2025Q4R6 v9 Data Characterization Report: Overall *All of Us* Cohort Demographics

Authors **1**
Summary **2**
Background and Purpose 2
Why is this report important to researchers/users? 2
Key Summary 2
Considerations to Data Sources or Approach used to collect data in CDRv9 4
Considerations to Data Generalizability in CDRv9 5
Data Availability Counts **5**
Purpose 5
Key Findings 5
Demographics **7**
Purpose 7
Key Findings 7
Genomics Data **8**
Purpose 8
Key Findings 8
Data type Definition **8**

Authors

Joyce Tang, Ryan Samarakoon, Sachin Bhut, Jun Qian, Jennifer Zhang, 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. Different versions of CDR are being made available to researchers (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 was 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 Us cohort as a whole, including high-level summary statistics related to demographic representation within the cohort and the availability of data. In addition, the report provides characterizations of participants who meet criteria for classification as underrepresented in biomedical research (UBR). UBR definitions can be found here: https://www.researchallofus.org/faq/how-does-all-of-us-assess-diversity-what-communities-does-all-of-us-consider-underrepresented-in-biomedical-research/. 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 All of Us dataset that is being made available to approved researchers and what may be potential biases within the data. Additionally, it also 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 **747,028** participants, including 31,798 self-identified American Indian/Alaska Native (AI/AN) participants, is now available in CDRv9 Controlled Tier. This is a **17.91%** increase compared to the CDRv8 Controlled Tier release in February 2025 (N= 633,547).

New Data types being made available in the CDRv9 which is now available to researchers starting Summer 2026

- 119,960 participants which have completed both Mental Health and Well-Being Surveys (MHWB) - Behavioral Health and Personality (BHP) & Emotional Health History and Well-Being (EHHWB)
- 139,637 participants which have completed the Behavioral Health and Personality (BHP) survey
- 137,946 participants which have completed the Emotional Health History and Well-Being (EHHWB)
- 53,613 participants which have completed **all** four Exploring the Mind (EtM) Tasks - Delay discounting task: Now or Later (delaydiscounting), Multiracial facial emotion recognition task: Guess the Emotion (emorecog), Flanker task: Left or Right (flanker), and Gradual-onset continuous performance task: City or Mountain (gradcpt)
- 97,712 participants which have completed **any** of the four Exploring the Mind (EtM) Tasks - Delay discounting task: Now or Later (delaydiscounting), Multiracial facial emotion recognition task: Guess the Emotion (emorecog), Flanker task: Left or Right (flanker), and Gradual-onset continuous performance task: City or Mountain (gradcpt)
- 70,860 participants which have completed the Delay discounting task: Now or Later (delaydiscounting)
- 83,884 participants which have completed the Multiracial facial emotion recognition task: Guess the Emotion (emorecog)
- 66,610 participants which have completed the Flanker task: Left or Right (flanker)
- 71,402 participants which have completed the Gradual-onset continuous performance task: City or Mountain (gradcpt)
- 58,806 participants which have Participant Mediated EHR: PTSC
- 24,386 participants which have Participant Mediated EHR: TPC (CE)
- 15,371 participants which have Participant Mediated EHR: CLAD
- 62,930 participants which have Fitbit Sleep Daily Summary*
- 59,764 participants which have Fitbit Sleep Level Short
- 9,969 participants which have Proteomics
- 8,980 participants which have RNA-Seq
- 99,838 participants which have Clinical Notes

*Fitbit Sleep data was previously provided in CDRv8. The sleep_daily_summary table includes an additional field minute_to_fall_asleep along with 3 additional tables: sleep_daily_summary_30dayavg, and sleep_daily_summary_counts, and sleep_daily_summary_ext.

Growth in CDRv9 since the last CDR release in controlled tier in Fall 2024

- 16.60% increase in the number of participants with Fitbit data available (refer to Table 1.1)
- 17.91% increase in the number of participants with any survey data available (refer to Table 1.1)
- 17.91% increase in the total number of participants with data available on Researcher Workbench (refer to Table 1.1)
- 18.04% increase in the number of participants with physical measurements data available (refer to Table 1.1)
- 23.85% increase in the number of participants with genomics data available (refer to Table 1.1)
- 22.44% increase in the number of participants with EHR data available (refer to Table 1.1)

Who is included in CDRv9?

CDRv9 includes participants who enrolled and consented up to and including January 1, 2025. Participants must complete The Basics survey to be included in the CDR. Inclusion 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 CDRv9

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. DV participant mediated EHR has a source of Participant Mediated EHR: PTSC, Participant Mediated EHR: TPC, or Participant Mediated EHR: CLAD.
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 CDRv9

Participant Portal: The participants included in this CDRv9 have been enrolled via two different participant portals. We now provide the participant portal origin flag to researchers in the src_id field. Participants who have src_id = "Participant Portal: PTSC", represent that they used PTSC (Vibrent) portal to register and enroll in the program. Participants who have src_id = "Participant Portal: TPC" represent that they used CareEvolution portal to register to enroll in the program. CareEvolution is the portal through which direct volunteer participants who often reside in areas where no 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 BYOD or WEAR Study.

Physical Measurements (PM): The program collects PM from three possible 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-19 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 the Bring-Your-Own-Device (BYOD) and WEAR Study approaches are included in this CDR. There is NO separate consent process for sharing the Fitbit data under BYOD approach. However, Fitbit data collected under WEAR study has a separate consent, which can be found in 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 participants who consent to be part of WEAR study but provide any 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. 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.

Deceased Reporting: 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. Currently, program reported cause of death from HealthPro is NOT provided in the CDR as it is collected as free text.

Considerations to Data Generalizability in CDRv9

The *All of Us Research Program* seeks to enroll participants from communities that have been historically underrepresented in biomedical research in order to build one of the world's largest and most diverse databases for health research. 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, there was 17.91% increase was observed in number of participants whose data is available in CDRv9 compared to CDRv8. Details on number of participants in overall as well as by data types (defined in Table 4.1) in CDRv9 and growth from CDRv8 can be found in Table 1.1. There are 31,798 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 31,798 participants who self-identify as AI/AN, 100% provide any survey data (PPI), 78.28% provide PM, 63.32% provide EHR, 10.09% provide Fitbit and 69.45% provide short read WGS or array data.

There are 45,460 participants who have consented to be a part of WEAR study, which was launched by the program where Fitbit devices were given to participants from underrepresented communities at no cost. Of 45,460 participants who consented to be part of WEAR study, 100% provide any survey data, 94.40% provide PM, 74.36% provide EHR, 75.77% provide Fitbit and 90.52% provide short read WGS or array data.

In the CDRv9 Controlled Tier, there are 43,374 participants in overall, 1,848 participants who self-identified as AI/AN and 23,043 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% of participants in the overall CDR and slightly more than 10% 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. We saw that there 425,831 participants in overall and 17,021 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, 42,037 (5.63%) 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 CDRv9. The data size for phenotypic data (Survey. EHR, PM and Fitbit) ranges from 6TB to 13TB and total row counts ranges from 127,526,968,343 to 216,213,226,757.

Table 1.1 All participants in current vs previous CDR, as well as who self-identify as AI/AN and WEAR participants in current CDR who provide different data types
Table 1.2 All participants in current vs previous CDR, as well as who self-identify as AI/AN and WEAR participants in current CDR who provide multiple data types
Table 1.3 Overview of the data size by data types in CDRv9
Figure 1.1a: Venn diagram of participants with survey responses, EHR data, PM, Fitbit data, and genomics data available
Figure 1.1b: Venn diagram of participants with survey responses, EHR data, PM, and genomics data available
Figure 1.2a Venn diagram of participants who self-identify as AI/AN with survey responses, EHR data, PM, Fitbit data, and genomics data available
Figure 1.2b Venn diagram of participants who self-identify as AI/AN with survey responses, EHR data, PM, and genomics data available
Figure 1.3 Venn diagram of WEAR participants with survey responses, EHR data, PM, Fitbit data, and genomics data available

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

Demographics

Purpose

The data below are intended to give researchers a high-level understanding of the CDRv9 participant cohort demographics. In response to The Basics survey (survey question wording can be found here), participants may self-report the following information: race, ethnicity, sex assigned at birth, sexual orientation, gender identity, age as of data cut-off date (January 1, 2025), educational attainment, income, employment, and self-identified categories of demographic descriptors.

The demographic data shown in **Tables 2.1 to 2.6** were extracted from the person and observation tables.

In **Tables 2.1 through 2.6**, **"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 2.6, "skip"** means the participant skipped the question. In **Tables 2.1 through 2.6,** "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 2.6 are mutually exclusive.

Key Findings

Overall, participants in CDRv9, 56.00% reported being White, 78.83% were non-Hispanic or Latino, 62.62% identified as female and 20.51% were aged between 60-69 years old. Further, 22.62% reported advanced degree education, 7.10% reported annual income >$200K and 38.86% reported being employed for wages (refer to **Tables 2.1-2.6**).

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 2.1-2.6**). 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 2.8 All, who self-identify as AI/AN and WEAR participants in current CDR by self-identified categories of demographic descriptors and data types available
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 sex assigned at birth 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 age group 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 educational attainment 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 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

Code used to generate the counts shown in the above tables can be found here (Coming Soon).

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 4.1**.

Key Findings

In CDRv9, 52.42% of participants who shared WGS data self-identified as White, 16.52% self-identified as Black, and 3.38% self-identified as Asian (**Table 3.1**). A similar distribution was observed for participants who shared array data (**Table 3.2**). 42,037 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 4.1**.

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 (Coming Soon).

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 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 survey that combines Family Health History and Personal Medical History); COPE; COVID-19 Vaccine (see line 37 below); Social Determinants of Health surveys; Exploring the Mind; Behavioral Health and Personality; and Emotional Health History and Well-Being."