The *All of Us* Research Study - Data Types, Data Access, and Data Analysis

SITC-NCI Computational Immuno-Oncology Webinar Series
December 7, 2022

Sheri D. Schully, Ph.D.
Deputy Chief Medical and Scientific Officer
*All of Us* Research Program
All of Us Research Program Mission

Our Mission

Accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us

Nurture partnerships for decades with at least a million participants who reflect the diversity of the U.S.

Deliver one of the largest, richest biomedical datasets that is broadly available and secure

Catalyze an ecosystem of communities, researchers, and funders who make All of Us an indispensable part of health research

Made possible by a team that maintains a culture built around the program’s core values
**Value**

**WHAT**

is the *All of Us* Research Program?

It will be **one of the largest, most diverse health databases** in history

**WHY**

should I care?

It combines **biological factors and social determinants** on a large, inclusive scale

**WHO**

benefits?

It equips researchers to make discoveries that **enable more precise approaches to care**, informing providers' recommendations & individuals' choices

**HOW**

do I access it?

It is now **available to researchers with institutional signoff** across a wide range of settings and career stages with just a computer and internet connection
## Current protocol

### Enroll, Consent and Authorize EHR
- Recruiting 18+ years old initially; plan to include children in future
- Online, interactive consent
- Includes authorization to share EHR data

### Answer Surveys
- **The Basics**
- **Overall Health**
- **Lifestyle**
- **Health Care Access & Utilization**
- **Personal and Family Medical History**
- **Social Determinants of Health**

**Additional surveys will be released on an ongoing basis**

### Provide Physical Measurements
- **Blood pressure**
- **Heart rate**
- **Height**
- **Weight**
- **BMI**
- **Hip circumference**
- **Waist circumference**

**Based on diverse sampling and capacity**

### Provide Biosamples
- **Blood (or saliva)**
- **Urine specimen**
- **Biosamples will be stored at the program’s biobank**

**Based on diverse sampling and capacity**

### Share data from Wearables/Digital Apps
- Share data from wearable fitness devices, starting with Fitbit

**Coming soon:**
- Integrated apps to track mood & cardio-respiratory fitness
Consent and EHR Authorization

• Participants must be 18 years or older
• Online video consent
• Consent process includes authorization to share EHR data with researchers

<table>
<thead>
<tr>
<th>Initial Data Types</th>
<th>Expanded Data Types (May Include)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Demographics</td>
<td>• Clinical notes</td>
</tr>
<tr>
<td>• Visits</td>
<td>• Radiology, cardiology, and other reports</td>
</tr>
<tr>
<td>• Diagnoses</td>
<td>• Mental health reports</td>
</tr>
<tr>
<td>• Procedures</td>
<td>• Substance abuse, alcohol use, and tobacco use</td>
</tr>
<tr>
<td>• Medications</td>
<td>• More laboratory results, potentially including genomics</td>
</tr>
<tr>
<td>• Laboratory visits</td>
<td></td>
</tr>
<tr>
<td>• Vital signs</td>
<td></td>
</tr>
</tbody>
</table>

Participant Surveys

Physical Measurements

Biosamples

Mobile/Wearable Tech
Participant surveys

**Open Surveys:**
- The Basics
- Overall Health
- Lifestyle
- Health Care Access & Utilization
- Personal & Family Health History
- Social Determinants of Health

**Closed Surveys:**
- COVID-19 Participant Survey Experience (COPE)
- Minute Survey on COVID-19 Vaccines

Additional surveys will be released on an ongoing basis
Physical measurements

Current Measurements:
• Blood pressure
• Heart rate
• Height
• Weight
• BMI
• Hip circumference
• Waist circumference
Biosamples

Examples:
- Blood
- Saliva
- Urine
Mobile/Wearable technology

Current Data:
• Data from wearable fitness devices, including Fitbit and Apple HealthKit
• Additional integrations in the future
• Data includes:
  ○ Heart Rate by Zone Summary
  ○ Heart Rate (Minute-Level)
  ○ Activity (Daily Summary)
  ○ Activity Intraday Steps (Minute-Level)
All of Us Data
Different data types available

Surveys
from 372,000+ participants
Including lifestyle, access to care, medical history, and data from nearly 100,000 participants on their experiences during the COVID-19 Pandemic

Physical Measurements
from 306,000+ participants
Blood pressure, heart rate, BMI and more

Electronic Health Records
from 258,000+ participants
Standardized to OMOP common data model

Genomics
98,600+ whole genome sequences
Only available via the Controlled Tier

Genomics
165,000+ genotyping arrays
Only available via the Controlled Tier

Wearables
(from 12,800+ participants)
Physical activity and heart rate from participants who have connected their Fitbit devices

Participant biosamples are forthcoming
Including the first batch of genomic data

Genomics

Only available via the Controlled Tier

98,600+ Whole Genome Sequences

165,000+ Genotyping Arrays

593,500,000+ Unique Variants
Represented and searchable through public genomic data browser

Genomics Analysis Tools
Hail and PLINK in addition to R, Python, and Jupyter Notebooks

Genomic Data is Paired with Rich Phenotypic Data

77,000+
Have Whole Genome Sequences + Electronic Health Records + Physical Measurements + Survey Responses

95,000+
Have Whole Genome Sequences + Physical Measurements + Survey Responses

78,200+
Have Whole Genome Sequences + Electronic Health Records

3,500+
Have Whole Genome Sequences + Fitbit Records Representing >30% of all participant Fitbit records
Selected Data Snapshots: Enrollment Progress

(Updated 12/4/22)

Participants: 558,000+
Participants who have completed initial steps of the program: 393,000+
Electronic Health Records: 339,000+
Biosamples: 411,000+

Enrollment Numbers

May 2017  March 2020  Oct 2022
Commitment to All of Us Participant Diversity

Geographic Diversity

Race and Ethnicity

- White
- Black, African American or African
- Hispanic Latino or Spanish
- Asian
- Other
- Prefer not to say

Data as of 12/4/22

Diversity
Includes racial and ethnic minorities as well as sexual and gender minorities, people with low income or limited education, and other groups.

50%+
Racial and Ethnic Minorities

80%+
Underrepresented in Biomedical Research
Accessing Data and Research Tools
Three Tiers

Public Tier
Contains only anonymized, aggregate data which is available to anyone through the Data Browser and Data Snapshots. Public Tier tools include — Data Browser, Research Projects Directory, Publications, Data Snapshots, and the Survey Explorer.

Registered Tier
Contains curated, anonymized, individual-level data which is available to registered researchers on the Researcher Workbench. Registered Tier tools include — Cohort Builder, Dataset Builder, Workspaces, Notebooks, and the Support Hub.

Controlled Tier
Available to approved researchers who have taken additional steps and training to access these data. Controlled Tier data includes: Genomic data, additional clinical fields in electronic health records, and additional demographic data from surveys that are suppressed or generalized in the Registered Tier.

*Counts between all three data tiers and Data Snapshots may vary because of the lag time associated with the data curation process.
How the Research Hub Works

1. Participants share their data with the All of Us Research Program through multiple sources. These data are sent to a secure cloud environment, managed by the Data and Research Center.

2. Participant data is received and funneled through a curation pipeline within a secure repository that connects to the Research Hub tools.

3. Anyone can visit the Research Hub to learn more about the types of data All of Us makes available for research. The Survey Explorer and Data Browser offer more information about the unique data elements and let visitors browse aggregated participant data.

4. Researchers register and apply for access to the Researcher Workbench to analyze data.

5. Registered researchers in the Researcher Workbench can create research projects using collaborative workspaces, cohort-building tools, interactive notebooks, and more.

6. Research underway can be viewed on the Projects Directory. Publications related to All of Us Research Program data are posted on the Publications page.
Data passport model

- *All of Us* employs a “data passport” model for the Registered and Controlled tiers, to grant researchers broad permission to explore the data for a wide range of studies.

- To facilitate collaboration and keep participant data secure, the Researcher Workbench is hosted in a Cloud-based system.
<table>
<thead>
<tr>
<th>Year</th>
<th>Data release timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>Data Browser (launched May 2019)</td>
</tr>
<tr>
<td></td>
<td>Interactive tool available to the public that provides summary statistics from the program’s growing database, allowing researchers to understand the characteristics of our participant population, <strong>explore the data types available and plan research questions</strong></td>
</tr>
<tr>
<td>2020</td>
<td>Researcher Workbench (launched May 2020)</td>
</tr>
<tr>
<td></td>
<td>Released of the first version of the All of Us dataset and the first set of researcher tools for beta testing</td>
</tr>
<tr>
<td></td>
<td>- Survey data</td>
</tr>
<tr>
<td></td>
<td>- Physical measurement data</td>
</tr>
<tr>
<td></td>
<td>- EHR data</td>
</tr>
<tr>
<td>2021</td>
<td>Expansion of data and tools (throughout 2021)</td>
</tr>
<tr>
<td></td>
<td>Expanded of the All of Us dataset with additional survey, measurement, and EHR data, including:</td>
</tr>
<tr>
<td></td>
<td>- <strong>Start of genotyping</strong> and whole genome sequencing</td>
</tr>
<tr>
<td></td>
<td>- <strong>Initial return of results</strong> to participants</td>
</tr>
<tr>
<td></td>
<td>&gt;1,000K researchers registered</td>
</tr>
<tr>
<td>2022</td>
<td>Controlled Tier &amp; Genomics (ongoing)</td>
</tr>
<tr>
<td></td>
<td>Launch our Controlled tier, available to approved researchers who have taken additional steps and training to access:</td>
</tr>
<tr>
<td></td>
<td>- <strong>Genomics data</strong></td>
</tr>
<tr>
<td></td>
<td>- <strong>Individual and granular-level</strong> information Available in Registered and Controlled tiers, with basic information available via our public data browser:</td>
</tr>
<tr>
<td></td>
<td>- ACS data linkage</td>
</tr>
<tr>
<td></td>
<td>- New surveys</td>
</tr>
<tr>
<td></td>
<td>&gt;1,000K researchers registered</td>
</tr>
<tr>
<td>2023 and beyond</td>
<td>Additional data types and tools (ongoing)</td>
</tr>
<tr>
<td></td>
<td>Expect to enroll 1M participants within 5-7 years of launch and continue to expand the All of Us dataset with new data releases, including over time:</td>
</tr>
<tr>
<td></td>
<td>- <strong>Assay data</strong></td>
</tr>
<tr>
<td></td>
<td>- <strong>Additional wearables data</strong></td>
</tr>
<tr>
<td></td>
<td>- Apple Healthkit</td>
</tr>
<tr>
<td></td>
<td>- <strong>New surveys</strong> (e.g., Mental Health &amp; Wellbeing)</td>
</tr>
<tr>
<td></td>
<td>- <strong>New Linkages to external data sources</strong></td>
</tr>
<tr>
<td></td>
<td>- Expanded researcher access</td>
</tr>
<tr>
<td></td>
<td>- Expanded ancillary study data</td>
</tr>
</tbody>
</table>

**Release of new versions of the Researcher Workbench**
Visit ResearchAllofUs.org to learn more about data sources and methodologies.

Use the Data Browser to view publicly available, de-identified data that can help you understand what’s available within the Researcher Workbench and what kind of scientific questions you can develop to inspire your research.
DataBrowser.ResearchAllofUs.org

- Provides summary statistics from the program’s growing database
- Open to everyone – no login!
- Allows participants to understand the makeup of the cohort
- Allows researchers to understand the characteristics of our participant population, explore the data types available and plan research questions

Interactive tool launched May, 2019
How to use the All of Us Data Browser

Search for specific keywords or browse using the different options underneath.
All of Us Data Browser: Dive in and spark your imagination

Available data gives insight into the participant cohort & research opportunities.

Hover mouse over charts & information icons for details & explanations.
All of Us Data Browser: A quick look

The Basics

This survey includes participant demographic information. Survey questions appear in the order in which participants took the survey.

Note: The data on this page are:
- Gathered directly from participants through electronic surveys
- Grouped into bins of 20 to protect privacy

For more information about this survey, please visit the Survey Explorer

Researchers can view the full surveys, including branching logic.
Cancer Counts (EHR N= >42,000)

- Malignant neoplastic disease: 42,080 (18.48%)
- Malignant neoplasm of skin: 12,560 (5.52%)
- Carcinoma in situ: 8,360 (3.67%)
- Malignant tumor of prostate: 6,410 (2.89%)
- Malignant neoplasm of female breast: 6,200 (2.78%)
- Basal cell carcinoma of skin: 3,000 (1.33%)
- Secondary malignant neoplastic ...
Comparing Female Breast Cancer Incidence Rate in AoU to US Incidence Rate
Comparing Male Prostate Cancer Incidence Rate in AoU to US Incidence Rate

Incidence Rate and SEER Incidence Rate for Male AoU participants

- **AoU Incidence Rate / Year**
- **SEER incidence rate**

<table>
<thead>
<tr>
<th>Year</th>
<th>AoU Incidence Rate</th>
<th>SEER Incidence Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017</td>
<td>0.0015</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td>0.0015</td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>0.0010</td>
<td></td>
</tr>
</tbody>
</table>
Immuno-Oncology use in AoU

Nivolumab (N=60)

Atezolizumab, avelumab, durvalumab, and cemiplimab all in database but in <20 participants

Nivolumab (N=60)

Pembrolizumab (N=160)
Use the Variant Search to explore allele frequencies for a gene or genomic region.

Explore data breakdowns by sex assigned at birth & age.

Drill down into specific variants to view select annotations and genetic ancestry associations.
Powerful analytical tools & user support

WORKSPACES
USES:
- Organizing research projects
- Collaboration
- Transparency

NOTEBOOKS
USES:
- Analyses
- Queries
Uses programming languages R and Python

DATASET BUILDER
USES:
- Pre-populated analyses
- Dataset previews

COHORT BUILDER
USES:
- Cohort creation

USER SUPPORT HUB
USES:
- Workbench training and resources
- Articles on Data and Tools
- Live office hours
Workspaces
Cohorts and Datasets

Cohorts
A cohort is a group of participants based on specific criteria.

Datasets
A dataset is a table containing data about a Cohort that can be exported for analysis.

Concept Sets
Concept sets are a set of medical concepts (e.g., blood pressure) that are of interest for a research study.
All of Us Data Dictionary

The Data Dictionary provides the following:

- A description for each data field;
- Information on whether the data in each field come from participant health records or from information the participants provide themselves, like survey data;
- Versioning data so you can see what has been changed, added, or removed since the previous curated dataset.

DATA SOURCES → DATA HARMONIZATION → DATA REFINEMENTS → CURATED DATA REPOSITORY + DATA DICTIONARY
Create an *All of Us* account

**Register to be an *All of Us* Researcher**

<table>
<thead>
<tr>
<th>STEP</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Learn more about the data available</td>
</tr>
<tr>
<td>2</td>
<td>Check for your institution's agreement</td>
</tr>
<tr>
<td>3</td>
<td>Register as a researcher</td>
</tr>
<tr>
<td>4</td>
<td>Verify your identity</td>
</tr>
<tr>
<td>5</td>
<td>Complete <em>All of Us</em> responsible conduct of research training</td>
</tr>
<tr>
<td>6</td>
<td>Sign data user code of conduct</td>
</tr>
</tbody>
</table>

[allofus.org/Register](http://allofus.org/Register)
No Login Required! Visit the Support Hub today to find answers to common questions about becoming a registered researcher, using the All of Us dataset, and more.

Find instructional materials about the All of Us Researcher Workbench
- Includes video tutorials, educational resources, release notes, and more

Search >250 useful articles
- Find information about data dictionaries, the Controlled Tier directory, how WGS and Array data are organized, and more

Connect with experts during weekly LIVE office hours
- Explore our calendar of dedicated office hours where researchers can talk about data types and tools with experts

researchallofus.org/support
Stay in touch to learn more

Sign up for our bimonthly newsletter

Subscribe to Research Roundup

Stay up-to-date on the latest news and insights from the *All of Us* Research Hub through our bimonthly email newsletter.

allof-us.org/RRSignup
Advantages and limitations

**Advantages**

- **DIVERSITY** of participants and data
- **GRANULARITY** of participant-provided data
- **SPEED** of data availability
- **ACCESSIBILITY** to broad group of researchers
- **POWERFUL** analytic capabilities
- **TEAM SCIENCE** approach
- **AVAILABILITY** of sample cohorts

**Limitations**

- **NOT NATIONALLY REPRESENTATIVE** (But reflective of the broad diversity of the U.S.)
- **DATA IS NOT YET COMPREHENSIVE** (It is continuing to expand)
- **NOT ALL PARTICIPANTS HAVE ALL DATA TYPES** AVAILABLE
- **LIMITED SET OF ENVIRONMENTAL/SOCIAL DETERMINANT DATA** (But more is coming through new surveys and linkages)
- **SOME KNOWLEDGE OF CODING REQUIRED**
All of Us Community and Provider Partner Network (as of August 2022)
All of Us Consortium Members (beyond community partners, as of August 2022)

HPO Network
- Health Care Provider Organizations

HPO Lite
- UT Health
- VCU Health

All of Us Wisconsin
- Marshfield Clinic Health System
- University of Wisconsin

All of Us California
- UC San Diego Health
- UCI Health

All of Us Pennsylvania
- University of Pittsburgh

All of Us New England
- Brigham and Women’s Hospital
- Mass General Brigham

All of Us Southeast Enrollment Center
- University of Florida Health

Trans America Consortium
- Northwestern Medicine
- UC San Diego Health

New York City Consortium
- Columbia University
- New York Presbyterian

Illinois Precision Medicine Consortium
- Washington University
- Northwestern Medicine

All of Us Southern Network
- University of Texas Health

University of Arizona and Banner Health
- Arizona Cancer Center of Arizona
- Banner Health

FQHCs (Federally Qualified Health Centers)
- Cherokee
- Sun River Health

Communications & Engagement
- WONDROS
- RTI International
- University of Utah

Participant Technology Systems Center (PTSC)
- Biobank
- Data & Research Center (DRC)

Genomics Partners
- Broad Institut
- University of Washington

Verily
- Color
- Mass General Brigham

Note: These are not approved logos and should not be repurposed on assets.