2022Q4R9 v7 Data Characterization Report: Overall *All of Us* Cohort Demographics

Authors

Hiral Master, Aymone Kouame, Hunter Hollis, Kayla Marginean, Kelsey Rodriguez

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

Summary

Background and Purpose

In July 2016, the National Institutes of Health gave initial funding to create the *All of Us* Research Program. The program strives to nurture relationships with participants, build a robust ecosystem of communities and researchers, and to deliver the largest and most diverse biomedical dataset. More details about the program can be found in the 2019 publication by *All of Us* Research Program Investigators. Importantly, registered users can access the *All of Us* Research Program's researcher-facing resource, i.e., the Curated Data Repository (CDR), via the Researcher Workbench, which is a secured cloud-based platform.
Researcher Workbench was launched on May 27, 2020 and the data is available to approved researchers. Data in the Controlled Tier (available to Registered Tier researchers since March 2022), contains genomic data and more granular demographic data compared to the data available in the Registered Tier (available to registered users since May 2020).
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). Lastly, it also provides the code used to generate reports, in the form of Jupyter Notebooks.

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 on the Researcher Workbench.

Key Summary

Data on 413,457 participants is now available in v7 controlled tier CDR. Thus ~11% growth compared to v6 controlled tier CDR (N= 372,397)
310,738 participants in V7 CDR meet the criteria for at least one UBR category. Thus ~11.4% growth compared to v6 controlled tier CDR (N= 278,882)

New Data types being made available in the v7 CDR which is now available to researchers starting

April 2023

- 33,049 participants provide Winter COPE survey
- 14,908 participants provide Fitbit Sleep Data
- 1,027 participants provide Long Read Whole Genome Sequence data (LrWGS), which were generated from biospecimens.
- 11,390 participants provide Structural Variant callset data, which were generated from short read WGS.

Growth in v7 CDR since the last CDR release in controlled tier in June 2022

- 89.57% increase in the number of participants who provide genomics data
- 21.51% increase in the number of participants who provide Fitbit data
- 11.42% increase in the number of participants meeting at least one UBR criteria
- 11.07% increase in the number of participants who provide EHR data
- 11.03% increase in the number of participants, which are being made available
- 10.73% increase in the number of participants who provide any survey data
- 10.24% increase in the number of participants who provide physical measurements data

Who is included in v7 CDR?

Participants in the *All of Us* program are only required to submit the primary consent and complete "The Basics" survey to have data included in the Curated Data Repository. Inclusion of any other data types is optional and may also depend upon a variety of additional considerations. For instance, completion of all other surveys, including Overall Health and Lifestyle (part of core surveys) is NOT mandatory for participants to be included in CDR. Therefore, completion rates for surveys vary by survey type. Furthermore, v7 CDR included participants who were enrolled in the *All of Us* Research Program and provided consent from May 2018 to July 1, 2022.

Considerations to Data Completeness in v7 CDR

Surveys: In this v7 CDR

97 (0.02%) do not have "The Basics" in the CDR, which are *currently under investigation*. These participants in the CDR had no rows for the Basics and therefore, are not present in either obs or ds_survey tables. Further, 3,936 participants (0.95%) had rows for the Basics but have 100% PMI Skip, no matching concept, null any combination of those things. Researchers need to be aware of these counts and adjust criteria according to the study design.

Electronic Health Records (EHR): The process of sharing EHR data with the *All of Us* Program varies depending on the type of participant, i.e., enrolled via Healthcare Provider Organization (HPO) vs. Direct Volunteer (DV). Currently, we only provide EHR data for the participants who are associated with an HPO funded by program to the researchers on the workbench. EHR records for participants who enroll via DV or see providers at non-funded HPOs have NOT been included. Therefore, EHR records *may not provide a complete record of care*. Additionally, the completeness of records that are submitted for inclusion in the CDR may vary depending on the process used to extract data from EHR vendors.

Considerations to Data Sources or Approach used to collect data in v7 CDR

Physical Measurements (PM): The program collects PM from two sources: EHR and, for patients paired with an Healthcare Provider Organization (HPO), an in-person visit for the collection of baseline physical measurements ("program physical measurements"). 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) approach is included in this CDR. There is NO separate consent process for sharing the Fitbit data under BYOD approach.

Considerations to Data Generalizability in v7 CDR

The *All of Us* Research Program intentionally over-samples UBR categories as part of its goal to provide one of the most diverse databases in existence. Therefore, researchers may observe that demographic characteristics of CDR data may not represent the US population, and thus researchers/users should be cautious if their study aims to generalize the findings to the US population.

We provide high-level overview metrics for the number of participants in overall CDR and by data types (**explained in table 6.1**) that are being made available to researchers/users to provide insights on data completeness 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 growth of 11.03% is observed in the number of participants whose data is available from April 2023 to June 2022. Details on the number of participants overall as well as by data types**(defined in Table 6.1)** in v7 CDR and growth from v6 CDR can be found in **Table 1.1**.

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

Data Availability Counts

Purpose

We also provide mutually exclusive counts of participants by data types for researchers/users 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. However, there are some participants who provide no data or 1 data type other than survey. In this CDR, 97 (0.02%) participants do not have "The Basics" in the CDR, which are currently

under investigation. These participants in the CDR had no rows for the Basics and therefore, are not present in either obs or ds_survey tables. Therefore, researchers need to be aware of these counts and adjust criteria according to the study design.

Table 1.1 Count of all participants in current vs. previous CDR who provide different Data types
Table 1.2 Count of all participants in current vs. previous CDR who provide multiple Data types
Figure 1.1: Count of participants in v7 CDR by data types

Code used to generate the counts shown in the above tables and figure can be found here:
https://workbench.researchallofus.org/workspaces/aou-rw-b3b105ae/dataqualityreportsc2022q4r9v7cdr/notebooks/preview/1.%20Summary%20of%20Participants%20By%20Data%20Type.ipynb

Demographics

Purpose

We provide high-level overview metrics to give researchers a high-level understanding of the demographic characterization for overall CDR and by data types (**explained in table 6.1**). Participants available in the CDR are characterized using following measures, which are extracted from the Basics Surveys: race, ethnicity, sex, gender identity, age at data cut-off date (i.e., July 1, 2022), educational attainment, income, and employment.

Key Findings

Overall, participants in v7 CDR, 55.42% reported being White, 78.16% were non-Hispanic or Latino, 60.36% identified as female and 21.06% were aged between 60-69 years old. Further, 21.55% reported advanced degree education, 6.5% reported annual income >$200K and 36.93% reported being employed for wages (refer to **Tables 2.1-2.8**).
These demographic characteristics were consistent for participants who provided any survey, EHR, PM, or genomics data, given the differences between demographics characteristics for overall sample and sample by data types was <10% (refer to **Tables 2.1-2.8**). However, the demographic characteristics for participants who provided Fitbit were not consistent with the overall characteristics (i.e., difference >10%). Specifically, participants who provided Fitbit data more frequently reported being White (80.94%), non-Hispanic or Latino (89.56%), more educated (37.36% with advanced degree), having higher annual incomes (11.63% reported >$200K), and being employed (51.99% reported employed for wages).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/users may find that the data from the *All of Us* Research Program may not represent the US population. Caution must be taken when generalizing the study findings. In turn, it is the responsibility of researchers to account for differences between the US population and the *All of Us* cohort through their own analysis, if needed.

Code used to generate the counts shown in the above tables can be found here: https://workbench.researchallofus.org/workspaces/aou-rw-b3b105ae/dataqualityreportsc2022q4r9v7cdr/notebooks/preview/2.%20Demographic%20Characteristics%20of%20Participants%20by%20Data%20Type.ipynb

UBR

Purpose

We provide an overview on UBR metrics for researchers to have a high level understanding on the diversity of the sample that is available on Researcher Workbench. Primarily, it is important that the *All of Us* Research program intentionally over-samples participants who are underrepresented in biomedical research. Thus, the data from the *All of Us* Research program should not be viewed as representative of the US population. The definitions for UBR categories were derived from prior work published by Mapes el al.

Key Findings

In v7 CDR, overall, 75.16% of participants (N=310,738) met at least one criteria of UBR (under-represented in biomedical research) definition. There was a 11.42% increase in the number of participants meeting at least one UBR criteria compared to June 2022 CDR, which had 278,882 participants meet at least one UBR criteria. For more details refer to **Table 3.1**.

Code used to generate the counts shown in the above tables can be found here: https://workbench.researchallofus.org/workspaces/aou-rw-b3b105ae/dataqualityreportsc2022q4r9v7cdr/notebooks/preview/3.%20UBR%20Breakdown.ipynb

Genomics Data

NOTE: WGS counts used for reporting purposes refers to short read whole genome sequence data

Purpose

In **tables 4.1 and 4.2**, we provide high-level overview metrics to give researchers a high-level understanding of the self-reported race/ethnicity breakdown for participants who provide genomics data (see the data type explained in **table 6.1**). The information on race and ethnicity was extracted using the Basics Survey. In **tables 4.3-4.5**, we also provide counts of participants who provide genomic data in addition to other data types, which are explained in **table 6.1.**

Key Findings

In v7 CDR, ~51%, 20% and 3% of participants who provide WGS data self reported being White, Black and Asian, respectively (refer to **table 4.1**). Similar distribution of race/ethnicity were observed for participants who provided array data (refer to **table 4.2**). 7,444 participants provided any WGS and Array AND any EHR AND and PM AND any PPI AND any Fitbit (refer to **table 4.5**) data types as defined in **table 6.1**.

Table 4.1 Participants counts who provide WGS data by self-reported race/ethnicity
Table 4.2 Participants counts who provide array data by self-reported race/ethnicity
Table 4.3 Count of Participants with WGS data and other data types (**please see note highlighted below)
Table 4.4 Count of Participants with array data and other data types
Table 4.5 Count of Participants with WGS data and array data

Code used to generate the counts shown in the above tables can be found here:
https://workbench.researchallofus.org/workspaces/aou-rw-b3b105ae/dataqualityreportsc2022q4r9v7cdr/notebooks/preview/5.%20Genomics%20by%20Race%20and%20Datatypes.ipynb

UBR by Datatype

Purpose:

We provide an overview on UBR metrics for overall CDR and/or by data types (explained in **table 6.2**) for researchers to have a high level understanding on the diversity of the sample that is available on Researcher Workbench. Primarily, it is important that the *All of Us* Research program intentionally over-samples participants who are underrepresented in biomedical research. Thus, the data from the *All of Us* Research program should not be viewed as representative of the US population.

Key Findings:

In v7 controlled tier CDR, overall, 75.16% of participants (N=310,738) met at least one criteria of UBR (under-represented in biomedical research) definition (refer to **Table 5.2**). These overall UBR counts were similar by PM, EHR, and survey data types as defined in **table 6.2** (i.e. the difference of <15%). However, the total count of participants meeting at least 1 UBR criteria was 53.53% (refer **Table 5.2**). We acknowledge that the difference of <15% threshold is arbitrary in nature and the results may vary based on a different threshold that may be used to determine the differences.
Similar findings were observed in Registered Tier CDR (refer **Table 5.1**)

Table 5.1 Participants counts who are classified as UBR (under-represented in biomedical research) in registered tier CDR v7 (R2022Q4R9)
Table 5.2 Participants counts who are classified as UBR (under-represented in biomedical research) in controlled tier CDR v7 (C2022Q4R9)

Code used to generate the counts shown in the above tables can be found here: https://workbench.researchallofus.org/workspaces/aou-rw-b3b105ae/dataqualityreportsc2022q4r9v7cdr/notebooks/preview/4.%20UBR%20by%20Datatypes.ipynb

Datatype Definition

We have used following definitions for data types in Tables 1.1 to 4.5
All Participants or Overall: All people in the CDR.
WGS: Participants with short read WGS data in current beta release, unless otherwise noted
Array: Participants with Array data in current beta release
Physical measurements (PM): Participants with any physical measurements in the CDR
EHR: Participants with any EHR data in the CDR
Fitbit: Participants with any FITBIT data (heart rate, activity summary, sleep and steps) in the CDR
Surveys/ PPI (participant provided information): Participants with any surveys refers to: 'The Basics', 'Lifestyle', 'Personal Medical History', 'Overall Health', 'Healthcare Access & Utilization', 'Personal and Family Health History' (new survey which is the combination of 'Family Health History' and 'Personal Medical History'), 'COPE', 'COVID-19 vaccine' and 'Social determinant of Health' surveys.

Data Types Definition for UBR by data types shown in Tables 5.1 and 5.2
All Participants: All people in the CDR.
WGS: Participants with short read WGS data in current beta release
Array: Participants with Array data in current beta release
Physical measurements: Participants with any physical measurements in the CDR
EHR: Participants with any EHR data in the CDR
Fitbit: Participants with any FITBIT data (heart rate, activity summary, sleep and steps) in the CDR
Surveys:
PPI Surveys 1 to 6: Participants with any data in surveys: 'The Basics', 'Lifestyle', 'Personal Medical History', 'Overall Health', 'Healthcare Access & Utilization', 'Personal and Family Health History' (new survey which is the combination of 'Family Health History' and 'Personal Medical History')

COPE Surveys: Participants with any data in the COPE surveys (any version), however, counts for the questions related to COVID-19 vaccines are not-included in the COPE survey.
COVID-19 Vaccine: Participants with any data in the Minute Survey on COVID-19 Vaccines OR any COVID vaccine related questions in COPE survey
Social Determinants of Health Survey: Participants with any data in the Social Determinants of Health Survey