speed up health research discoveries, enabling new kinds of individualized health care. To make this possible, the program is building one of the world's largest and most diverse databases for health research. More details about the program can be found in the 2019 publication by *All of Us* Research Program Investigators: https://www.nejm.org/doi/full/10.1056/NEJMsr1809937 .

The *All of Us* dataset, called the Curated Data Repository (CDR), is stored on the Researcher Workbench, a secure, cloud-based platform, which was launched on May 27, 2020. The program offers tiered access to the data. Researchers' institutions must first have agreements in place with *All of Us* before they can register to use the Researcher Workbench's Registered and Controlled Tiers. Data are made available to researchers in versions as participant and linked data are collected and curated (refer to data dictionaries to learn about all the CDRs).

The Registered Tier dataset includes individual-level data from electronic health records (EHR), wearables, and surveys, as well as physical measurements taken at the time of participant enrollment. The Controlled Tier dataset includes data available on the Registered Tier, as well as genomic data and expanded demographic, survey, and EHR data. Genomic data includes short-read whole genome sequences (WGS), long-read sequences, structural variants and genotyping arrays.

The primary purpose of this report is to provide information on how to contextualize, characterize, and appropriately leverage the complex, multifaceted, and unprecedented resource that is the *All of Us* CDR. This report includes a characterization of the *All of U*s cohort as a whole, including high-level summary statistics related to demographic representation within the cohort and the availability of data. Finally, the report includes a link to access the code used to generate the reports within the Researcher Workbench as a Jupyter Notebook file.

Why is this report important to researchers/users?

This report provides a high-level summary of the available *All of Us* dataset and what may be potential biases within the data. Additionally, it provides the detailed methodology, including the code and findings that were used to generate the report using the data available in the Researcher Workbench.

Key Summary

Data from **633,547** participants, including 26,728 self-identified American Indian or Alaska Native (AI/AN) participants, are now available in CDR version 8 (v8) Controlled Tier. This is **a 53.23%** increase compared to the CDR version 7 (v7) Controlled Tier release in April 2023 (N= 413,457).

New Data types being made available in the CDRv8 which is now available to researchers starting Fall 2024

- 633,547 participants' data are now available in observation_period table, which is a derived table

and consists of 633,547 rows of data. It can be used to examine the span of time for which a participant provides available data points in all OMOP domains in CDRv8 (including EHR, survey and PM). It provides the very first (start) and the last (end) dates for each participant for any available OMOP data domains.

- 615,224 participants with self-reported survey responses about racial and ethnic subcategories, such as European, Fijian, Korean, Somali, Spanish, etc., as part of The Basics survey data. Please note: Free-text responses have been suppressed, including Tribal affiliation.
- 437,942 participants with self-reported responses to questions about disability status, as part of the data for the The Basics and Life Functioning surveys.
- 354,090 participants' data are now made available in the condition_era table, which is a derived table and consists of 5,6714,078 rows of data. It provides the span of time through which a participant is assumed to have a given condition. It provides the start and end dates of conditions recorded in participant record per condition based on SNOMED hierarchy.
- 333,030 participants' data are now made available in the drug_era table, which is a derived table and consists of 30,789,168 rows of data. It provides the span of time through which a participant is exposed to a particular active ingredient. It provides the start and end dates of conditions recorded in participant records per ingredient level.
- 63,496 participants with self-reported responses to the new Life Functioning survey.
- 62,548 participants have registered via the program's Participant Center and so utilized an alternative online portal, hosted by Scripps Research and CareEvolution to enroll in the program and provide survey data.
- 55,831 participants with self-reported height measurement.
- 55,086 participants with self-reported weight measurement.
- 40,776 participants who consented to be part of the program's Wearables Enhancing *All of Us* Research (WEAR) Study where Fitbit devices are given to participants from underrepresented communities at no cost.
- 26,728 participants who self-identify as American Indian or Alaska Native (AI/AN). Participants may self-identify with more than one category in response to the first The Basics survey question. All of Us does not ask for verification of Tribal enrollment or descendancy from participants who self-identify as AI/AN. Researchers cannot access self-reported Tribal affiliation. All of Us will never allow researchers to access participants' Tribal affiliation without approval from Tribal Nations and their respective or designated institutional review boards. Researchers should review our supplemental guide on working with data shared by AI/AN participants before working with any data from these participants.
- 19,703 participants who provided their EHRs through the Participant Portal ("participant mediated EHR data") (For additional information, refer to EHR section within "Considerations to Data Completeness in CDRv8")
- 4,546 participants with deceased status information sourced from HealthPro, a program portal for collecting participant data from program staff (see "CDRv8 data collection (data sources and approaches) considerations" for more information)

Growth in CDRv8 since the last CDR release in controlled tier in Spring 2023

- 278% increase in the number of participants with Fitbit data available (refer to Table 1.1)
- 53% increase in the number of participants with any survey data available (refer to Table 1.1)
- 53% increase in the total number of participants with data available on Researcher Workbench (refer to Table 1.1)
- 51% increase in the number of participants with physical measurements data available (refer to Table 1.1)
- 43% increase in the number of participants with genomics data available (refer to Table 1.1)

- 37% increase in the number of participants with EHR data available (refer to Table 1.1)

Who is included in CDRv8?
CDRv8 includes participants who enrolled and consented prior to October 1, 2023. Participants must complete The Basics survey to be included in the CDR. Collection of other data types is optional and may depend on other factors. For example, some participants may not consent to share their EHR data.

Considerations to Data Completeness in CDRv8

Surveys: In CDRv8, there are 63,496 responses to the new Life Functioning survey. This survey was developed to collect disability response data from participants who completed The Basics survey before disability measures were added. The Life Functioning survey items are sourced from the disability items in the American Community Survey (ACS). We recommend that researchers analyze the Life Functioning survey data along with data from the six disability measures captured in The Basics survey. Researchers can read more about the Life Functioning survey on the User Support Hub.

Electronic Health Records (EHR): The source of EHR data varies depending on how participants are engaged and enroll in the program. Participants within the catchment zone of a program-funded Health Care Provider Organization (HPO) are enrolled in the program through that organization. The transfer of EHR data for those participants is mediated by the HPO with which they are affiliated. Participants who join through an HPO may also choose to share their EHR data through the All of Us Participant Portal ("participant-mediated EHR"). To address the possibility of duplicate records when EHR data are provided from both sources, data transferred directly from the HPO are made available in the CDR, but participant mediated EHR data are suppressed. Participants who do not reside within the catchment zone of a program-funded HPO are enrolled as "direct volunteer" (DV) participants. DV participants can also provide participant-mediated EHRs. DV participants may provide access to any or all EHR data.
EHR data from either source may provide an incomplete record of care. The quality and completeness of participants' EHR data may vary.

Considerations to Data Sources or Approach used to collect data in CDRv8

Participant Portal: The participants included in this CDRv8 have been enrolled via two different participant portals. The participant portal origin flag is provided to researchers in the src_id field. Participants who have src_id = "Participant Portal: PTSC", represent that they used the Participant Technologies and Services Center (PTSC, hosted by VIbrent Health) portal to register and enroll in the program. Participants who have src_id = "Participant Portal: TPC" represent that they used the Participant Center (TPC, hosted by Scripps Research with CareEvolution) portal to register to enroll in the program. TPC is the portal through which direct volunteer participants who often reside in areas where no participating health care organization (HPO) sites are available for enrollment into the program but some overlap is possible. Once enrolled in either portal, participants may complete consents, surveys, request a salivary kit to donate biosample, EHR data, and participate in Bring-Your Own-Device (e.g., BYOD)' or the WEAR Study to share Fitbit data.

Physical Measurements (PM): The program collects PM from three sources: EHRs, an in-person visit for the collection of baseline physical measurements ("program physical measurements"), and participant-provided (self-reported) height and weight measurements. In Q2 of 2022, a survey was launched to remotely collect self-reported PM. The rationale for collecting this self-reported PM data was

to address a gap in data missingness by allowing participants who may be experiencing barriers to attending in-person clinic visits, such as Covid restrictions or mobility. For the PM data collected in the EHR, researchers should be aware that units of measure are inconsistent across HPOs, so researchers will need to normalize units. However, rates of outlier values for measures of height and weight are very low.

Fitbit: Currently, Fitbit data collected under the program's Bring-Your-Own-Device (BYOD) and WEAR Study methods are included in this CDR. There is NO separate consent process for sharing the Fitbit data under BYOD approach. However, Fitbit data collected under the WEAR study has a separate consent, which can be found in the WEAR study table. It is important to note that Fitbit data from BYOD and WEAR participants are NOT mutually exclusive. For instance, if participants withdraw from WEAR, their Fitbit data prior to withdrawal will be included in the final dataset based on the protocol. Further, they have the opportunity to share their Fitbit data under BYOD, which collects historical data, should they decide to sign up for BYOD. There are 173 WEAR study participants who provided Fitbit data (i.e., device, heart rate, sleep, OR activity) before WEAR consent start date, which is expected since they may sign up to share data under BYOD. Additionally, participants who consent to be part of WEAR study may NOT provide any Fitbit data (i.e., device, heart rate, sleep, OR activity) given device ordering workflow challenges, abandonment, system scaling issues, or participants never wore or sync their devices with the portal. In the CDRv8 Registered Tier, 9,889 participants (src_id = Participant Portal: PTSC) and 773 participants (src_id = Participant Portal: TPC) who provide WEAR consent but did NOT provide any Fitbit data, which includes device, activity, sleep or heart rate. In the CDRv8 Controlled Tier, 9,951 participants (src_id = Participant Portal: PTSC) and 779 participants (src_id = Participant Portal: TPC) who provide WEAR consent but did NOT provide any Fitbit data, which includes device, activity, sleep or heart rate.

Mortality data: All death records are now provided in aou_death table. Deceased status information is now available from 2 sources: EHR (src_id = EHR sites) and HealthPro, a program portal for collecting participant data by program staff (src_id = Staff Portal : HealthPro). Deceased status information has historically been sourced from EHR (if available). In September 2020, *All of Us* Research Program launched deceased status reporting in HealthPro. Starting in the CDRv8, this additional source of participant deceased status is made available to researchers. Program reported cause of death is collected as free text and not currently available..

Derived tables: Three derived tables, i.e., observation_period, drug_era and condition_era have been accurately populated and are now made available in CDRv8. These tables may enable users to run the queries behind the OHDSI tools (listed here) on the Researcher Workbench.

Considerations to Data Generalizability in CDRv8

All of Us seeks to build one of the world's largest and most diverse databases for health research and accelerate precision health research. Inclusion criteria for enrollment is intentionally broad.The demographic characteristics of participants with data available in the CDR may NOT entirely represent the U.S. population. Researchers should be cautious when aiming to generalize study findings to the U.S. population.

NOTE: WGS counts used for reporting purposes, refers to short-read whole genome sequence data, unless otherwise noted.

Data Availability Counts

Purpose

We provide high-level overview metrics for the number of participants in the overall CDR and by data types (**explained in table 4**) that are being made available to researchers to provide insights on data completeness and data sizes in the current release. This information will help inform researchers about potential biases that they might need to account for as they design their studies.

Key Findings

Overall, a 53.23% increase was observed in the number of participants whose data are available in CDRv8 compared to CDRv7. Details on the number of participants overall as well as by data types (defined in Table 4) in CDRv8 and growth from CDRv7 can be found in Table 1.1. There are 26,728 participants who self-identify as American Indian or Alaska Native (AI/AN) alone or in combination with one or more other categories in The Basics survey. Of 26,728 participants who self-identify as AI/AN, 100% provided any survey data (PPI), 78.51% provided PM, 61.61% provided EHR, 9.9% provided Fitbit and 66.14% provided short read WGS or array data.

There are 40,776 participants who have consented to be a part of the Wearables Enhancing *All of Us* Research (WEAR) study, which was launched by the program in which Fitbit devices were given to participants at no cost. To be eligible for the WEAR Study, participants had to be enrolled in the program and identify with one or more communities underrepresented in biomedical research. Additional requirements included completing The Basics survey and having access to a smartphone or tablet. Of 40,776 participants who consented to be part of WEAR study, 100% provide any survey data, 93.57% provide PM, 67.83% provide EHR, 72.45% provide Fitbit and 82.93% provide short read WGS or array data.

In the CDRv8 Controlled Tier, there are 30,601 participants in overall, 1,219 participants who self-identified as AI/AN and 16,401 participants who consented to be part of WEAR study provided key data types - any survey (PPI), PM, EHR, Fitbit and genomics (i.e., short read WGS OR array) (Figure 1.1a, 1.2a and 1.3). However, we acknowledge that less than 10% participants in overall CDR and in AI/AN cohort provide Fitbit data. Therefore, we removed the Fitbit data type category in overall CDR and AI/AN cohort and investigated the data availability. There are 339,452 participants overall and 13,636 participants who self-identified as AI/AN who provide any survey, PM, EHR and genomics (i.e., short read WGS OR array) data (Figure 1.1b and 1.2b).

We also provide mutually exclusive counts of participants by data types for researchers to have insights into the number of participants that provide multiple data types (refer Table 1.2). For instance, 7,444 (1.8%) participants provided all 6 data types, i.e., PPI, EHR, PM, Fitbit and Genomics data. It is important to note that completing the The Basics survey is required before participants can provide any other data types.

Table 1.3 shows the data sizes for different data types available in CDRv8. The data size for phenotypic data (Survey. EHR, PM and Fitbit) ranges from 2 GB to 6TB and total row counts ranges from 12,010,420 to 127,526,968,343.

Code used to generate the counts shown in the above tables and figure can be found here.

Demographics

Purpose

The data below provide an overview of CDRv8 participant demographics. In response to The Basics survey (survey question wording can be found here), participants may self-report demographic characteristics such as race, ethnicity, sex assigned at birth, and self-identified categories of demographic descriptors.

The race and ethnicity data in **Tables 2.1 and 2.2** were extracted from the person table. The demographic data shown in **Tables 2.3 to 2.9** were extracted from the ds_survey tables.

In **Table 2.1**, the **"not specified"** category may refer to participants who only selected ethnicity descriptors or who did not respond to the question at all. In **Tables 2.2 through 2.8, "not specified"** means a participant selected "prefer not to answer," or the category has value "none indicated" OR "no matching concept." In **Tables 2.1 through 2.8, "skip"** means the participant skipped the question. In **Tables 2.1 through 2.8**, "none of these" or "additional options" refers to participants who selected "None of these fully describe me." Participants could then provide free-text responses. Free-text responses are suppressed unless otherwise noted. Further, counts for the categories shown in **Tables 2.1 through 2.8** are mutually exclusive.

We acknowledge that participants only see one multi-select question "which categories describe you?" in The Basics survey to determine self-reported race and ethnicity. To align the research data with the collection methods, a new column named "Self-reported categories" has been added in the person table and is reported in **Table 2.8**. In future reports, we will NOT report **Table 2.1 and 2.2**.

Key Findings

Overall, participants in CDRv8, 56.45% reported being White, 79.39% were non-Hispanic or Latino, 62.50% identified as female and 20.66% were aged between 60-69 years old. Further, 22.39% reported advanced degree education, 6.81% reported annual income >$200K and 38.54% reported being employed for wages (refer to **Tables 2.1-2.7**).

These demographic characteristics were consistent for participants who provided any survey, EHR, PM, Fitbit or genomics data, given the differences between demographics characteristics for overall sample, WEAR and sample by data types was <10% (refer to **Tables 2.1-2.8**). We acknowledge that the difference of <10% threshold is arbitrary in nature and the results may vary based on a different threshold that may be used to determine the differences. Further, depending on the study design, the researchers may find that the data from the All of Us Research Program may not represent the U.S. population. Caution must be taken when generalizing the study findings. It is the responsibility of researchers to account for differences between the U.S. population and the All of Us cohort through their own analysis, if needed.

Table 2.1 Participants in current vs. previous CDR, as well as who self-identify as AI/AN and WEAR participants in current CDR by self-reported race and data types available
Table 2.2 Participants in current vs. previous CDR, as well as who self-identify as AI/AN and WEAR participants in current CDR by self-reported ethnicity and data types available
Table 2.3 Participants in current vs. previous CDR, as well as who self-identify as AI/AN and WEAR participants in current CDR by self-reported sex assigned at birth and data types available
Table 2.4 Participants in current vs. previous CDR, as well as who self-identify as AI/AN and WEAR participants in current CDR by self-reported age group and data types available
Table 2.5 Participants in current vs. previous CDR, as well as who self-identify as AI/AN and WEAR participants in current CDR by self-reported educational attainment and data types available
Table 2.6 Participants in current vs. previous CDR, as well as who self-identify as AI/AN and WEAR participants in current CDR by self-reported Income and data types available
Table 2.7 Participants in current vs. previous CDR, as well as who self-identify as AI/AN and WEAR participants in current CDR by self-reported employment and data types available
Table 2.8 All, who self-identify as AI/AN and WEAR participants in current CDR by self-identified categories of demographic descriptors and data types available

Code used to generate the counts shown in the above tables can be found here.

Genomics Data

NOTE: WGS counts used for reporting purposes refers to short-read whole genome sequence data, unless otherwise noted.

Purpose

Tables 3.1 and 3.2 provide an overview of the self-reported race and ethnicity of participants who have shared genomic data. Participants' self-reported race and ethnicity was extracted from responses to The

Basics survey. **Tables 3.3-3.5** provide counts of participants who provide genomic data in addition to other data types.

Data type definitions are available in **Table 4**.

Key Findings

In CDRv8, 52.38% of participants who shared WGS data self-identified as White, 17.15% self-identified as Black, and 3.16% self-identified as Asian (**Table 3.1**). A similar distribution was observed for participants who shared array data (**Table 3.2**). 28,960 participants shared WGS, array, EHR, PM, PPI, and Fitbit data (**Table 3.5**).

A list of acronyms is available in the **Table of Contents** and data types are defined in **Table 4**.

Table 3.1 Participant counts for WGS data by self-reported race and ethnicity
Table 3.2 Participants counts for array data by self-reported race and ethnicity
Table 3.3 Participant counts for WGS data and other data types
Table 3.4 Participant counts for array data and other data types
Table 3.5 Participant counts for WGS and array data and other data types

Code used to generate the counts shown in the above tables can be found here.

Data type Definition

Table 4 Definitions of data types and cohort shown in Tables 1.1 through 3.5

- **Total Participants:** All participants in the CDR.
- **Self-identify AI/AN:** Participants who self-identify as American Indian or Alaska Native (AI/AN). Participants may self-identify with more than one category in response to first The Basics survey question. All of Us does not ask for verification of Tribal enrollment or descendancy from participants who self-identify as American Indian/Alaska Native.
- **WEAR:** Participants who consented to be the part of the WEAR Study. The program provides WEAR participants a Fitbit device at no cost. To be eligible for the WEAR Study, participants must have already enrolled in the program and identify with one or more communities underrepresented in biomedical research. They must also meet additional requirements, including completing The Basics survey and having access to a smartphone or tablet. The enrollment to WEAR study ended on December 1, 2024.
- **WGS:** Participants with short-read whole genome sequence data available in the CDR, unless otherwise noted.
- **Array:** Participants with array data available in the CDR.
- **Physical measurements (PM):** Participants with any physical measurements available in the CDR.
- **EHR:** Participants with any electronic health record data available in the CDR.
- **Fitbit:** Participants with any Fitbit data (i.e., activity summary OR steps intraday OR heart rate summary OR heart rate minute level OR sleep summary OR sleep sequence level) available in the CDR. It does NOT include participants who provide Fitbit device data but does NOT have any other Fitbit data associated with it.

- **Surveys/PPI (participant provided information):** "Participants with data available for any survey in the CDR.

  Survey data may include responses from: The Basics; Lifestyle; Personal Medical History; Overall Health; Health Care Access & Utilization; Family Health History, Personal and Family Health History (a new survey that combines Family Health History and Personal Medical History); COPE; COVID-19 Vaccine (see line 37 below); and Social Determinants of Health surveys."