National Institutes of Health

All of Us RESEARCH PROGRAM | The Future of Health Begins With You

Curating Data for the *All of Us* Research Program

All of Us Research Program

Mission

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

How

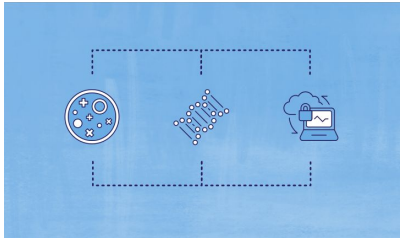
- Collect genomic and EHR data for 1+ million participants

http://allofus.nih.gov

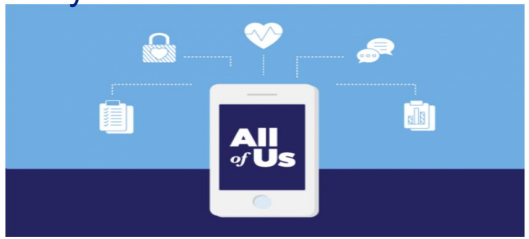
All of Us Research Program - Components

Biobank

Genome Center

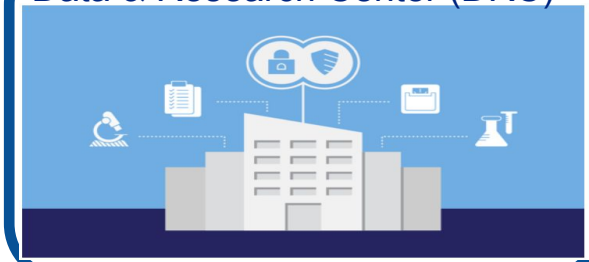


Participant Technology Systems Center

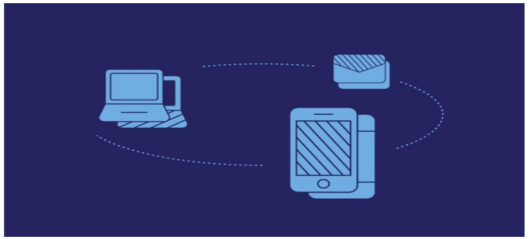


Health Care Provider Organizations (HPO)

Data & Research Center (DRC)



Participant Center



All of Us Research Program

U.S. Department of Health & Human Services › National Institutes of Health

NIH National Institutes of Health
All of Us Research Program

ABOUT ⌄ FUNDING ⌄ NEWS, EVENTS, & MEDIA JoinAllofUs.org › Search

The future of health begins with All*of*Us

The *All of Us* Research Program is a historic 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, environment, and biology, researchers will uncover paths toward delivering precision medicine.

WATCH VIDEO

We're interested in your research ideas for our upcoming Research Priorities Workshop.

LEARN MORE

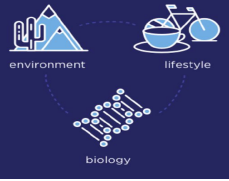
Sign up to be notified of announcements, events, funding news and more.

SUBSCRIBE

We are building a research program of 1,000,000+ people

The mission of the *All of Us* Research Program is to accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us.

ABOUT THE SCALE & SCOPE



All of Us

Submit New Idea

WELCOME IDEAS FAQ

Recent Popular

AOU RESEARCH PRIORITIES USE CASES

Could the Precision Medicine Initiative Reduce Social Inequality?

The proposed study would identify microbial communities associated with stigmatized conditions such as idiopathic malodor or mental health.

Submitted by Irene Gabashvili (@drirene) on 17th Jan | 31 comments

64 votes

IDEATE

AOU RESEARCH PRIORITIES USE CASES

The effects of alcohol (and other drug) consumption on major somatic diseases and psychiatric disorders and their treatments.

Alcohol affects nearly all major diseases. Failing to capture lifetime alcohol exposure will adversely affect the goals of All of Us.

There are better questions about consumption, that should be asked of previous year, typical year and heaviest year.

Questions should be asked about problems due to alcohol, and those should be focused on lifetime.

Additional questions about other drugs should also be included.

Submitted by Howard Edenberg (@edenberg) on 12th Jan | 10 comments

50 votes

IDEATE

AOU RESEARCH PRIORITIES USE CASES

What is the contribution of dietary patterns and nutritional status to chronic disease susceptibility and prevention?

The top four causes of deaths in the U.S. are diet related and many other troublesome diseases are affected by diet, It will be important therefore to capture nutritional status, food environment, and repeated 24h (ASA24s) dietary intakes along with measures that can validate dietary measures of public policy interest including added sugar, energy, protein, sodium, iron, vitamins, dietary supplements, potassium, fruits ...more »

Submitted by Christopher Lynch (@measurenutrition) on 19th Dec 2017 | 17 comments

41 votes

IDEATE

AOU RESEARCH PRIORITIES USE CASES

How do autoimmune diseases start?

Many autoimmune diseases are characterized by the presence of autoantibodies. These may be present long before disease starts, but what triggers the transition from autoimmunity (the presence of the antibodies) to disease is unknown, but is likely to include both environmental exposures and genetic susceptibility. What are the risk factors for the transition to disease? Are these modifiable? In which patients is the risk ...more »

Submitted by Robert Carter (@carterrob) on 18th Oct 2017 | 5 comments

40 votes

IDEATE

S Stokes, Michael
Campaign Owner

Campaign Funnel

ALL STAGES 254 IDEAS

IDEATE 254 IDEAS

Leaderboard [-]

M	Mary McDonald 2039 points	1
N	NICHD All of Us Team 826 points	2
	Irene Gabashvili 353 points	3
J	Jane Atkinson 280 points	4
M	Mohamed Salah Noshi 268 points	5
	Abraham Palmer 222 points	6
C	Christopher Lynch 201 points	7
E	Howard Edenberg 181 points	8
	David Nation 168 points	9
S	Susan Plawsky 160 points	10

[View All]

Subscribe [-]

New Ideas RSS Feed
Top Ideas RSS Feed
New Comments RSS Feed

Campaigns

AoU Research Priorities Use

powered by ideascale

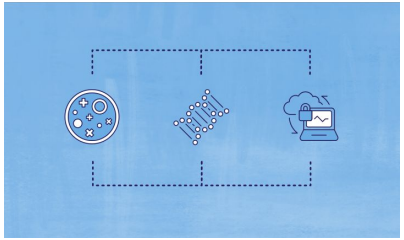
Data and Research Center

- Awardees
  - Vanderbilt University
  - Broad Institute
  - Verily (Google)

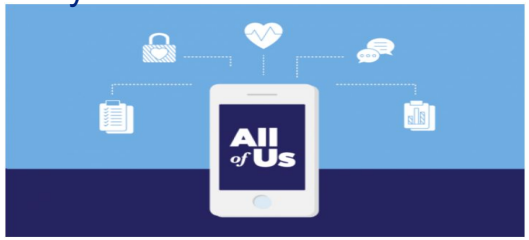
All of Us Research Program - Components

Biobank

Genome Center

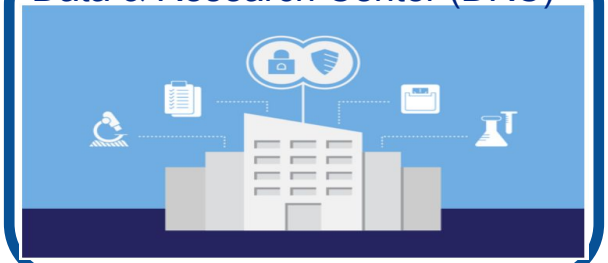


Participant Technology Systems Center

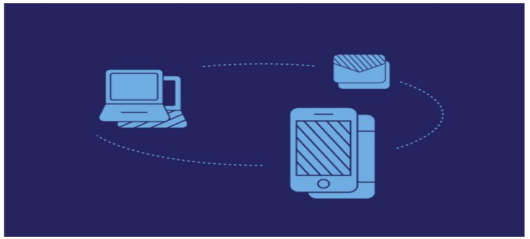


Health Care Provider Organizations (HPO)

Data & Research Center (DRC)



Participant Center



EHR Data

Curation

Curated Data Repository (CDR)

- The CDR is the resource that contains all study data that researchers can access.

- Currently collects data from:
  - Participant Provided Information (PPI)
  - Physical Measurements (PM)
  - EHR Information uploaded by research medical centers (RMC) and federally qualified health centers (FQHC)

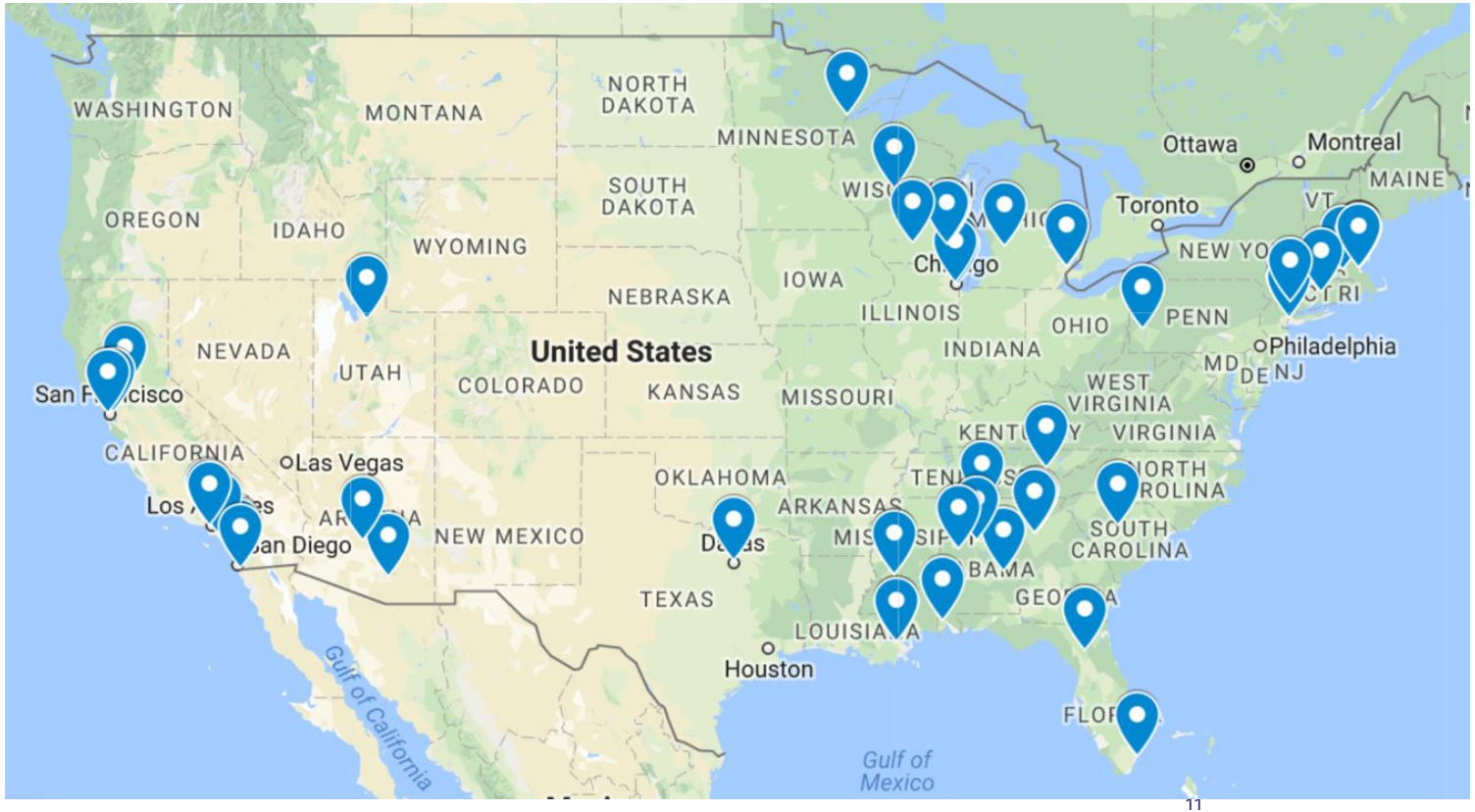
Curation Objectives

1) Develop and implement a pipeline to generate CDR

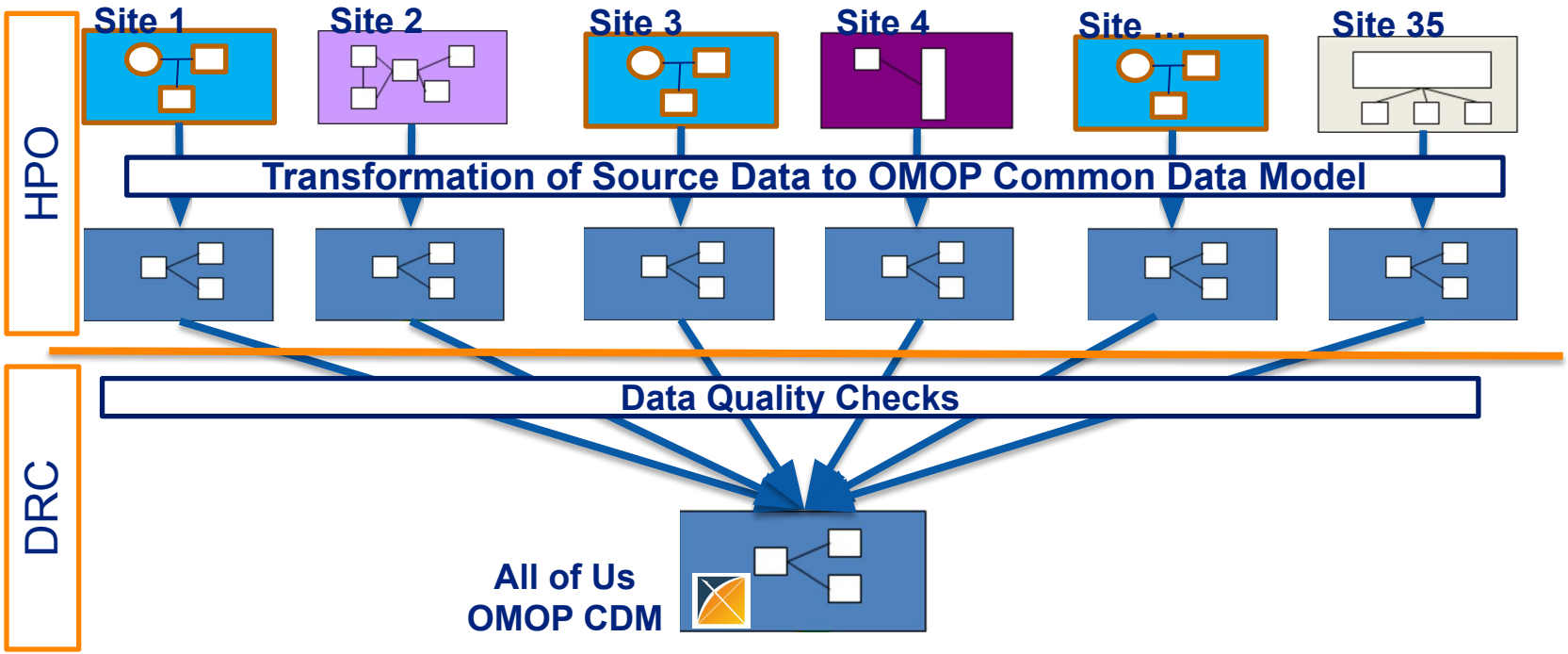
2) Create data quality checks
   a) Concordance, Completeness, Plausibility, and Currency

3) Extend infrastructure to receive, store, and harmonize data from new data sources (i.e. wearables).

All of Us Research Program HPO Sites



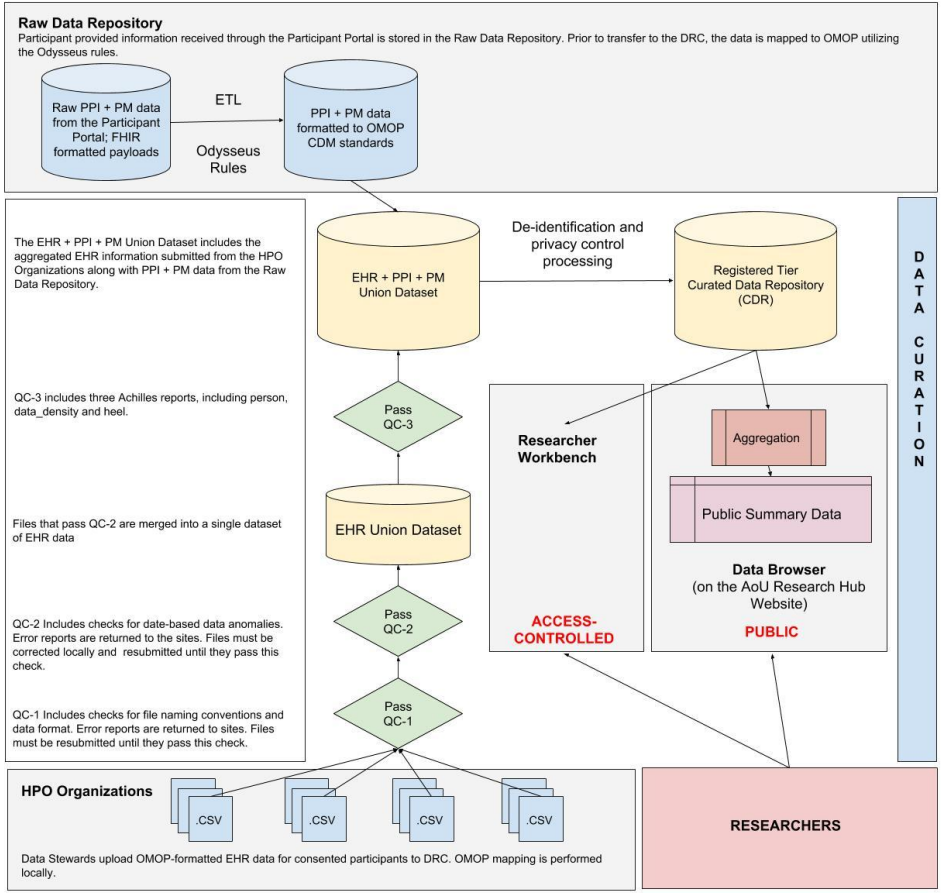
Aggregating EHR Data



Source: https://www.ohdsi.org/data-standardization/

Curation Overview

- Three phases
  - File structure checks
  - Site level content error checking
  - Aggregate data checks

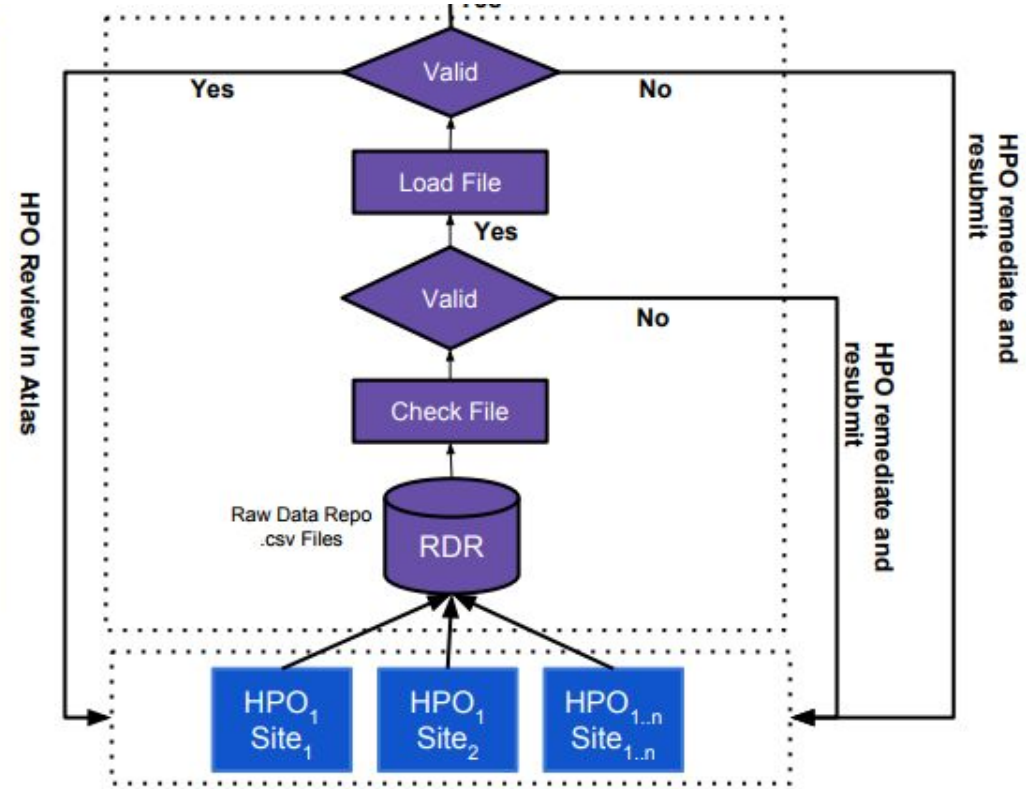


Quality Check 1 (QC-1): File Validation & Initial Data Quality Checks

QC-1:

- Processing happens at the DRC or Locally

- File validation against specification
  - File names
  - Column names and order
  - Column type



Local File Validation

- https://github.com/all-of-us/aou-ehr-file-check
- Checks
  - File names
  - Column names and order
  - Column type

README.md

AoU EHR Submission Validator

Validate submissions for the All of Us data sprints

Requirements

- Python 2.7.* or Python 3 (download from here and install)
- pip (download get-pip.py and run `python get-pip.py`)

Installation / Configuration

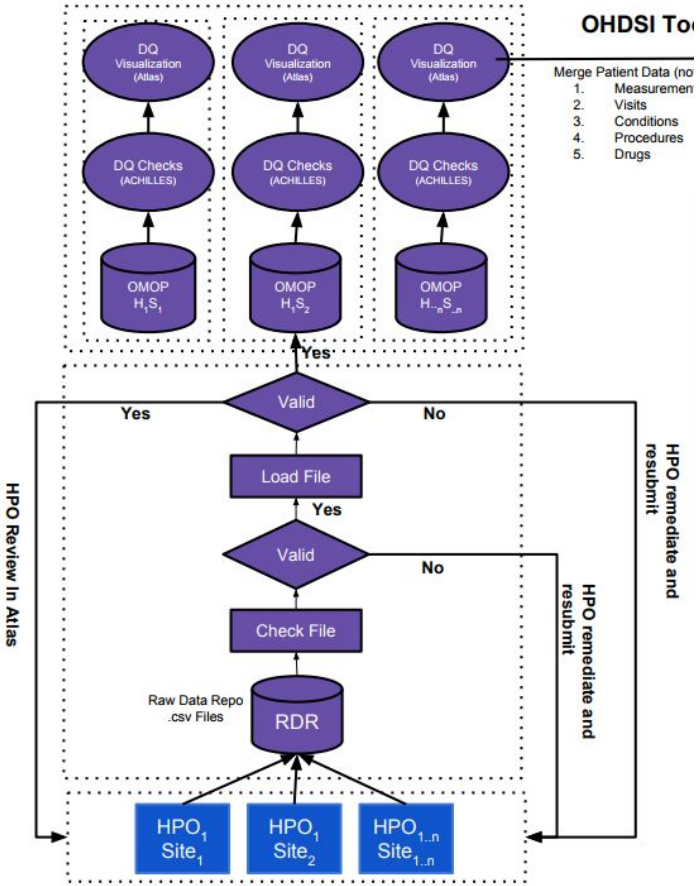
- Install requirements by running

```
pip install -r requirements.txt
```

Quality Check 2 (QC-2): Comprehensive Pre-Aggregation Data Quality Checks

QC-2:

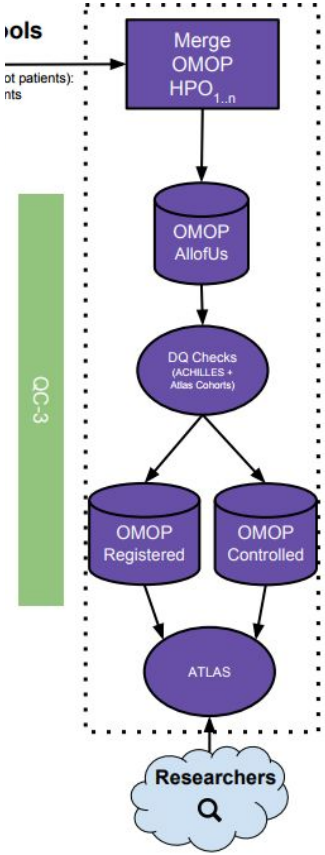
- Processing happens at the DRC; checks occur at the site level, pre-aggregation
- More in-depth checks are completed using ACHILLES tool and custom checks.
- Identifies abnormalities, such as visit before date of birth
- Reports are returned to sites, up to sites to correct errors



Quality Check 3 (QC-3): Final Post-Aggregation Data Quality Checks

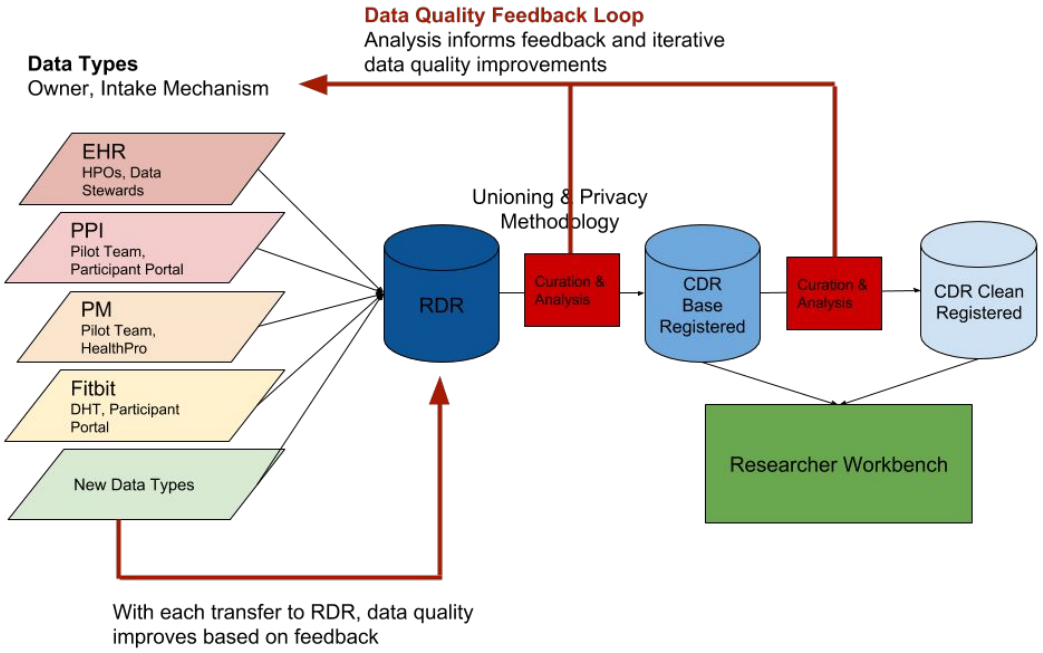
QC-3:

- In between QC-2 and QC-3 data is aggregated
- QC-3 occurs post-aggregation across all site data
- Duplications are removed
- Checks for phenotype completeness
- At this stage, DRC corrects any errors identified; sites are not involved
- Data is then “tiered” into access levels and provided to researchers with the appropriate level of access



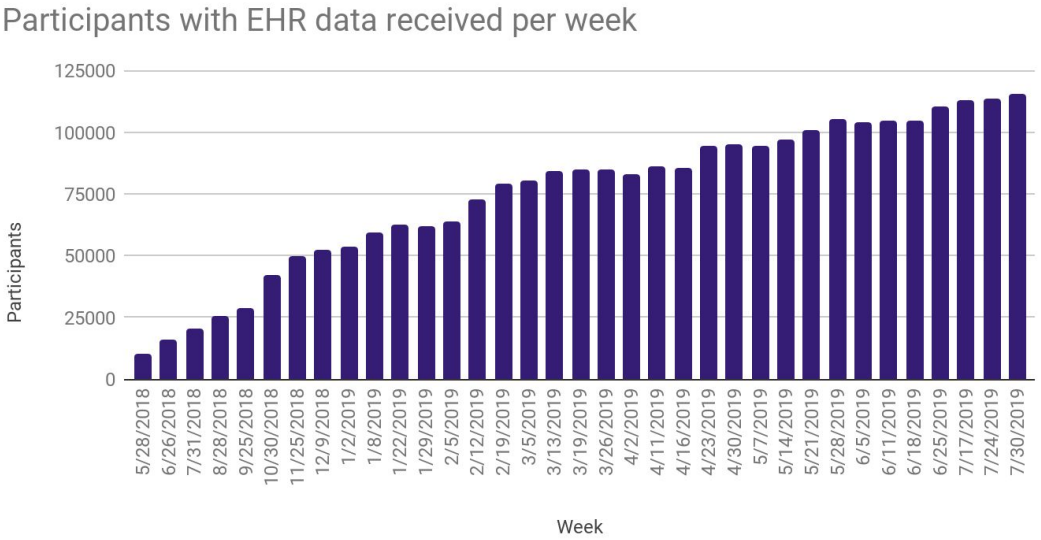
EHR Operations Data Quality Feedback Loop

- 2x monthly EHR Operations calls with HPO Data Stewards
- With each data transfer, a set of standard reports are provided:
  - Achilles Reports:
    - Person
    - Achilles_Heel
    - Data_Density
- Additional feedback about baseline data quality requirements
  - Automated reports are being developed
  - 1:1 interactions with HPOs to improve data quality



EHR Counts

- **115K** Total number of participants with EHR data transferred to DRC
- **34** Sites
  - **6** FQHCs
  - **28** Sites part of an RMC

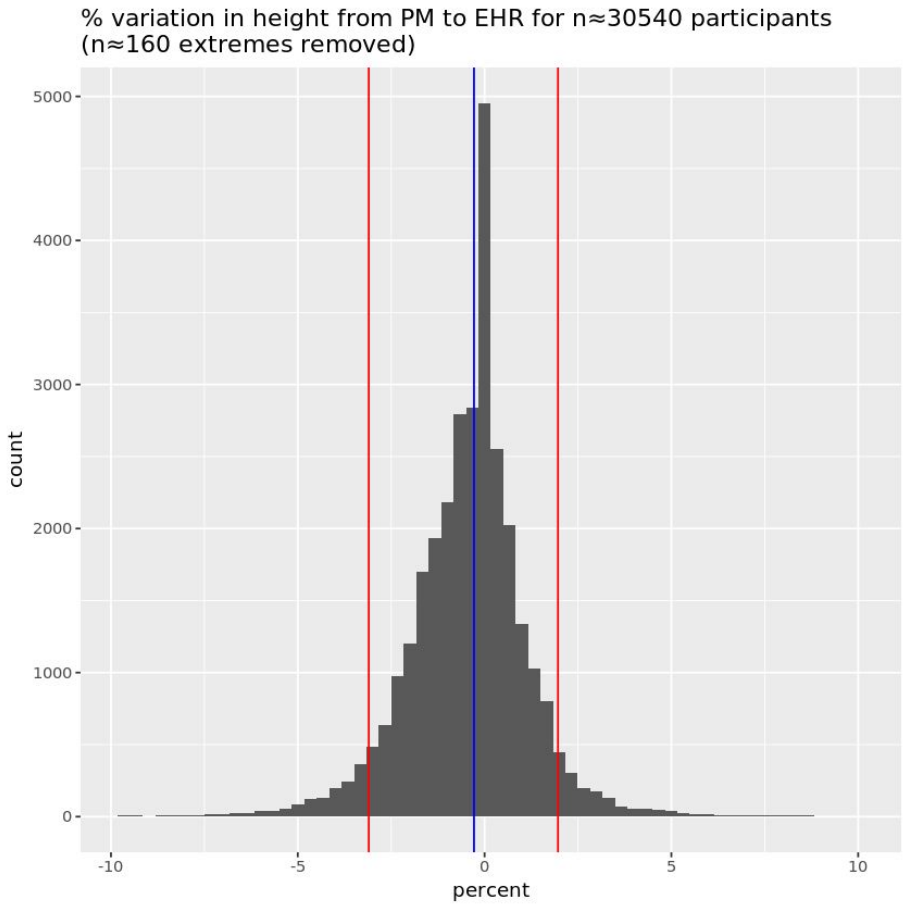


Analysis 5: ***Case Study:*** Height Comparison (EHR + Physical Measurements)

- Physical Measurements data: height measured in centimeters.
- EHR data: most recent height for each individual, normalized to centimeters using 2.54cm = 1in.
- Percent difference of the two, using PM Height as the standard:

  $$\frac{\langle \text{PM Height} \rangle - \langle \text{EHR Height} \rangle}{\langle \text{PM Height} \rangle}$$

- Red lines represent 5th and 95th percentiles (-3.1% and 2.0% or -5.2cm to 3.3cm).
- Blue line is median (-0.4% or -0.48cm)



Scaling Curation Efforts

= new data types

Current

Curation Co-Chairs

- Baseline EHR Ops & Curation Development Team
- Baseline Quality Assurance, EHR + PPI + PM Team
- Baseline Privacy for Registered Tier, EHR + PPI + PM Team

Next 18 Months

EHR Curation
Product Manager
- EHR Quality Team
- Development Team

Sys Admin / Sys Ops Mgr

Data Quality Dashboards Team

Quality Assurance
Product Manager
- **Quality Assurance Methodology Teams** *Teams by Data Types*
- **Dev Teams** *Teams by Data Types*

Quality Control

Data Validation Teams

Privacy
Product Manager
- **Privacy Methodology Teams** *Teams by Data Types*
- **Dev Teams** *Teams by Data Types*

In support of EHR, PPI, PM data only, 1 data tier

Additional data types added + 2 tiers data

Next 36 Months

EHR Curation
Product Manager
- EHR Quality Team
- Development Team

Sys Admin / Sys Ops Mgr

Data Quality Dashboards Team

Quality Assurance
Product Manager
- **Quality Assurance Methodology Teams** *Teams by Data Types*
- **Dev Teams** *Teams by Data Types*

Privacy
Product Manager
- **Quality Assurance Methodology Teams** *Teams by Data Types*
- **Dev Teams** *Teams by Data Types*

QC

QC

Data Validation

Data Validation

Tools

Data Browser (Public)

https://databrowser.researchallofus.org/

Search Across Data Types

Keyword Search

Data based on Curated Data Repository (CDR) dated 11/13/2018 with 116,460 total participants.

FAQs

Introductory Videos

User Guide

EHR Domains:

Conditions
13,614
medical concepts
36,260 participants in this domain
View Top Conditions

Drug Exposures
14,967
medical concepts
33,440 participants in this domain
View Top Drug Exposures

Labs and Measurements
7,733
medical concepts
32,480 participants in this domain
View Top Labs and Measurements

Procedures
13,229
medical concepts
35,320 participants in this domain
View Top Procedures

Survey Questions:

The Basics
14
survey questions
104,440 participants in this domain
Survey includes participant demographic information.
View Complete Survey

Overall Health
16
survey questions
101,420 participants in this domain
Survey provides information about how participants report levels of individual health.
View Complete Survey

Lifestyle
7
survey questions
100,460 participants in this domain
Survey includes information on participant smoking, alcohol and recreational drug use.
View Complete Survey

Help

All of Us RESEARCH PROGRAM | The Precision Medicine Initiative
RESEARCHER WORKBENCH

Workspaces > Workspace > Cohorts >

Add a Cohort

ABOUT | COHORTS | CONCEPTS | NOTEBOOKS

Include Participants

ADD CRITERIA

Program Data
- Surveys
- Physical Measurements

Domains
- Demographics
- Conditions
- Procedures
- Drugs
- Measurements
- Visits

All of Us RESEARCH PROGRAM The Precision Medicine Initiative
RESEARCHER WORKBENCH

Workspaces > Workspace > Cohorts >

ABOUT

Include Participants

ADD CRITERIA

ADD CRITERIA

CONDITIONS | SEARCH SNOMED | MODIFIERS

ICD9 CODES

diabetes

- 001-999.99 DISEASES AND INJ
- V01-V91.99 FACTORS INFLUEN
- E000-E999.9 EXTERNAL CAUS

- 250 Diabetes mellitus
- 250.00 Diabetes mellitus without mention ... plication, type II or ...
- 250.02 Diabetes mellitus without mention o ... lication, type II or ...
- 250.01 Diabetes mellitus without mention of complication, type I [juv...
- V77.1 Screening for diabetes mellitus
- 250.60 Diabetes with neurological manifestations, type II or unspeci...
- 357.2 Polyneuropathy in diabetes
- 250.03 Diabetes mellitus without mention of complication, type I [ju...
- 250.40 Diabetes with renal manifestations, type II or unspecified typ...
- 250.80 Diabetes with other specified manifestations, type II or unspe...
- 250.62 Diabetes with neurological manifestations, type II or unspeci...
- 250.50 Diabetes with ophthalmic manifestations, type II or unspecifi...
- 250.42 Diabetes with renal manifestations, type II or unspecified typ...
- 250.90 Diabetes with unspecified complication, type II or unspecifie...

Selected Criteria

CANCEL NEXT FINISH

All of Us RESEARCH PROGRAM | The Precision Medicine Initiative
RESEARCHER WORKBENCH

Workspaces > Workspace > Cohorts >

ABOUT

Include Participants

ADD CRITERIA

ADD CRITERIA

CONDITIONS | SEARCH SNOMED | MODIFIERS

ICD9 CODES

Diabetes mellitus

- 240 Simple and unspecified goiter 6,581
- 241 Nontoxic nodular goiter 16,227
- 242 Thyrotoxicosis with or without goiter 8,235
- 243 Congenital hypothyroidism 655
- 244 Acquired hypothyroidism 48,771
- 245 Thyroiditis 10,383
- 246 Other disorders of thyroid 7,050
- 249 Secondary diabetes mellitus 2,108
- 250 Diabetes mellitus 80,157
- ther disorders of pancreatic internal secretion 10,253
- isorders of parathyroid gland 4,824
- 253 Disorders of the pituitary gland and its hypothalamic control 6,997
- 254 Diseases of thymus gland 311
- 255 Disorders of adrenal glands 8,827
- 256 Ovarian dysfunction 10,083
- 257 Testicular dysfunction 6,410
- 258 Polyglandular dysfunction and related disorders 488
- 259 Other endocrine disorders 5,937
- 260 Kwashiorkor 317

Selected Criteria

CANCEL NEXT FINISH

All of Us RESEARCH PROGRAM | The Precision Medicine Initiative
RESEARCHER WORKBENCH

Workspaces > Workspace > Cohorts >

ABOUT

Include Participants

ADD CRITERIA

ADD CRITERIA

CONDITIONS | SEARCH SNOMED | MODIFIERS

ICD10 CODES

Type 2 diabetes mellitus

- E03 Other hypothyroidism 14,002
- E04 Other nontoxic goiter 3,555
- E05 Thyrotoxicosis [hyperthyroidism] 1,318
- E06 Thyroiditis 1,821
- E07 Other disorders of thyroid 870
- E08 Diabetes mellitus due to underlying condition 268
- E09 Drug or chemical induced diabetes mellitus 153
- E10 Type 1 diabetes mellitus 4,078
- E11 Type 2 diabetes mellitus 22,187
- E13 Other specified diabetes mellitus 331
- E15 Nondiabetic hypoglycemic coma 5
- E16 Other disorders of pancreatic internal secretion 766
- E20 Hypoparathyroidism 218
- E21 Hyperparathyroidism and other disorders of parathyroid gland 1,103
- E22 Hyperfunction of pituitary gland 571
- E23 Hypofunction and other disorders of the pituitary gland 698
- E24 Cushing's syndrome 191
- E25 Adrenogenital disorders 81
- E26 Hyperaldosteronism 98

Selected Criteria

ICD9

Group 250Diabetes mellitus

ICD10

OR Group E10Type 1 diabetes mellitus

OR Group E11Type 2 diabetes mellitus

CANCEL | NEXT | FINISH

All of Us RESEARCH PROGRAM | The Precision Medicine Initiative
RESEARCHER WORKBENCH

Workspaces > Workspace > Cohorts >

Add a Cohort

ABOUT | COHORTS | CONCEPTS | NOTEBOOKS

Include Participants

DELETE GROUP

Contains Conditions Codes | 82,661

ADD CRITERIA

Group Count: 82,661

AND

DELETE GROUP

ADD CRITERIA

Program Data
- Surveys
- Physical Measurements

Domains
- Demographics
- Conditions
- Procedures
- Drugs
- Measurements
- Visits

And Exclude Participants

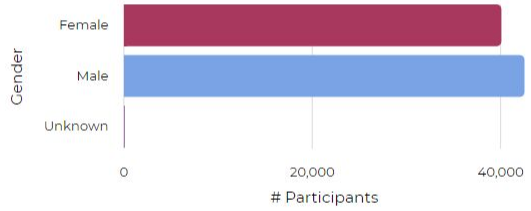
DELETE GROUP

ADD CRITERIA

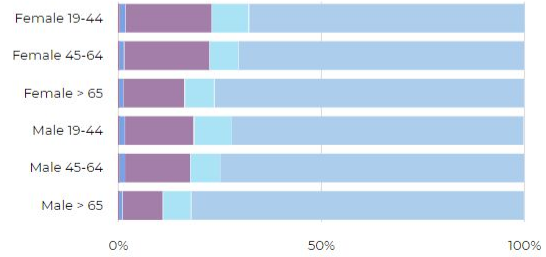
Total Count: 82,661

SAVE COHORT

Results by Gender



Results By Gender, Age Range, and Race



Researcher Workbench (Private)

All of Us RESEARCH PROGRAM | The Precision Medicine Initiative
RESEARCHER WORKBENCH

K Karthik Natarajan

Researcher Workbench

⚠ In order to get access to data and tools please complete the following steps:

STEP 1

Turn on Google 2-Step Verification

Add an extra layer of security to your account by providing your phone number in addition to your password to verify your identity upon login.

✓ COMPLETED

STEP 2

Complete Online Training

Complete mandatory compliance training courses on how data should be used and handled.

✓ COMPLETED

STEP 3

Login to eRA Commons

Link to your eRA Commons account to the workbench to gain full access to data and tools.

LOGIN

Quick Tour & Videos

How to Use the All of Us Researcher Workbench

See all documentation

Working Within Researcher Workbench

Workspace

Working with Notebooks

jupyter A1c comparison (unsaved changes)

File Edit View Insert Cell Kernel Navigate Widgets Help

Cluster aou-rw-stable-32/all-of-us not found | Not Trusted | Python 3

Run | Code

Step 1: Load libraries to import data, conduct analyses, and produce plots

In [3]:
```
# load libraries to allow data import
from aou_workbench_client.cdr.model import *
from aou_workbench_client.data import load_data
from IPython.display import display, HTML
# load scientific computing library
import numpy as np
#load 2D plotting library
import matplotlib.pyplot as plt
```

Step 2: Enter input variables

1. "name_of_cohort_x" => The cohort you want to reference in this Notebook.
2. "table_name" => The OMOP table you would like to pull from.
3. "name_of_concept_set" => The concept set you want to reference in this Notebook.

In [4]:
```
# input variables
# cohorts to reference in this notebook
name_of_cohort_a = "DM with a1c"
name_of_cohort_b = "No DM with a1c"
# the OMOP table to pull from
table_name = Measurement
# concept set to reference in this notebook
name_of_concept_set = "Hemoglobin A1c"
```

jupyter A1c comparison Autosave Failed!

File Edit View Insert Cell Kernel Navigate Widgets Help

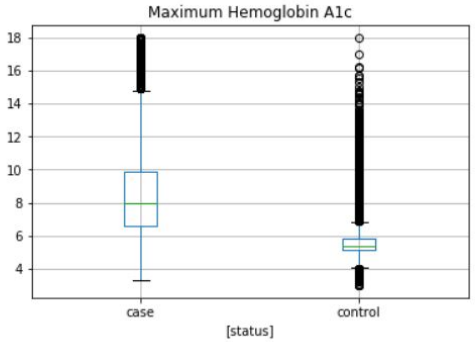
Cluster aou-rw-stable-32/all-of-us not found Trusted Python 3

Run Code

Step 5: Generate boxplots to compare maximum A1C values between the groups.

In [9]:
```
plot=a1cs.boxplot(by='status').set_title("Maximum Hemoglobin A1c")
plt.suptitle("")
```

Out[9]:
```
Text(0.5, 0.98, '')
```



In [ ]:

Topics Still Debating

- What is a “Concept Set”?
- How to convey what OMOP is to the average user?
  - OHDSI and AoU videos and Book of OHDSI will help
- How to distill the quality of the data to the researcher?
- How to track progress across sites?

Questions