Research Information Series

Introduction to Data Science
Access and Resources

Janet Robishaw, Ph.D.
Joanne Krasnoff, Ph.D.
Oge Marques, Ph.D.

November 28, 2022
Transformational Research Enterprise

Patient Data
Transdisciplinary Team & Approaches
Technology Innovations

Identifying Causal Genes and Gene-X Interactions
Understanding Phenotypic and Genetic Heterogeneity
Developing New Diagnostic Tools and Drugs
Creating Polygenic Risk Algorithms
Access to Patient Data: FAU Medicine & Hospital Partners

Prep
- eClinicalWorks
- Power Bi

Stage 1
- Automated workflows
- Raw data dump into local tables

Stage 2
- Star schema data model
- Data entry by date, cleaning, deidentification
- Dashboards
- Relational
- Searchable

Stage 3

FAU Medicine
Marcus Center for Integrative Health
Mangurian Center for Brain Health
Broad Access to Patient Data:

- UK Biobank >500K Patients
- Strengths of EHR Warehouses:
  - No need for patient recruitment or data collection
  - Large sample size
  - Diverse population
  - Generalizability
  - Multi-site studies
FAU Medicine Data Warehouse

- Demographics
- Diagnoses
- Procedures
- Medications
- Labs & vitals
Welcome to the world’s largest, living ecosystem of real-world data and evidence for the life sciences and healthcare industries.

https://trinetx.com/
TriNetX

- Network is constantly refreshed and growing
- 120+ HCOs in 19 Countries
- Anonymized EMR-based-networks of 100m+ patients
- 70B+ date and patient-indexed clinical observations
- Free of charge to healthcare organizations

https://trinetx.com/
Local Research
Provide data and analytics to develop study protocols and pursue grants

Collaboration
Participate in multi-site investigator-initiated trials with peer research organizations

Sponsored Trials
Attract industry-sponsored clinical trials and funding for local research and development
FAU Available Networks

- **Florida Atlantic University**: FAU Medicine patients, on-line analytics, download data sets
- **APAC**: Asia-Pacific healthcare organizations, 6 HCOs, on-line analytics, no dataset download
- **COVID-19 Research**: Research network participants plus several EU sites, which is why there are more patients on this one than plain Research. On-line analytics only, no dataset download
- **EMEA**: Europe/Middle-East/Africa healthcare organizations, 21 HCOs, on-line analytics, no dataset download
- **Global Collaborative**: global healthcare organizations, 93 HCOs, on-line analytics, no dataset download
- **LATAM**: Latin America healthcare organizations, 14 HCOs, on-line analytics, no dataset download
- **Research**: US healthcare organizations plus other global HCOs that allow download data set capabilities, 72 HCOs, on-line analytics, data set download
- **US Collaborative**: Only US sites, 55 HCOs, on-line analytics, no data set download
Live Demo
Next steps:

- Working with TriNetX and OIT – FAU Medicine QA and QC + SSO login
- Platform information and training site
- Honest Data Broker process for providing datasets

Stay tuned for further information on roll-out...
Oge Marques, PhD
Professor
College of Engineering and Computer Science
Charles E. Schmidt College of Medicine
Florida Atlantic University
The future of health begins with you.

The All of Us Research Program is inviting one million people across the U.S. to help build one of the most diverse health databases in history. We welcome participants from all backgrounds. Researchers will use the data to learn how our biology, lifestyle, and environment affect health. This may one day help them find ways to treat and prevent disease.

JOIN NOW!

https://allofus.nih.gov
How the Program Works

1. Participants Share Data
2. Data is Protected
3. Researchers Study Data
4. Participants Get Information
5. Researchers Share Discoveries
The Participant side

https://www.joinallofus.org/
What is the All of Us Research Program?

ALL OF US IS A RESEARCH PROGRAM FROM THE NATIONAL INSTITUTES OF HEALTH (NIH). It is seeking one million or more people from across the U.S. By doing so, it hopes to one day help speed up medical research. People who join will share information about their health, habits, and what it’s like where they live. By looking for patterns, researchers may learn more about what affects people’s health.

https://www.joinallofus.org/
The All of Us Research Program is building a dataset to help transform the future of health research by equipping researchers with expansive health data from diverse populations, especially those underrepresented in biomedical research. More than 2,800 health researchers are leveraging this one-of-a-kind dataset to improve understanding of health and disease, identify opportunities to reduce disparities, and enable more precise approaches to care.

On this page:

- About the Research Hub
- Benefits for Researchers

https://allofus.nih.gov/get-involved/opportunities-researchers
Welcome to the *All of Us* Research Hub

The *All of Us* Research Program, led by the National Institutes of Health, is building one of the largest biomedical data resources of its kind. The *All of Us* Research Hub stores health data from a diverse group of participants from across the United States.

Registered researchers can access *All of Us* data and tools to conduct studies to help improve our understanding of human health.

[https://www.researchallofus.org](https://www.researchallofus.org)
What is All of Us (Research Hub)?

- Centralized, secure, cloud-based platform
- Allows researchers across a wide range of institutions and at all career levels to execute rapid, hypothesis-driven research
- Any US-based academic, non-profit, or healthcare organization can enter and utilize the All of Us Research Hub
- The All of Us data is curated into research datasets
- Tiered access (public, registered, controlled)
Why multiple tiers?

Public Tier
- Helps exposure
- Increase popularity
- Early assessment of usefulness of data
- However... many limitations (multiples of 20, etc.)

Registered Tier
- Enforces control
- Offers access to a wider range of data
- Greater granularity

Controlled Tier
- Contains genomic data in the form of whole genome sequencing (WGS) and genotype arrays
- Genotype arrays previously suppressed demographic data fields from EHRs, surveys and unshifted dates of events

Source: https://www.researchallofus.org/
Data Snapshots

Data Snapshots showcase the scale and diversity of the *All of Us* Research Program participant cohort. The snapshots provide participant demographics, geographic distribution, and more. We update the snapshots daily.

- **527,000+** Participants
- **322,000+** Electronic Health Records
- **390,000+** Biosamples Received

[View More Data Snapshots](https://www.researchallofus.org/data-tools/data-snapshots/)
The All of Us Research Program: Data quality, utility, and diversity

Andrea H. Ramirez • Lina Sulieman • David J. Schlueter • ... • Anthony Philippakis
Dan M. Roden • the All of Us Research Program

Open Access • DOI: https://doi.org/10.1016/j.patter.2022.100570 • Check for updates

https://doi.org/10.1016/j.patter.2022.100570
Data browser

• The Data Browser provides interactive views of the publicly-available All of Us Research Program participant data.
  – **Electronic Health Record (EHR) data** are derived from reports by health care providers.
  – **Genomic data** are derived from biosamples provided by participants.
  – **Physical measurements** are taken at the time of participant enrollment.
  – **Data from survey responses** and **wearables data** are collected from participants on an ongoing basis.

In order to protect participant privacy, personal identifiers are removed, aggregate data are rounded to counts of 20, and only summary demographic information is included.

Source: https://databrowser.researchallofus.org/
Using the Data Browser

Search Across Data Types

Search: sleep apnea

Data includes 372,380 participants and is current as of 6/6/2022.

EHR Domains

Conditions

12 matching medical concepts
227,740 participants in this domain

Procedures

4 matching medical concepts
221,860 participants in this domain

View Conditions

View Procedures
Using the Data Browser

Survey Questions

1 matching survey questions

142,100 participants in this domain

This survey includes information about past medical history, including medical conditions and approximate age of diagnosis.

View Complete Survey

1 matching survey questions

145,620 participants in this domain

Survey includes information about the medical history of a participant's immediate biological family members.

View Complete Survey
Using the Data Browser

Survey Questions

Personal Medical History

1 matching survey questions
142,100 participants in this domain
This survey includes information about past medical history, including medical conditions and approximate age of diagnosis.

View Complete Survey

Family Health History

1 matching survey questions
145,620 participants in this domain
Survey includes information about the medical history of a participant's immediate biological family members.

View Complete Survey
Using the Data Browser

142,100 1

Participants completed this survey matching of 465 questions available

Has a doctor or health care provider ever told you that you have or had any of the following respiratory conditions? (select all that apply)

See Answers >
## Using the Data Browser

Has a doctor or health care provider ever told you that you have or had any of the following respiratory conditions? (select all that apply)

<table>
<thead>
<tr>
<th>Answer</th>
<th>Concept Code</th>
<th>Participant Count</th>
<th>% Answered out of 142100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respiratory Conditions: No Lung Condition</td>
<td>1384615</td>
<td>85,920</td>
<td>60.46%</td>
</tr>
<tr>
<td>Respiratory Conditions: Asthma</td>
<td>43528439</td>
<td>27,240</td>
<td>19.17%</td>
</tr>
<tr>
<td>Respiratory Conditions: Sleep Apnea</td>
<td>43529837</td>
<td>22,960</td>
<td>16.16%</td>
</tr>
<tr>
<td>Respiratory Conditions: Other Lung Condition</td>
<td>1384564</td>
<td>6,960</td>
<td>4.90%</td>
</tr>
<tr>
<td>Skip</td>
<td>903096</td>
<td>6,280</td>
<td>4.41%</td>
</tr>
<tr>
<td>Respiratory Conditions: Chronic Lung</td>
<td>1384567</td>
<td>3,780</td>
<td>2.66%</td>
</tr>
<tr>
<td>Did not answer</td>
<td>0</td>
<td>≤ 20</td>
<td>0.01%</td>
</tr>
</tbody>
</table>
Using the Data Browser

<table>
<thead>
<tr>
<th>Answer</th>
<th>Concept Code</th>
<th>Participant Count</th>
<th>% Answered out of 22960</th>
</tr>
</thead>
<tbody>
<tr>
<td>How Old Were You Sleep Apnea: Adult</td>
<td>1385003</td>
<td>19,100</td>
<td>83.19%</td>
</tr>
<tr>
<td>How Old Were You Sleep Apnea: Older Adult</td>
<td>1385493</td>
<td>2,980</td>
<td>12.98%</td>
</tr>
<tr>
<td>How Old Were You Sleep Apnea: Elderly</td>
<td>1384967</td>
<td>360</td>
<td>1.57%</td>
</tr>
<tr>
<td>How Old Were You Sleep Apnea: Adolescent</td>
<td>1385492</td>
<td>280</td>
<td>1.22%</td>
</tr>
<tr>
<td>How Old Were You Sleep Apnea: Child</td>
<td>1385417</td>
<td>160</td>
<td>0.70%</td>
</tr>
<tr>
<td>Skip</td>
<td>903096</td>
<td>120</td>
<td>0.52%</td>
</tr>
</tbody>
</table>
### Using the Data Browser

**Are you still seeing a doctor or health care provider for Sleep Apnea?**

<table>
<thead>
<tr>
<th>Answer</th>
<th>Concept Code</th>
<th>Participant Count</th>
<th>% Answered out of 22960</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep Apnea Currently: Yes</td>
<td>43530234</td>
<td>14,920</td>
<td>64.98%</td>
</tr>
<tr>
<td>Sleep Apnea Currently: No</td>
<td>43529540</td>
<td>7,900</td>
<td>34.41%</td>
</tr>
<tr>
<td>Skip</td>
<td>903096</td>
<td>120</td>
<td>0.52%</td>
</tr>
</tbody>
</table>
## Using the Data Browser

### Are you currently prescribed medications and/or receiving treatment for Sleep Apnea?

<table>
<thead>
<tr>
<th>Answer</th>
<th>Concept Code</th>
<th>Participant Count</th>
<th>% Answered out of 22960</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rx Meds for Sleep Apnea: Yes</td>
<td>43530212</td>
<td>14,640</td>
<td>63.76%</td>
</tr>
<tr>
<td>Rx Meds for Sleep Apnea: No</td>
<td>43529517</td>
<td>8,160</td>
<td>35.54%</td>
</tr>
<tr>
<td>Skip</td>
<td>903096</td>
<td>140</td>
<td>0.61%</td>
</tr>
</tbody>
</table>
Using the Data Browser

Survey Questions

Personal Medical History

1
matching survey questions
142,100 participants in this domain
This survey includes information about past medical history, including medical conditions and approximate age of diagnosis.

View Complete Survey

Family Health History

1
matching survey questions
145,620 participants in this domain
Survey includes information about the medical history of a participant's immediate biological family members.

View Complete Survey
Using the Data Browser

145,620 1

Participants completed this survey matching of 104 questions available

Has anyone in your family ever been diagnosed with the following lung conditions? Think only of the people you are related to by blood. Select all that apply.

See Answers >
Using the Data Browser

Has anyone in your family ever been diagnosed with the following lung conditions? Think only of the people you are related to by blood. Select all that apply.

<table>
<thead>
<tr>
<th>Answer</th>
<th>Concept Code</th>
<th>Participant Count</th>
<th>% Answered out of 145620</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respiratory Condition: Asthma</td>
<td>43528438</td>
<td>27,840</td>
<td>19.12%</td>
</tr>
<tr>
<td>Respiratory Condition: Chronic Lung Disease</td>
<td>43528553</td>
<td>27,620</td>
<td>18.97%</td>
</tr>
<tr>
<td>Respiratory Condition: Sleep Apnea</td>
<td>43529836</td>
<td>21,740</td>
<td>14.93%</td>
</tr>
<tr>
<td>None</td>
<td>903095</td>
<td>12,660</td>
<td>8.69%</td>
</tr>
<tr>
<td>Don't Know</td>
<td>903087</td>
<td>3,980</td>
<td>2.73%</td>
</tr>
<tr>
<td>Skip</td>
<td>903096</td>
<td>1,400</td>
<td>0.96%</td>
</tr>
<tr>
<td>Prefer Not To Answer</td>
<td>903079</td>
<td>60</td>
<td>0.04%</td>
</tr>
<tr>
<td>Did not answer</td>
<td>0</td>
<td>≤ 20</td>
<td>0.02%</td>
</tr>
</tbody>
</table>
Using the Data Browser

<table>
<thead>
<tr>
<th>Answer</th>
<th>Concept Code</th>
<th>Participant Count</th>
<th>% Answered out of 21740</th>
</tr>
</thead>
<tbody>
<tr>
<td>Father Respiratory Condition: Sleep Apnea</td>
<td>43529833</td>
<td>8,780</td>
<td>40.39%</td>
</tr>
<tr>
<td>Sibling Respiratory Condition: Sleep Apnea</td>
<td>43529838</td>
<td>7,500</td>
<td>34.50%</td>
</tr>
<tr>
<td>Mother Respiratory Condition: Sleep Apnea</td>
<td>43529835</td>
<td>5,700</td>
<td>26.22%</td>
</tr>
<tr>
<td>Grandparent Respiratory Condition: Sleep Apnea</td>
<td>43529834</td>
<td>3,120</td>
<td>14.35%</td>
</tr>
<tr>
<td>Son Respiratory Condition: Sleep Apnea</td>
<td>43529839</td>
<td>1,580</td>
<td>7.27%</td>
</tr>
<tr>
<td>Daughter Respiratory Condition: Sleep Apnea</td>
<td>43529832</td>
<td>880</td>
<td>4.05%</td>
</tr>
<tr>
<td>Skip</td>
<td>903096</td>
<td>240</td>
<td>1.10%</td>
</tr>
</tbody>
</table>
Researcher Workbench

- The workbench is a cloud-based platform where registered researchers can access Registered and Controlled Tier data.

- It has powerful tools which support data analysis and collaboration.
Data Sources

- Electronic Health Records
- Biosamples and Bioassays
- Surveys
- Physical Measurements
- Mobile Health Data (Wearable Devices)
Data Methods

DATA CURATION

DATA SOURCES → DATA HARMONIZATION → DATA REFINEMENTS → CURATED DATA REPOSITORY + DATA DICTIONARY
The workbench includes:

- **Workspaces** which are a collaborative environment, where registered researchers can access, store and analyze data for specific research purposes. This allows them to organize research projects and join forces in order to tackle more challenging tasks.
- **Notebooks** are where researchers with R or Python experience can perform high-powered analysis within the All of Us datasets using the integrated, cloud-based Jupyter Notebook environment.
- **Dataset builder** which allows researchers to search and save collections of health information about cohorts called concept sets. It is extremely beneficial when performing pre-populated analysis and dataset reviews.
- **Cohort Builder** is a custom point-and-click tool allowing researchers to create, review and annotate groups of participant data or cohorts within the dataset. The primary focus of the cohort builder is ultimately cohort creation.
- **Workbench User Support Hub** contains a variety of resources, for example, training materials, that demonstrate how to use the tools and navigate the data within the Researcher Workbench. This is used for learning, support and guides.
Research Projects Directory

3,058 ACTIVE PROJECTS
This information was updated 11/23/2022

The Research Projects Directory includes information about all projects that currently exist in the Researcher Workbench to help provide transparency about how the Workbench is being used. Each project specifies whether Registered Tier or Controlled Tier data are used.

Note: Researcher Workbench users provide information about their research projects independently. Views expressed in the Research Projects Directory belong to the relevant users and do not necessarily represent those of the All of Us Research Program. Information in the Research Projects Directory is also cross-posted on AllOfUs.nih.gov in compliance with the 21st Century Cures Act.

Search By:  
Project title: fau

2 projects have 'fau' in the project title
Transdisciplinary research

All of Us
RESEARCHER WORKBENCH

Data
Analysis
About

Cohorts

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

All of Us Participants

→

Your Cohort

Participant ID 1
Participant ID 2
Participant ID 3

Datasets

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

Your Cohort

Data About Your Cohort

Participant ID 1
Participant ID 2
Participant ID 3

Medication
Labs

ID 1
Med 1
Labs 1

ID 2
Med 2
Labs 2

ID 3
Med 3
Labs 3

NSF-Funded Traineeship in Data Science Technologies and Applications
Cohort 2 Starts Fall 2022
Questions?

omarques@fau.edu
Reminder: Please complete Qualtrics Evaluation
Thank you