Deploying Cloud-based Platforms and Analytic Tools to Support Covid-19 and Beyond

August 6, 2020 | 12:00 – 1:30 pm ET
About the Series:

- Made possible with funding from the Gordon and Betty Moore Foundation
- To foster collaboration between specialty societies and academia, we are grateful to collaborate with the Association of Academic Medical Colleges

Continue the Conversation:

- Use #COVIDRegistries when tweeting about the webinar series
- Follow @CMSSMed and visit CMSS.org for frequent updates
CMSS WEBINAR SERIES
Advancing Clinical Registries to Support Pandemic Treatment and Response

Today’s Webinar:

Deploying Cloud-based Platforms and Analytic Tools to Support Covid-19 and Beyond

Moderator:

William J. Marks, Jr
MD, MS-HCM
Head of Clinical Science & Head of Neurology, Verily Life Sciences;
Adjunct Clinical Professor of Neurology & Neurological Sciences, Stanford University School of Medicine

Host:

Helen Burstin, MD, MPH, MACP
Chief Executive Officer
Council of Medical Specialty Societies (CMSS)

Panelists:

David Glazer
Engineering Director
Verily Life Sciences

Andrea Ramirez, MD
Assistant Professor of Medicine,
Vanderbilt University of Medicine

Chris Treml
Director of Operations, Data Science Institute, American College of Radiology
Biomedical Analysis in the Cloud

multi-modal, multi-source, multi-tool, multi-use

David Glazer, Verily Life Sciences
CMSS Webinar, 6-aug-2020
**Traditional approach**

*Bring data to researchers*

- Discourages shared research
  - “Weakest link” security
  - Huge infrastructure needed
  - Pay for multiple copies
  - Bespoke & unsupported tools

**Cloud-centric approach**

*Bring researchers to data*

- Facilitates collaboration
  - Centralized security controls
  - Accessible to all researchers
  - Decreased cost of storage
  - Shared tool ecosystem
Uniting the biomedical ecosystem

Vision
Enable the next generation of collaborative biomedical research

Principles
Modular
composed of functional components with well-specified interface

Community driven
created by many groups to foster a diversity of ideas

Open
open-source licenses, software, arch to enable extensibility & reuse

Standards based
consistent with standards developed by coalitions such as GA4GH
Terra is a scalable & secure platform for biomedical researchers to access data, run analysis tools, and collaborate.
Terra’s Value Proposition

**Integrated data**

Platform enables the integration and secure access of disparate data types and datasets, reducing data latency, costs, quality issues, and inefficiencies associated with fragmentation.

**Integrated tools**

Platform integrates proprietary and third-party tools to run leading batch and interactive analysis, reducing the challenges associated with multiple systems and frameworks.

**Collaborative workspaces**

Platform enables rich collaboration with sophisticated provenance, security controls, and compliance for faster and more effective analysis within and across organizations.
Terra’s architecture primarily serves 3 roles:

- **Data Production & Curation**: Data Generators
- **Tool Creation & Publication**: Tool Developers
- **Biomedical Researchers**: Biomedical Researchers

These roles interact through the Terra Platform API, which consists of:

- **Data Management (Data Mgmt)**
- **Workspaces**
- **Tools Management (Tools Mgmt)**

In addition, there are Cloud Services that support these interactions.
Partners & Datasets
Tools currently available in Terra

- Cromwell
- jupyter
- hail
- R Studio
- TensorFlow
- Keras
- Bokeh
- Spark
- seaborn
- biopython
- scikit-learn
- NumPy
- Dockstore
- Create, Share, Use
Thousands of researchers use Terra every month to analyze biomedical data
Terra use case: 

All of Us Research Program
All of Us Research Program

- Engage **1,000,000 or more** U.S. research participants
- Share tissue samples, genetic data, lifestyle information, electronic health records
- Pioneer a new model of research that emphasizes engaged research participants, responsible data sharing, and privacy protection

Research focuses on the intersection of 3 factors:

- Environment
- Lifestyle
- Biology

All of Us and the All of Us logo are service marks of the U.S. Department of Health and Human Services (HHS).
All of Us Program Components

Health Care Provider Organizations

Data and Research Center

VANDERBILT UNIVERSITY MEDICAL CENTER

Biobank

Genome Centers

Participant Technology Systems Center
Data Curation Process

1. **DATA SOURCES**
2. **DATA HARMONIZATION**
3. **DATA REFINEMENTS**
4. **CURATED DATA REPOSITORY**
5. **DATA DICTIONARY**
Researchers Invited to Give Feedback on Initial Dataset and Tools

In partnership with our participants—now nearly 350,000 and counting—we’re working to build one of the world’s largest and most diverse datasets to advance health research. Today, I’m happy to announce that we’ve opened our research platform, the All of Us Researcher Workbench, for beta testing. Now, researchers can begin using our initial dataset and tools in studies and tell us what’s working and what we can improve. This moment is an important step in our effort to accelerate new discoveries.
Research Hub → Researcher Workbench

PUBLIC

- Data Browser >
- Survey Explorer >
- Help Desk >

REGISTERED RESEARCHERS >

RESTRICTED

RESEARCHER WORKBENCH

- Data Dictionary >
- Cohort Builder >
- Notebooks >
- ...and more! >
The Workbench contains the tools researchers need to learn about, access and analyze *All of Us* data.
Within the Researcher Workbench, researchers run analyses, individually or collaboratively, in “Workspaces.”
Researchers have access to integrated support

Workspace

Cohort Builder

Dataset Builder

Jupyter Notebook

All of Us USER SUPPORT HUB

Help Desk

Please select your issue

Ask a question or report an issue.

Share your feedback.
Once you create your workspace, you can create a cohort for your study using Cohort Builder, where you will select participants to include or exclude from your study based on criteria of interest for your research.
After creating your cohort, you can check your work using a Cohort Review Set.
Cohorts + Concept Sets + Values = Datasets

➢ Once you have your cohort, you can start to put together your dataset.
➢ Your dataset is made up of a cohort, a concept set, and selected values.

Datasets

Build a dataset by selecting the variables and values for one or more of your cohorts. Then export the completed dataset to Notebooks where you can perform your analysis.

1. Select Cohorts (Participants)
   - Prepackaged Cohorts
     - All Participants
   - Workspace Cohorts
     - Major Depression Cohort

2. Select Concept Sets (Rows)
   - Prepackaged Concept Sets
     - Demographics
     - Mental Health Concept Set
   - Workspace Concept Sets
     - All Surveys
     - Survey

3. Select Values (Columns)
   - Survey
     - person_id
     - survey_datetime

Learn more in the data dictionary

SAVE AND ANALYZE
Once you select your cohort, your concept set(s), and values, you now have your dataset.

Now you can preview your dataset before saving it and exporting it to a notebook for analysis.

* synthetic data
Jupyter Notebooks can be used for analysis in R and Python.

- Notebooks come with ready-made code that can be inserted called “Snippets”
- Each user gets a virtual machine (computing system) per workspace
**Tools@Beta:** Powerful, Flexible Tools for Reproducible Science

Workspace ➔ Cohort Builder ➔ Dataset Builder ➔ Jupyter Notebook
Presented on Behalf of the Data and Research Center (DRC)
Terra use case:
UK Biobank
AMP PD - Beta 1 - Getting Started

The purpose of this workspace is to provide getting started information and notebooks for researchers

Last changed: 9:03 PM

AMP PD Demographics

Sample workspace for notebooks on top of AMP PD demographics data.
The workspace is access controlled using the amp-pd-researchers

Last changed: Aug 23

Baseline Health Study

This workspace contains sample notebooks for analyzing Baseline Health Study data.

You may view notebooks in this

Last changed: Feb 7

Kathiresan Lab UK Biobank

Overview
We reproduced on Terra a subset of the analyses in:

> Haas, ME et al. (2018) **Genetic

Last changed: Feb 7

Kathiresan Lab UK Biobank defaux copy

Overview
We reproduced on Terra a subset of the analyses in:

> Hass, ME et al. (2018) **Genetic

Last changed: Jan 22

dG copy -- Baseline Health Study

This workspace contains sample notebooks for analyzing Baseline Health Study data.

Last changed: Feb 7

test

No description added

try-saturn

A workspace for trying out Terra functionality as it evolves.
Overview

We reproduced on Terra a subset of the analyses in:

Haas, ME et al. (2018) Genetic Association of Albuminuria with Cardiometabolic Disease and Blood Pressure. AJHG volume 103, issue 4, p461-473. doi:10.1016/j.ajhg.2018.08.004

At a high level:

- researchers[] have taken most columns from the raw UK Biobank phenotypes and run GWAS
- in this paper, mhaas et. al. have combined several phenotypes into one that is clinically robust
- they then performed GWAS against that derived phenotype, generating very good results

For additional overview, please see mhaas slide deck from a lab meeting while the research was in progress.

Analysis Artifacts

Notebooks are collaboratively written in this Terra workspace.

R environment setup

Install R packages and notebook extensions specific to this set of analyses. Each time the user starts a new Leonardo instance (such as when resizing) the user must run this notebook to set up.

- notebook
- html version of notebook

Curation - Load UK Biobank provided data to Cloud Platform

Wrangle the data provided by UK Biobank in to Cloud Storage and/or BigQuery as appropriate for the expected usage patterns.
Perform the linear regression

```python
In [22]:
    covar_cols = ['sa.pheno.age', 'sa.covar.array', 'sa.covar.PC1',
                   'sa.covar.PC2', 'sa.covar.PC3', 'sa.covar.PC4', 'sa.covar.PC5', 'sa.covar.PC6',
                   'sa.covar.PC7', 'sa.covar.PC8', 'sa.covar.PC9', 'sa.covar.PC10']

In [23]:
    # For the original analysis, the destination was
    # gs://ukbb_v2/projects/mhass/ACR/UKB.v2.n382500.InACRadjS.021618.tsv
    result_destination = '/'.join(['gs://uk-biobank-sek-data-us-east1/gwas-results/InACR',
                                   'timestamp',
                                   '.', 'join(['UKB.v2.n382446.InACRadjS',
                                              'datestamp',
                                              'tsv'])])

    result_destination

Out[23]:
    'gs://uk-biobank-sek-data-us-east1/gwas-results/InACR/20181121_194957/UKB.v2.n382446.InACRadjS.20181121.tsv'

In [24]:
    vds_linassoc = (vds
                   .linreg3(['sa.pheno.InACRadjS'],
                             covariates=covar_cols,
                             use_dosages=True,
                             variant_block_size=10)
                   .drop_samples())

    print('Starting linear regression:')
    vds_linassoc.export_variants(result_destination,
                                  'Variant = v, bgen.snp_id = va.varid, rsid = va.rsid, INFOScore_QCTOOL = va.INFOScore_QCTOOL,
                                  vCallRate = va.qc.callRate, pHWE = va.qc.pHWE, Cohort_AC = va.qc.AC, Cohort_AF = va.qc.AF',
                                  'browser')
```

```
Terra use case:
COVID-19 Research

Data & Tools for COVID-19/SARS CoV2 analysis
See this article for a summary of available resources.
Terra showcase

New and interesting

COVID-19_Broad_Viral_NGS
Massachusetts has been severely impacted by the COVID-19 pandemic, with 101,163 cases and 7,085 deaths as of June 2, 2020. Seventy percent of the state’s 6.9 M population lives in the city.

COVID-19
This workspace contains COVID-19 genomic data and workflows that will enable you to perform viral genomic analysis. This workspace will be routinely updated with new, additional data as it becomes available.

COVID-19_cross_tissue_analysis
The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, underscores the urgent need to identify molecular mechanisms that mediate viral entry, propagation, and tissue

Featured workspaces

Introduction-to-TCGA-Dataset
# Workspace Overview
Practical accessing and analysing controlled-access TCGA data with example analysis tools. Data processing

DNA-methylation-preprocessing
### DNA-methylation-preprocessing
Suite of tools to conduct methylation data analysis. Methods from this workspace can be used for alignment.

Bioconductor
Explore common Bioconductor packages that can be used to perform bulk RNA differential expression analyses or manipulate single-cell RNA-seq data.

GATK4 example workspaces

Germline-CNVs-GATK4
### GATK Best Practices for Germline Copy Number Variation
An analysis to detect germline copy number variants in exome sequence.

Variant-Functional-Annotation-With-Functocator
### GATK Best Practices for Functocator
**Functocator** (FUNCtional annOTATOR) analyzes variants for their function and writes the analysis to a specified output file.

Variant_Calling_Spark_Multicore
### GATK Best Practices for Variant Calling with Spark on a Multicore Machine
This workspace highlights a pipeline for
COVID-19 Viral Genomics

From .BAM to NextStrain Tree and GenBank Data Submission

- Your .BAM files
  - wdl
  - Perform Assembly
- Perform GenBank Data Submission
- Your .fasta files + Broad/ MGH .fasta + GenBank .fasta

2019-2020 COVID-19 Samples

Diagram showing the distribution and diversity of COVID-19 samples from 2019 to 2020.
COVID-19 Cross Tissue Analysis

Integrated analyses of single-cell atlases reveal age, gender, and smoking status associations with cell type-specific expression of mediators of SARS-CoV-2 viral entry and highlights inflammatory programs in putative target cells


METHODS: Code availability
Data and an interactive analysis examining the co-expression of genes across datasets can be accessed via the open-source data platform, Terra at https://app.terra.bio/#workspaces/kco-incubator/COVID-19_cross_tissue_analysis.

ABOUT THE WORKSPACE
The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, underscores the urgent need to identify molecular mechanisms that mediate viral entry, propagation, and tissue pathology in distinct cell types across organs. The surface receptor angiotensin-converting enzyme 2 (ACE2) and the associated proteases, transmembrane protease serine 2 (TMPRSS2) and Cathepsin L (CTSL), were previously identified mediators of SARS-CoV cellular entry.

Experimental Overview
We use single-cell RNA-seq (scRNA-seq) across diverse tissues to assess the cell-type-specific expression of ACE2, TMPRSS2, and CTSL. We identify specific subsets of respiratory epithelial cells as putative targets of viral infection, including subsets of epithelial cells in the nasal passages, lung and airways. Additionally, we detect expression in other tissues that may serve as routes of viral transmission, including the gut and corneal epithelia, and in cells potentially associated with COVID-19 clinical pathology including cardiomyocytes, olfactory sustentacular cells, and renal epithelial cells.

For more details about the resources presented in this Terra Workspace, please use this link to refer to the official manuscript hosted on BioRxiv.
Platforms ⇒ Rapid Response
How do we prepare for the unexpected?

- Collaborative platforms allow fast sharing of new tools, knowledge, and data.
  - Example: COVID-19 viral genomics and tissue expression workspaces
  - Example: All of Us COPE surveys

- General-purpose platforms allow fast special-purpose analysis.
  - Example: UK Biobank surveyed their participants about COVID-19, and provided some of the first strong evidence about genetic factors in disease susceptibility and severity.
  - Example: All of Us is running serology assays on biosamples collected in early 2020, to look for early COVID-19 exposure patterns.
Thank you!
Novel Partnerships and Research Tools to Support Covid-19

Chris Treml
American College Radiology
Who We Are

• Member based organization
  • ~40,000 members

• Center for Research and Innovation
  • Over 140 staff
  • 2 million+ images process annually
Operating Partnership Types
SINGLE PARTY

PRIMARY /SECONDARY

FEDERATED
Single Party

SIMPLE

WIDELY USED

CONTROL
Primary/Secondary

“SUB-CONTRACTOR”

COMPLEMENTARY EXPERTISE

MID COMPLEXITY
• Virtual registry
• >40 data elements collected
• Groups
  • NIBIB
  • Gen3
  • ACR
  • RSNA
  • AAPM
  • TCIA
• Independence on collection
• Top level abstraction layer
• Governance
  • Unifying data elements
  • Unifying subject IDs
  • Site enrollment
  • Timelines
• NIBIB
NIBIB COVID-19 Public Portal

Access Gateway

- Unrestricted Public Access
- Integrated Search
- API

Common Data Model (Local Mappings)

Common QC Protocols (Augment local workflow)

Protocols for acquisition and De-Identification

Data Types: DICOM, EMR, Lab, ...

- Integrated Registry (images and clinical data)
- Real-time Upload & Federated Queries
- Real-world Testing of AI

- Image Repository and limited clinical data
- Retrospective Batch Upload
- AI Competitions

- Image Repository for NCI-funded trial activity
• >40 data elements collected
• Demographics
• Vitals
• Labs
• Images
• Diagnosis
• Repeatable
• Single Instance
CONNECT
• Anonymization
• QC
• Mapping
• Automatic abstraction
• Local control
Wrapping up

- Different partnership models, different benefits
  - Single, Primary/Secondary, Federated
- COVID-19 presenting areas for high scalability
  - MIDRC example
- ACR Connect
  - Last mile platform for data gathering
All of Us Researcher Workbench
CMSS Webinar: Deploying Cloud-based Platforms and Analytic Tools to Support Covid-19 and Beyond
August 6, 2020

Andrea H Ramirez, MD, MS
Vanderbilt University Medical Center
andrea.h.ramirez@vumc.org

@AndreaRamirezMD
The All of Us Research Program is a historic, longitudinal effort to gather data from one million or more people living in the United States to accelerate research and improve health. By taking into account individual differences in lifestyle, socioeconomics, environment, and biology, researchers will uncover paths toward delivering precision medicine – or individualized prevention, treatment, and care – for all of us.

“All of Us is among the most ambitious research efforts that our nation has undertaken!”

NIH Director Francis Collins, M.D., Ph.D.

The All of Us Research Program is part of the broader Precision Medicine Initiative.
On behalf of the Data and Research Center team!
The All of Us Research Program

Nurture relationships with one million or more participant partners, from all walks of life, for decades.

Our mission
To accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us.

Catalyze a robust ecosystem of researchers and funders hungry to use and support it.

Deliver the largest, richest biomedical dataset that is easy, safe, and free to access.
The Research Hub: Access & Analyze All of Us Data

Registered Tier, Participant-level Data

Public, Aggregate Data
Research Hub: Public Tools for Browsing Aggregate Data

Data Browser

Data Snapshots

Survey Explorer
Research Projects Directory

This list, updated 3/31/2020 shows how researchers are using All of Us data.

*Note: Researcher Workbench users fill out their research project information independently. Any views, opinions, or other comments made in the Directory belong to the relevant research project team. These views, opinions, or comments do not necessarily represent the All of Us Research Program’s beliefs.

**Information in the Research Projects Directory is also cross-posted on AllofUs.gov in compliance with the 21st Century Cures Act.

There are currently 1,657 active workspaces. This information was updated on 4/6/2020.

SORT BY TITLE: ABC | DEF | GHI | JKL | MNO | PQR | STU | VWX | YZ | 0-9

PROJECT PURPOSE(S)
- Population Health

SCIENTIFIC QUESTIONS BEING STUDIED
Not available.

SCIENTIFIC APPROACHES
Not available.

ANTICIPATED FINDINGS
Not available.

DEMOGRAPHIC CATEGORIES OF INTEREST
Not available.

RESEARCH TEAM
Owner: Oscar Carlton - PI

https://www.researchallofus.org/research-projects-directory/?section=stu
Research Hub → (Beta) Researcher Workbench

- PUBLIC
  - Data Browser
  - Data Snapshots
  - Survey Explorer

- RESTRICTED
  - Data Access Approval
  - Data Dictionary
  - Concept Set Selector
  - Cohort Builder
  - Notebooks
  - Help Desk
What to Expect as a Beta Researcher

1. Currently, an institutional agreement & eRA Commons account must be in place.

2. Feedback is welcome.

3. The tools are continuing to evolve.

4. The program cohort is actively growing, and so is our data.

5. The goal is to be a true game changer for understanding health.
What data are available now?
Enroll, Consent and Authorize EHR
- Recruiting 18+ years old initially; plan to include children in future
- Online, interactive consent
- Includes authorization to share Electronic Health Record (EHR) data

Answering Surveys
- Initial surveys: The Basics, Overall Health, Lifestyle, Health Care Access & Utilization, Family Medical History, Personal Health History
- Additional surveys will be released on an ongoing basis.

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

Provide Biosamples*
- Blood (or saliva, if blood draw is unsuccessful)
- Urine specimen
- Biosamples will be stored at the program’s biobank

Wearables and Digital Apps
- Share data from wearable fitness devices, starting with Fitbit
- More integrations to come. E.g., integrated apps to track mood & cardio-respiratory fitness

Available in the Current Dataset

*Based on diverse sampling and capacity

Our Data Is Growing. Here are the Current Data Types
## Research Data Available Now

### Data Type

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Participant Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>&gt;224,000</td>
</tr>
<tr>
<td>Physical Measurement</td>
<td>&gt;188,000</td>
</tr>
<tr>
<td>Electronic Health Record</td>
<td>&gt;127,000</td>
</tr>
</tbody>
</table>

### Count of participants with multiple data types

- Survey: 34823 participants
- EHR: 771 participants
- Physical Measurement: 126362 participants
- Electronic Health Record: 62187 participants
A. Ramirez, L. Suleiman, D. Schlueter, et al., *The All of Us Research Program: data quality, utility, and diversity*, medRxiv 2020.05.29.20116905; doi:
### Data@Beta: Data Quality

<table>
<thead>
<tr>
<th>Description</th>
<th>EHR Ever Smoking OR (95% CI)</th>
<th>Survey Ever Smoking OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top 3 Increased risk effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer of the bronchus; lung</td>
<td>4.94 (4.11, 5.95)</td>
<td>3.19 (2.65, 3.84)</td>
</tr>
<tr>
<td>Cancer within the respiratory system</td>
<td>4.94 (4.12, 5.92)</td>
<td>3.15 (2.62, 3.78)</td>
</tr>
<tr>
<td>Malignant neoplasm of bladder</td>
<td>2.36 (1.87, 2.98)</td>
<td>1.76 (1.42, 2.18)</td>
</tr>
<tr>
<td><strong>Top 3 Decreased risk effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vascular hamartomas and non-neoplastic nevi</td>
<td>0.51 (0.42, 0.62)</td>
<td>0.55 (0.48, 0.64)</td>
</tr>
<tr>
<td>Nevus, non-neoplastic</td>
<td>0.52 (0.43, 0.64)</td>
<td>0.57 (0.49, 0.66)</td>
</tr>
<tr>
<td>Benign neoplasm of skin</td>
<td>0.53 (0.49, 0.58)</td>
<td>0.62 (0.58, 0.66)</td>
</tr>
</tbody>
</table>

A. Ramirez, L. Suleiman, D. Schlueter, et al., *The All of Us Research Program: data quality, utility, and diversity*, medRxiv 2020.05.29.20116905; doi:
A. Ramirez, L. Suleiman, D. Schlueter, et al., *The All of Us Research Program: data quality, utility, and diversity*, medRxiv 2020.05.29.20116905; doi:
How do you get access?
Interested in becoming a Beta Researcher?

Visit researchallofus.org to apply!
How the Application Process Works

**STEP 1**
CHECK FOR YOUR INSTITUTION’S AGREEMENT

Check that your institution has signed the Data Use and Registration Agreement. If you do not see your institution listed, please [fill out the form](https://www.researchallofus.org/institutional-agreements/) to initiate the process.

**STEP 2**
REGISTER AS A RESEARCHER

Complete your researcher profile, sign Terms of Services, and agree to the Privacy Policy. Please note, a Data Use and Registration Agreement must be in place, and you must have an eRA Commons account.

**STEP 3**
CONNECT eRA COMMONS

Connect your eRA Commons account upon applying to the Researcher Workbench.

**STEP 4**
COMPLETE All of Us RESPONSIBLE CONDUCT OF RESEARCH TRAINING

Understand more about our privacy safeguards and the ethics surrounding the use of participant data.

**STEP 5**
SIGN AUTHORIZED USER CODE OF CONDUCT

Each user must agree and sign an Authorized User Code of Conduct.

[https://www.researchallofus.org/institutional-agreements/](https://www.researchallofus.org/institutional-agreements/)
Currently, 165 institutions are in the access pipeline with 110 having completed agreements!

Institutions are Diverse:
- Non-profit Orgs
- Federally Qualified Health Centers
- Historically Black Colleges & Universities
- R2 Public Universities
- Major Academic Medical Centers
- Disease Advocacy Groups
- And growing....
Ok now what?
Welcome to the Researcher Workbench

The Researcher Workbench contains the tools researchers need to learn about, access and analyze participant-level All of Us data.
Searchable Knowledge Base, Community Forums, and Help Desk
These workspaces demonstrate how computable electronic phenotypes can be implemented within the All of Us dataset using examples of previously published phenotype algorithms. You can open the workspaces to view them or “duplicate” the workspaces to edit and execute the algorithms.
We implemented a validated phenotyping algorithm for breast cancer from eMERGE network.

1. We found 3748 total breast cancer cases
   a. 3714 female cases (99.1%)
   b. 34 male cases (0.9%)
2. Algorithm utilized diagnosis/history codes in the OMOP observation and condition occurrence tables
3. This algorithm will be made available to researchers as part of the Phenotype library
   a. Cohort builder implementation and notebook query implementation match

**Issues encountered:**
1. Be careful about how you specify certain codes in cohort builder vs. notebooks
Tutorial Workspace Example: Working with Survey Data

**Issues Encountered**

- Data is formatted long
- “Child questions” are not always answered by the intended participants
- Several tables contain survey information [DS_SURVEY, CONCEPT, OBSERVATION, OBSERVATION_EXT] and field names may vary
- May be joined together, but that takes computational time

**Ex. How to query survey data**

1. View a pdf of the **Overall Health** Survey using [Survey Explorer](#).

2. Identify question of interest:

   **How confident are you filling out medical forms by yourself?**
   - Extremely / Quite a bit / Somewhat / A little bit / Not at all / Skip

3. Go to [OHDSI ATHENA](#) to find the associated ID, Code, and Name

4. In a Notebook, query the **All of Us** custom DS_SURVEY table

   ```sql
   WHERE question_concept_id = 1585766
   ```

5. Output frequency table of answers

6. Plot counts of responses
All of Us Demonstration Project Subcommittee

The DRC AoU Demonstration Projects will use Research Hub tools to characterize and validate the AoU cohort and data in order to demonstrate the quality, utility, and diversity of AoU data and tools. AoU Demonstration Projects will replicate known, previously published findings, and will not aim to discover new disease classifications, relationships, or risk associations -- our role is to support the community’s use of data, not to make first discoveries.

Members: Andrea Ramirez (Chair), Kelly Gebo, Cheryl Clark, Elizabeth Cohn, Lucila Ohno-Machado, Mine Cicek, Eric Boerwinkle, Sheri Schully, Steve Mockrin
Goal: Fully executed research projects demonstrating the utility and validity of AoU data timed to publish at data platform launch, not novel discovery work.

**Phase 1**  
[DRC] Description, Replication, Utility Assessment

**Phase 2**  
[Consortium] Expanded Description, Replication, Utility Assessment

**Phase 3**  
[Consortium] Future Preceding new data types
Engaging a large consortium, covering broad health areas

<table>
<thead>
<tr>
<th>All of Us Awardee</th>
<th>No. of Approved Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans-American Consortium for the Health Care Systems Research Network (TACH)</td>
<td>4</td>
</tr>
<tr>
<td>NYC Precision Medicine Consortium</td>
<td>4</td>
</tr>
<tr>
<td>The Participant Center</td>
<td>5</td>
</tr>
<tr>
<td>Asian Health Coalition</td>
<td>2</td>
</tr>
<tr>
<td>University of Arizona</td>
<td>4</td>
</tr>
<tr>
<td>New England Precision Medicine Consortium</td>
<td>4</td>
</tr>
<tr>
<td>California Precision Medicine</td>
<td>2</td>
</tr>
<tr>
<td>Stanford University</td>
<td>1</td>
</tr>
<tr>
<td>Illinois Precision Medicine Consortium</td>
<td>5</td>
</tr>
<tr>
<td>South East Enrollment Center (SEEC)</td>
<td>4</td>
</tr>
<tr>
<td>Data and Research Center (DRC)</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Health Areas</th>
<th>No. Projects in Health Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular Disease</td>
<td>7</td>
</tr>
<tr>
<td>Cancer</td>
<td>5 (2 overlaps with other area)</td>
</tr>
<tr>
<td>Diabetes and Obesity</td>
<td>6 (3 overlap with other areas)</td>
</tr>
<tr>
<td>Mental Health</td>
<td>2</td>
</tr>
<tr>
<td>Wellness</td>
<td>2</td>
</tr>
<tr>
<td>Opioids and Pain</td>
<td>2</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>1</td>
</tr>
<tr>
<td>Chronic lower respiratory disease</td>
<td>1</td>
</tr>
<tr>
<td>Neurodegenerative condition and cognition</td>
<td>1</td>
</tr>
<tr>
<td>Neurodegenerative condition and cognition</td>
<td>1</td>
</tr>
<tr>
<td>Other areas: Environmental exposures, Health disparities, Infectious disease, Body temperature, Women’s health, Pediatric data description, Hematologic disorders</td>
<td>13</td>
</tr>
<tr>
<td>Gap addressing projects: Ophthalmology</td>
<td></td>
</tr>
<tr>
<td>Compare Census and CDC data in AoU</td>
<td></td>
</tr>
<tr>
<td>Rare diseases</td>
<td>3</td>
</tr>
</tbody>
</table>
When patients enroll, EHRs from childhood may come along.

Questions:
- How much data is there?
- What data types are there?
- Can the data be used for pediatric research?
D38: Exploration of Pediatric Data from All of Us

Patients with:

- Drugs: 4070
- Conditions: 4865
- Visits: 5718
- Procedures: 4537
- Measurements: 3633
**Goal:** To replicate PheRS approach for three mendelian diseases including cystic fibrosis (CF), hereditary hemochromatosis (HH) and sickle cell (SC) anemia in allofus (AOU) cohort.

**Contributors:**

Jun Qian, Lisa Bastarache, David Schlueter, Janey Wang, Andrea Ramirez
Vanderbilt University Medical Center


Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. *J Am Med Inform Assoc.* 2019 Dec 1;26(12):1437-1447.
Raw PheRS (A) and studentized residuals of the raw PheRS (B) for three diseases in AOU EHR data. The regression model, PheRS~bs(age)+sex+race+unique_encounter_years was used for the r_PheRS calculation (B). P values were calculated using Wilcoxon rank sum test.
D44: Comparison of family health history gathered in survey and electronic health records in *All of Us* Registered Tier Data

- **Background and Methods**
  - **Family health history** is integral to clinical care and important to integrated precision medicine care.
  - Prior studies show gaps in data from Electronic Health Record (**EHR**) structured fields.
  - Prior studies have been assisted by **free text extraction**.
  - Surveys have shown great promise using common data models.

*All of Us has both data sources for family health history information.*
Data and Research Center – Family History in EHR vs. Surveys

Is EHR confirming what we already knew from surveying participants? Or are we getting new information?
Results and Conclusions

Participants with any Family History data

Total: ~52,700

EHR

- 28% (~14,800)

PPI

- 10% (~5,400)
- 62% (~32,500)
50 overall attendees, 26 demo project team members
9 out of 11 Consortium Awardees represented
22 different projects presented both on-site and remotely (of 38 total)
Discussions around feedback mechanism and new beta features
Discussions around authorship and alpha access next steps and publication of projects internally in Featured Workspaces
Data Release Timeline

*All of Us* is targeting a bold timeline. Projected dates are subject to change and dependent on security and usability testing.

**Data Browser Launch**
- **May 6, 2019**
  - Launch of an interactive tool available to the public that provides summary statistics from the program’s growing database.

**Researcher Workbench Beta Launch**
- **May 2020**
  - Release of the first version of the *All of Us* dataset and the first set of researcher tools for beta testing.
  - The dataset includes:
    - Survey data
    - Measurement data
    - Electronic health record (EHR) data

**Dataset Expansion**
- Expansion of the *All of Us* dataset with additional survey, measurement, and EHR data.
- Start of genotyping and whole genome sequencing; initial return of results to participants

**Additional Data Types and Tools 2021 & beyond**
- Expect to enroll 1M participants within 5-7 years of launch.
- Continue to expand the *All of Us* dataset with new data releases, including over time:
  - genomics data
  - wearables data (such as Fitbit)
  - new surveys
  - assay data
  - linkages to external data sources, etc.
- Release new versions of Researcher Workbench with more tools.
Data Release Timeline

*All of Us is targeting a bold timeline. Projected dates are subject to change and dependent on security and usability testing.

- Data Browser Launch May 6, 2019
  Launch of an interactive tool available to the public that provides summary statistics from the program’s growing database.

- Researcher Workbench Beta Launch May 2020
  Release of the first version of the All of Us dataset and the first set of researcher tools for beta testing.
  The dataset includes:
  - Survey data
  - Measurement data
  - Electronic health record (EHR) data

- Dataset Expansion
  Expansion of the All of Us dataset with additional survey, measurement, and whole genome sequencing; initial return of results to participants.

- Additional Data Types and Tools 2021 & beyond
  Expect to enroll 1M participants within 5-7 years of launch.
  Continue to expand the All of Us dataset with new data releases, including over time:
  - genomics data
  - wearables data (such as Fitbit)
  - new surveys
  - assay data
  - linkages to external data sources, etc.
  Release new versions of Researcher Workbench with more tools.
NIH’s All of Us Program Joins Fight Against COVID-19

Posted on June 16th, 2020 by Dr. Francis Collins

The future of health begins with you

We've learned so much about coronavirus disease 2019 (COVID-19), but there's still much more that we need to learn in order to defeat this devastating pandemic. Among the critical questions: why do some young people who are healthy and have no other health issues get as sick as someone who has chronic conditions?
Pivot to COVID-19

0. Enrollment paused in March

1. Serology study (see map)

2. EHR curation effort

3. New survey: longitudinal collection on health and well-being

“Aside from physical health concerns, the pandemic may be stressful to many of us. It can be difficult to cope with fear and anxiety, changing daily routines, a general sense of uncertainty, financial and economic hardships, social isolation, and/or stigmatization. I join you in acknowledging there are other unforeseen effects stemming from this global health pandemic.

Francis Collins, M.D., Ph.D., Director
Data Release Timeline

*All of Us* is targeting a bold timeline. Projected dates are subject to change and dependent on security and usability testing.

**Data Browser Launch**
*May 6, 2019*
Launch of an interactive tool available to the public that provides summary statistics from the program’s growing database.

**Researcher Workbench Beta Launch**
*May 2020*
Release of the first version of the *All of Us* dataset and the first set of researcher tools for beta testing.
The dataset includes:
- Survey data
- Measurement data
- Electronic health record (EHR) data

**Dataset Expansion**
Expansion of the *All of Us* dataset with additional survey, measurement, and whole genome sequencing; initial return of results to participants.

**Additional Data Types and Tools**
*2021 & beyond*
Expect to enroll 1M participants within 5-7 years of launch.
Continue to expand the *All of Us* dataset with new data releases, including over time:
- genomics data
- wearables data (such as Fitbit)
- new surveys
- assay data
- linkages to external data sources, etc.
Release new versions of Researcher Workbench with more tools.
Interested in becoming a Beta Researcher?

Visit researchallofus.org to apply!
All of Us Community and Provider Partner Network (as of January 2020)
All of Us Consortium Members (beyond community partners, as of March 2020)

The Participant Center

- Scripps Research Translational Institute
- Medscape
- patientslikeme
- Bloodworks Northwest
- BlueCross BlueShield
- New York Blood Center
- Quest Diagnostics
- Walgreens
- WebMD
- ACHA
- National Blood Collaborative Blood Center
- Commit for Life
- UK HealthCare
- Sage
- Health Partners
- All of Us California
- UC San Diego Health
- UCI Health
- UC Davis Health
- Cedars-Sinai
- USC University of Southern California
- UC Health
- Northwestern Medicine
- Brigham and Women’s Hospital
- Partners HealthCare
- Massachusetts General Hospital
- Boston Medical Center
- Rush University Medical Center
- Cook County Health
- University of Illinois Hospital & Medical Center
- UI Health
- All of Us New England
- All of Us Pennsylvania
- University of Pittsburgh
- Temple University
- University of Arizona
- Arizona Health
- Banner Health
- All of Us Wisconsin
- Madison Health
- University of Arizona Health System
- All of Us Pennsylvania
- University of Arizona Health Banner Health

HPO Network

(Health Care Provider Organizations)

- RMCs
- All of Us California
- UC San Diego Health
- UCI Health
- UC Davis Health
- Cedars-Sinai
- USC University of Southern California
- UC Health
- Northwestern Medicine
- Brigham and Women’s Hospital
- Partners HealthCare
- Massachusetts General Hospital
- Boston Medical Center
- Rush University Medical Center
- Cook County Health
- University of Illinois Hospital & Medical Center
- UI Health

- Trans America Consortium
- All of Us New England
- All of Us Pennsylvania
- University of Arizona and Banner Health

- All of Us Wisconsin
- Madison Health
- University of Arizona Health System

- New York City Consortium
- Columbia University Medical Center
- New York Presbyterian
- Weill Cornell Medicine

- All of Us Southern Network
- Trans America Consortium
- All of Us Wisconsin
- All of Us Pennsylvania
- All of Us New England
- New York City Consortium

- FQHCs (Federally Qualified Health Centers)
- VA Medical Centers
- VA
- U.S. Department of Veterans Affairs
- Cherokee
- Community Health Centers
- Community Health Centers Inc.
- Emory
- University of Alabama at Birmingham
- Health Services
- Jackson
- University of Florida
- University of Washington
- HudsonAlpha
- Baylor College of Medicine
- Johns Hopkins Medicine
- UTHealth
- Broad Institute
- color
- Partners HealthCare
- University of Utah

All of Us Participant Technology Systems Center (PTSC)

- Mayo Clinic
- Data & Research Center (DRC)
- Vanderbilt University Medical Center
- Broad Institute
- verily

Genomics Partners

- Columbia University Medical Center
- Northwell Medicine
- UTHealth
- ACHA
- WebMD
- ACHA
- National Blood Collaborative Blood Center
- Commit for Life
- UK HealthCare
- Sage
- Health Partners
- All of Us California
- UC San Diego Health
- UCI Health
- UC Davis Health
- Cedars-Sinai
- USC University of Southern California
- UC Health
- Northwestern Medicine
- Brigham and Women’s Hospital
- Partners HealthCare
- Massachusetts General Hospital
- Boston Medical Center
- Rush University Medical Center
- Cook County Health
- University of Illinois Hospital & Medical Center
- UI Health

Communications & Engagement

- WONDROS
- HCM
- U of Utah
- HealthEvolution
- careMD
- careMD Evolution
It takes *All of Us*...
For more information…

Thank you!

ResearchAllofUs.org
(to apply as a researcher!)

JoinAllofUs.org

@AllofUsResearch
#JoinAllofUs

andrea.h.ramirez@vumc.org
@AndreaRamirezMD

AllofUs.nih.gov

Precision Medicine Initiative, PMI, All of Us, the All of Us logo, and “The Future of Health Begins with You” are service marks of the U.S. Department of Health and Human Services.
Additional Slides
Screenshot Walk Thru
## Underrepresented in Biomedical Research (UBR) Definitions

<table>
<thead>
<tr>
<th>Category</th>
<th>All of Us</th>
<th>UBR definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ancestry</strong> <em>(Race/Ethnicity)</em></td>
<td>Participant identifies as any race or ethnicity other than White, including Multi-Ancestry (any combination of 2 or more races)</td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>Participant is under 18 or 65 or over at time of consent</td>
<td></td>
</tr>
</tbody>
</table>
| **Sexual & Gender Minorities** | **Biological sex at birth**: Participant self-reports intersex as their sex assigned at birth  
**Gender Identity (GI)**: Participant identifies as anything other than man or woman (e.g., non-binary, transgender, etc.)  
**Sexual Orientation (SO)**: Participant identifies as anything other than straight (e.g., gay, lesbian, bisexual, etc.) |                                                                                 |
| **Income**                | Participant with household income equal to or below two times the Federal Poverty Level |                                                                                 |
| **Educational Attainment** | Participant is currently without a high school diploma or GED              |                                                                                 |
| **Geography**             | Participant’s zip code is within the list of Rural/Non-Metropolitan zip codes (per HRSA’s Federal Office of Rural Health Policy) |                                                                                 |
| **Access to Care**        | Participant cannot readily use the health care system or pay for needed care |                                                                                 |
| **All of Us**             |                                                                             |                                                                                 |
Questions & Answers

Please submit all questions through the question box.
Summary & Evaluation

• Thank you to all our panelists.
• A recording of the webinar will be available on the CMSS website in the coming weeks.
• Please compete a short evaluation following the webinar.
• For more information, contact info@cmss.org.
CMSS WEBINAR SERIES

Advancing Clinical Registries to Support Pandemic Treatment and Response

The series will address key questions related to the rapid development, deployment and implementation of Covid-19 focused clinical registries and clinical repositories by specialty societies and academia.

SUMMER 2020 | FREE TO ATTEND

Upcoming Webinars:

Aug. 12
Prioritizing Patient Engagement and Inclusion of Patient-generated Data

Sept. 1
Using Clinical Registries to Address Disparities in Covid-19
Upcoming Webinar:

Prioritizing Patient Engagement and Inclusion of Patient-Generated Covid-19 Data

Aug. 12 | 1:00 - 2:30 pm ET

CMSS WEBINAR SERIES
Advancing Clinical Registries to Support Pandemic Treatment and Response

Moderator:
Susannah Fox
Principal, Internet Geologist LLC;
Former Chief Technology Officer,
US Department of Health and
Human Services

Panelists:
Gina Assaf
Independent digital design consultant, Leading the Patient-Led Research Group within Body Politic

Emily Sirotich
COVID-19 Global Rheumatology Alliance; Canadian Arthritis Patient Alliance; Department of Health Research Methods, Evidence, and Impact, McMaster University

Host:
Helen Burstin, MD, MPH, MACP
Chief Executive Officer
Council of Medical Specialty Societies (CMSS)

Gary Wolf
Co-Founder, Quantified Self

Hannah Davis
Researcher with the Patient-Led Research Group within Body Politic