# 2022Q2R2 v6 Data Characterization Report: Overall AoU Cohort Demographics

## Table of Contents

# Authors

Hiral Master, Sofia Labrecque, Aymone Kouame, Kayla Marginean, Kelsey Rodriguez

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

# 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[1]. 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, 2021 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). For more details, refer to Appendix 1: Controlled vs. Registered Tiers. The CDR for both tiers are refreshed bi-annually. These refreshes include data from additional participants, data updates for existing participants, and the expansion and/or addition of data types across all enrolled participants.

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.

## Completeness

One key consideration for researchers/users who wish to leverage participant data across multiple data types is the number of participants who have records across the full breadth of data types required for their study. 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, 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, only participants associated with an HPO are able to share their EHR data.* The technology for DV participants to share EHR data is under development. When the technology is available, DV participants will authorize the transfer process for their records. For participants who enter the program via an HPO and provide their consent to share EHR data, the HPO will transfer their data to the DRC. However, only HPOs who are funded by the program submit data from their institution for inclusion in the CDR. Participants who see providers at non-funded HPOs in the area where they live will not have those records 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. Records of care provided to the program may be incomplete. Therefore, participants may have varying representation of data types across their full records in the program.

## Generalizability

Another key consideration for researchers is related to generalizability. 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.

# Important Takeaways

- We present this data characterization report to ensure our stakeholders (researchers, participants, and consortium members) can meaningfully interpret the current state of data offerings available to registered users.
- On June 22**,** 2022, data on 372,397 participants were made available to researchers who have access to controlled tier data. There has been **a 12.4% increase** in the number of participants since the data in the controlled tier was first launched in March 2022.
- In the current June 2022 release, 273,756 (73.5%) participants met at least one UBR criteria. There was a 10% increase in the number of participants meeting at least one UBR criteria in current CDR compared to controlled tier data released in  March 2022, which had 247,578 participants meeting at least one UBR criteria.

- The June 2022 release is considered a data “refresh” and major changes include the addition of new survey data from the Program’s Social Determinants of Health (SDOH) and COVID-19 Minute surveys in both Controlled and Registered Tiers.
- Additionally, select EHR-derived COVID-19 vaccine concepts were unveiled within the Registered Tier in the June 2022 refresh. Details on the concept ids that were unveiled can be found by clicking this link.

# Detailed Report

## Methodological Highlights

**Rationale**: We first provide the detailed methodology for researchers to understand protocols that were employed by the program to collect the data. Further, the section below also provides the details on methodology that was used to create this report, to help researchers/users. These methods (including code) can be used by researchers/users for replication purposes or may provide insights into developing their study design.

**The controlled tier (C2022Q2R2, version 6) dataset that was released on June 22, 2022 was used to generate this report.** Specifically, it included participants who were enrolled in the *All of Us* Research Program from May 2018 to January 1, 2022. The data is available to registered Researcher Workbench Users. The total participant counts in the controlled and registered tiers are slightly different (<1% difference) due to privacy and cleaning rules (refer to Appendix 1) that differ across the two datasets. Participants can provide the following data types:

1. **Participant Provided Information (PPI)** is data collected via surveys. When a participant enrolls in the *All of Us* Research Program, they complete core surveys (The Basics, Lifestyle and Overall Health). The participant may choose to complete additional surveys as they become available throughout the life of the program, including Family History, Personal Medical History, Healthcare Access & Utilization surveys, Social Determinants of Health, COVID-19 Participant Experience (COPE) Survey and COVID-19 Minute surveys. For more details refer to Appendix 3.
2. **Physical Measurements (PM)** are measurements taken at the time of participant enrollment including height, weight, body mass index, and waist circumference. For more details refer to Appendix 4.
3. **Electronic Health Records (EHR)** refers to the digital version of a patient's medical history that is maintained by the provider over time. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is used to standardize all EHR data and includes data for the following domains: Measurement, Procedure, Observation, Drug Exposure, Device Exposure, and Condition. For the purpose of this report, participants providing data for at least one of the EHR domains are considered to have EHR data. For more details refer to Appendix 5.

4. **Digital Health Technologies (DHT)** refers to data from Digital Health Technologies. Currently, only data from **Fitbit** is available to the users on the Researcher Workbench. For more details refer to Appendix 6.
5. **Genomics Data** refers to WGS (Whole Genome Sequence) and array data, which are available to registered users.  For more details refer to Appendix 7.

All the data processing and analyses used to generate this report were conducted within the Jupyter Notebook (Python programming language) on the Researcher Workbench. Detailed methodology (including the code used to clean the data and generate results) can be accessed by registered Researcher Workbench users by clicking this link to the workspace. Specifically, the workspace consists of following jupyter notebooks:

1. Summary of Participants By Data Type
2. Demographic Characteristics of Participants By Data Types
3. UBR Breakdown of Participants

# Key Insights

## Overall

**Purpose and Rationale**:
We provide high-level overview metrics for the number of participants 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 highlights:**
It is important to note that completing the “The Basics” survey is required before participants can provide any other data types. Overall, data on **372,397** participants who have (at minimum) completed “The Basics” survey are available in the Controlled Tier. Of these participants who provide “The Basics” survey,

- **306, 179** (82.2%) participants additionally provided PM or Physical Measurement data, which consist of measurements taken at the time of participant enrollment including height, weight, body mass index, and waist circumference
- **258,145** (69.4%) participants additionally provided data in at least one of the EHR or Electronic Health Records domains: Visits, Conditions, Procedures, Measurements, Drug Exposures, and Device Exposure.
- **165,072** (44.3%) participants additionally provided genotyping array variant data and 98,558 (26.3%) participants  provided WGS (Whole Genome Sequence) data.
- **12,844** (3.5%) participants additionally provided Fitbit data, i.e.,  data in at least one of the heart_rate_summary, heart_rate_minute_level, steps_intraday, and activity_summary tables.

- The above number of participants for the data types (PM, EHR, Array, WGS and Fitbit) are **NOT** mutually exclusive.

Overall, a growth of **12%** was seen in the number of participants whose data is available from June 2022 to March 2022. For more details refer to Table 2a in Appendix 2. Therefore, researchers/users now have access to more participants' data.

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. For instance, 2,732 (0.83%) participants provided **all 6 data types**, i.e., PPI, EHR, PM, Fitbit and Genomics data.

There were 79,425 (21.3%) participants who provided **only** 5 data types. Specifically,

- 1,800 (0.48%) participants provide all phenotypic data, which includes PPI, EHR, PM and Fitbit. Currently, the Fitbit data provided in June 2022 CDR release, represents the data collected via Bring Your Own device Program ONLY. To clarify, only participants who own Fitbit, consent to share their EHR data and would like to contribute their Fitbit data to the program are being made available to registered researchers. This may be one potential reason for low counts when examining combination for multiple data types, where Fitbit data is additionally included.
- 75,854 (20.37%) provided Genomics (WGS and Array) and phenotypic data, which includes PPI, EHR and PM data only.

There were 76,456 (20.5%) participants who provided **only** 4 data types. Specifically,

- 1,800 (0.5%) participants **only** have PPI, EHR, PM and Fitbit data and have **NOT** provided any Genomics data
- 55,346 (14.9%) of participants provided Array, EHR, PM and PPI data **only** while 2,352 (0.63%) participants provided WGS, EHR, PM and PPI data **only**. It is important to note that the program strived to provide array data for every WGS given array generations is quicker than WGS. Therefore, it can be seen that the number of participants with array data is higher compared to those with WGS only. However, sometimes array generation fails for not meeting the call rate threshold of 0.98, even though these passed clinical call rate QC at the GCs. Hence there are around 2,000+ participants who provide WGS but do not provide array data. In addition, WGS samples are not a strict subset of the array samples. There are possible duplicate samples in both WGS and Arrays. The independent maximal set of related samples are provided as auxiliary data on the Research Workbench for WGS. Researchers should consider removing these samples, based on the recommendation from this genomic data quality report, which is published on the User Support Hub.

There were 124,709 (33.5%) participants who provided **only** 3 data types. Specifically,

- 113,117 (30.38%) **only** provided PPI, EHR and PM
- 255 (0.07%) **only** provided PPI, EHR and Fitbit

There were 30,922 (8.3%) participants who provided **only** 2 data types. Specifically,

- 23,246 (6.24%) **only** provided PPI and PM
- 3,884 (1.04%) **only** provided PPI and EHR
- 3,460 (0.93%) **only** provided PPI and Fitbit

Lastly, 58,153 (15.6%) participants **only** have PPI and have not provided any other data types, which seems reasonable since providing "The Basics Survey" data is sufficient for credit for this data type.

***For more details see the jupyter notebook Summary of Participants by Data Type.***

## Demographics

**Purpose and Rationale**:
We provide high-level overview metrics to give researchers a high-level understanding of the demographic characterization. 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., January 1, 2022), educational attainment, income, and employment.

**Key highlights:**
Overall, participants who completed the Basics survey (PPI), 54.0% reported being White, 77.4% were non-Hispanic or Latino, 60% identified as female and 21% were aged between 60-69 years old. Further, 20.9% reported advanced degree education, 6.2% reported annual income >$200K and 36.3% reported being employed for wages.

These demographic characteristics were consistent for participants who provided PPI, EHR, PM, or genomics data, given the differences between demographics characteristics for overall sample and sample by data types was ≤10% . 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%), non-Hispanic or Latino (88.1%), more educated (37.2% with advanced degree), having higher annual incomes (11.4% reported >$200K), and being employed (52.0% 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 AoU cohort through their own analysis, if needed.

***For more details see the jupyter notebook labeled "Demographic Characteristics of Participants By Data Types"***

## UBR

**Purpose and Rationale**:
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.

**Key highlights:**
Overall, **73.5%** of participants (N=273,756) met at least one criteria of UBR (under-represented in biomedical research) definition. These participants were classified as UBR. Those who do not meet the definition of UBR are considered RBR (represented in biomedical research). The definitions for UBR categories were derived from prior work published by Mapes el al[2].

Participants who are UBR in race and ethnicity categories have the strongest representation in the CDR (almost 50/50 split to those traditionally represented by biomedical research [RBR] participants). Overall, however, RBR participants have much higher representation than UBR participants in the cohort do based on individual categories, including in the income, age, sexual and gender minority, and education categories;however, over 80% of all participants are included in any of the individual UBR groups.

Overall, 273,756 participants met the UBR criteria in June 2022 CDR. Thus, there was a **10% increase** in the number of participants meeting at least one UBR criteria compared to March 2022 CDR, which had 247,578 participants meet at least one UBR criteria. For more details refer to Table 2b in Appendix 2. Of all the demographic categories, disparities in representation were most prominent in the race/ethnicity, income and age categories. To further elaborate,

- 43.8% participants (N=163,277) were classified as UBR who self-reported their race as non-white race or their ethnicity as Hispanic/Latino.
- 25.9% participants (N=96,324) were classified as UBR who self-reported their annual household income as less than $25k.
- 24.4% participants (N=91,026) were classified as UBR with age >= 65 years at the time of consent.
- <10% of participants were classified as UBR based on definitions for educational attainment (i.e., less than GED or high school degree), sex (i.e., self-reported their biological Sex at Birth as intersex) and Sexual & Gender Minorities (i.e., self-reported their biological Sex at Birth as neither male nor female, their gender identity as non-binary or different than biological Sex at Birth, or their sexual orientation as non-straight sexual).

***For more details see the jupyter notebook labeled "UBR Breakdown"***

# References

1. All of Us Research Program Investigators. (2019). The "All of Us" research program. New England Journal of Medicine, 381(7), 668-676.
2. Mapes, B. M., Foster, C. S., Kusnoor, S. V., Epelbaum, M. I., AuYoung, M., Jenkins, G., ... & All of Us Research Program. (2020). Diversity and inclusion for the All of Us research program: A scoping review. PloS one, 15(7), e0234962.

# Appendices

## 1. Controlled vs. Registered Tiers

The Controlled Tier includes all of the information available in the Registered Tier, as well as genomic data, additional clinical fields from EHRs, and additional demographic data that are suppressed or generalized in the Registered Tier.

The Controlled Tier will also have more granular data on UBR (underrepresented in biomedical research) populations, as well as precise date and geolocation information. In particular, the Controlled Tier provides access to more accurate individual-level information about the participant (see Table 1).

For more details on differences in access and privacy models for registered and controlled tier data can be found in the *All of Us* Controlled Tier Dataset v6 CDR Release Notes.

Table 1. Data Privacy Model for Registered and Controlled Tiers

| **Data Element** | **Registered Tier** | **Controlled Tier** |
|---|---|---|
| Explicit identifiers | Suppress | Suppress |
| Free text fields in survey and unstructured clinical documents | Suppress | Suppress |
| Dates (of events) | Random shift<br>*Backward by a random number between 1 to 365* | As Collected (unshifted) |
| Date of Birth | Random shift<br>*Backward by a random number between 1 to 365* | Generalize to year of birth |
| Date of Death | Random shift<br>*Backward by a random number between 1 to 365* | As Collected (unshifted) |
| Data of participants age >89 | Suppress | As Collected |

| | | |
|---|---|---|
| Geolocation | Generalize to US state | Generalize to first 3 digits of zip code |
| Marital status | As Collected | As Collected |
| Living situation<br>*PPI (survey): Where are you currently living?* | Suppress | As Collected |
| Own or rent | As Collected | As Collected |
| Higher level Race/Ethnicity<br>*Eg: Asian, White, Black, MENA etc* | Generalize | As Collected |
| Race/Ethnicity subcategory<br>*Eg: Hmong, Fillipino, Caribbean* | Suppress | Suppress |
| Sex at birth (PPI)* | Generalize | As Collected *<br>*Includes all branching logic questions* |
| Gender identity (PPI) | Generalize | As Collected *<br>*Includes all branching logic questions* |
| Sexual orientation (PPI) | Generalize | As Collected *<br>*Includes all branching logic questions* |
| Race/Ethnicity (EHR) | Suppress<br>*Value from EHR is suppressed to harmonize with PPI data* | As Collected |
| Sex/Gender (EHR) | Suppress<br>*Value from EHR is suppressed to harmonize with PPI data* | As Collected |
| ICD codes indicative of suppressed sex/gender<br>*List of codes here* | Suppress | As Collected |
| **Data Element** | **Registered Tier** | **Controlled Tier** |
| Education | Generalize | As Collected |
| Employment status | Generalize | As Collected |
| Annual household income | As Collected | As Collected |
| Death cause<br>*i.e., Death cause noted in the EHR, including relevant diagnosis codes* | Suppress | As Collected |
| Diagnosis codes subject to public knowledge<br>*List of codes here* | Suppress | As Collected |
| ICD Codes indicative of motor vehicle accidents<br>*ICD9 E80*-E84*, ICD10 V** | Suppress | Suppress |

| | | |
|---|---|---|
| Active duty military status | Suppress | As Collected |
| Born in US or not | As Collected | As Collected * |
| Genomic data<br>*Includes program-generated Whole Genome Sequencing and Array data* | Suppress | As Collected |
| COVID EHR data | Suppress<br>*COVID EHR drug concepts with valid OMOP start dates of 1/1/2021 or later are suppressed. Concepts with default 1/1/1970 start dates are also suppressed.* | As Collected ^ |

*Note: "As Collected" indicates that there will be no change to the data for the purpose of privacy protection.*
**Free text responses will be suppressed.*

## 2. Change Metrics

The following tables provide the overall participants counts by data types as well as those meeting UBR definitions using the controlled tier data that was launched in March 2022 vs. the current data refresh in June 2022.

### Table 2a. Participant counts by data types for the data released in June 2022 (Current CDR) vs. March 2022 (Previous CDR)

| | **Count of Participants in Current CDR (C2022Q2R2)** | **Count of Participants in Previous CDR (C2021Q3R6)** | **change** | **change%** |
|---|---|---|---|---|
| Total Participants in CDR | 372,397 | 331,382 | +41,015 | +11% |
| | | | | |
| **BY DATA TYPES** | | | | |
| Fitbit | 12,844 | 11,681 | +1,163 | +9% |
| WGS | 98,558 | 98,622 | -64 | -0% |
| Array | 165,072 | 165,208 | -136 | -0% |
| EHR | 258,415 | 224,413 | +34,002 | +13% |

| | | | | |
|---|---|---|---|---|
| PM | 306,179 | 269,702 | +36,477 | +12% |
| PPI | 372,397 | 331,362 | +41,035 | +11% |
| None | 0 | 20 | -20 | -100% |

Table 2b. Participant counts meeting UBR and RBR definitions for the data released in June 2022 (Current CDR) vs. March 2022 (Previous CDR)

| | **Count of Participants in Current CDR (C2022Q2R2)** | **Count of Participants in Previous CDR (C2021Q3R6)** | **Change** | **Change%** |
|---|---|---|---|---|
| Total Participants in CDR | 372,397 | 331,382 | +41,015 | +11% |
| **BY DIVERSITY CATEGORIES** | | | | |
| UBR Overall | 273,756 | 247,578 | +26,178 | +10% |
| RBR Overall | 98,755 | 83,875 | +14,880 | +15% |
| UBR Race Ethnicity | 163,277 | 150,508 | +12,769 | +8% |
| RBR Race Ethnicity | 209,120 | 180,874 | +28,246 | +14% |
| UBR Sex | 218 | 210 | +8 | +4% |
| RBR Sex | 372,179 | 331,172 | +41,007 | +11% |
| UBR Sexual And Gender Minorities | 35,364 | 31,199 | +4,165 | +12% |
| RBR Sexual And Gender Minorities | 337,297 | 300,345 | +36,952 | +11% |
| UBR Age | 91,026 | 80,034 | +10,992 | +12% |
| RBR Age | 281,371 | 251,348 | +30,023 | +11% |
| UBR Education | 34,241 | 32,319 | +1,922 | +6% |
| RBR Education | 338,156 | 299,063 | +39,093 | +12% |

| | | | | |
|---|---|---|---|---|
| UBR Income | 96,324 | 90,568 | +5,756 | +6% |
| RBR Income | 276,073 | 240,814 | +35,259 | +13% |

## 3. Participant-Provided Information: Surveys

Participant-provided information is collected via surveys and is intended to augment data collected from other sources, such as EHRs. Surveys are developed and deployed through a process that includes prioritization of scientific domains, content creation (based on sourcing of items from well-established studies and the literature), pilot evaluation and refinement but not independent validation of the combined instruments, and scheduled deployment (Cronin et al). Surveys are deployed on the secure online *All of Us* Participant Portal. Initial surveys include questions on participants' demographics, health, and lifestyle. Additional surveys on more specific subjects are added regularly. Survey content and information about source instruments are available on the *All of Us* Research Hub Survey Explorer.

The following survey modules have been implemented and are included in the June 2022 Registered and Controlled Tier CDR:

- The Basics
- Overall Health
- Lifestyle
- Health Care Access and Utilization
- Personal and Family Health History
- COVID-19 Participant Experience Survey (COPE)
- Minute Survey on COVID-19 Vaccines
- Social Determinants of Health (SDOH)

## 4. Physical Measurements

The program collects physical measurements from two sources: EHRs and, for patients paired with an Healthcare Provider Organization (HPO), an in-person visit for the collection of baseline physical measurements ("program physical measurements"). A characterization of physical measures captured in EHRs is provided in the EHR section of this report. The metrics below focus on physical measures collected as part of the in-person program enrollment process.

Program physical measurements include height, weight, waist and hip circumferences, heart rate, and blood pressure and help provide a more complete picture of participant's health. These measurements are collected only once, unless an error with the measurements is identified and the participant must return for a re-measurement. As part of the enrollment process after site pairing, the Participant Portal offers eligible participants the option to schedule an in-person appointment at their paired site. At that visit, biospecimens are collected and program physical measurements are collected by site staff using a clinical application called HealthPro.

Participants must first pair with a site in the Participant Portal to qualify for program physical measurements. Participants who sign up through an HPO are paired with their respective HPO sites. Direct Volunteers (DVs) are paired with the site closest to their own address.

For donors choosing to donate whole blood, a maximum of two blood pressure and heart rate measurements are taken, with the second measurement taken only if the first measurement is out of range. For apheresis donors, these measurements and weight measurements are taken when the blood is drawn.

The online portal alerts participants when they enter values that fall outside of normal ranges. The alerts are customized to address each physical measurement, and users have the option to override the alert (and provide necessary medical attention if it is an emergency) and enter the value if it falls within the validation range (which is determined by physical plausibility).

# 5. Electronic Health Records

The *All of Us* Research Program provides grants to a number of Healthcare Provider Organization (HPO) that are funded to recruit and enroll participants as well as to transfer EHR data for participants who have consented to provide it. These awardees include large Regional Medical Centers (RMCs), Federally Qualified Health Centers (FQHCs), and the Department of Veterans Affairs (VA). The Program has also funded a number of HPOs that are not responsible for the recruitment and enrollment of participants but only for the transfer of EHR data for eligible, consented participants. For participants who enter the program via an HPO and provide their consent to share EHR data, the HPO will transfer EHR data it has on paired participants to the DRC. When the technology is available, DV participants will authorize the transfer process for their records.

Because data aggregation and transfer occurs at the organizational rather than the awardee level, data provenance within the CDR is indicated at the organizational level. It is important for researchers to understand the provenance of EHR data, as there may be implications that should be taken into account within their study designs.

Due to the coordination across the many participating organizations, EHR data is collected from awardees quarterly. Each submission has a data cutoff date. Data should be current up to the cutoff date for each deadline and should include all eligible, consented participants up to that date. The data cutoff dates for the quarterly submissions are as follows:

- Q1: January 1
- Q2: April 1
- Q3: July 1
- Q4: October 1

# 6. Digital Health Technologies

The use of digital health technologies such as Fitbits has risen exponentially over the past decade. The data from these technologies can be used to gain further insights on participant's health. All participants who have already provided primary consent to be part of *All of Us* also have an opportunity to provide the digital health data, which are collected via  Fitbit and Apple wearable devices. There are plans to expand the data collection via other wearable devices.

The program employs the Bring-Your-Own-Device (BYOD)  approach to collect Fitbit and iOS HealthKit data. However, the program has also initiated the WEAR study, which provides Fitbits to participants who do not already own a device. Currently, the work on incorporating the WEAR data into the Raw Data Repository is in process; the data will be made available to researchers in future date.

There is currently no separate consent process for sharing the digital health data under BYOD; the data sharing is contingent on participants' devices syncing steps. Participants can log onto the *All of Us* Participant Portal and visit the Sync Apps & Devices tab. Participants without Fitbit devices can also take part by creating a free Fitbit account online and manually adding information to share with the program. Participants can choose what type of data to share and can stop sharing at any time; choosing to sync or link a device with *All of Us* is considered implied consent for sharing that digital health data type. Unsyncing or unlinking a device is considered implied withdrawal of that device; the data is no longer collected and transferred to DRC. If participants resyncs or relinks the device, the data collection will start over.

Fitbit devices provide heart rate (at the minute level), heart rate summary statistics, steps taken in a day (at the minute level), and summaries of daily activity and sleep. Users can also supply data on daily weight and daily food, water, and macronutrient consumption through Fitbit. The PTSC delivers BYOD Fitbit data files to a Google Bucket in the RDR on a daily basis. Typically, one .json file represents one data type for one day for one participant. Currently, the file contents are then parsed into a series of tables for data types, including:

a. Heart rate (by zone summary)
b. Heart rate (at the minute level)
c. Activity: daily summary
d. Activity: intraday steps (at the minute level)

Fitbit minimum viable product (MVP) was launched in December 2020, with data on approximately 9,000 participants. In September 2021 when data was refreshed, data on approximately ~11,600 users who had provided data as of April 1, 2021 became available to Researcher Workbench users in the Registered Tier. In March 2022, the data became available in the controlled tier as part of the controlled tier launch.

In the Registered Tier, person_ids are replaced with research_ids and all datetime fields are subject to date-shifting and data truncation based on participant age, in line with EHR data protocols. However, the Fitbit data in the Controlled Tier is not date-shifted and truncated by participant age. In both the Registered and Controlled Tiers, Fitbit data are provided as supplemental (non-OMOP) tables similar to cb_ and ds_ datasets.

# 7. Genomics

The first release of *All of Us* Research Program's genomic data contains 165,208 array samples and 98,622 whole genome sequencing (WGS) samples. The June 2022 release contains variant call format (VCF) files, Hail MatrixTables and auxiliary data (only for WGS) such as joint callset QC information, relatedness, genetic predicted ancestry, and variant annotations. Quality control processes were performed both independently and across samples.

Genomic data QC is split into three conceptual areas:

1. **Consistency** -- The uniformity of protocols at each GC that reduce the probability of batch effects and that normalize the data across GCs.
2. **Single Sample QC** -- QC processes run for each sample independently. These catch major errors, such as sample swaps or sample contamination.
3. **Joint Callset QC** (WGS only) -- QC processes executed on the joint callset, which uses information across samples to flag samples and filter variants.

All arrays and WGS have passed fingerprint concordance checks.