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
- Data & Research Center (DRC)
- Genome Center
- Participant Technology Systems Center
- Health Care Provider Organizations (HPO)
- Participant Center
**AU 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 depression, alcoholism, or mental health.

Submitted by Wayne Saba (All of Us) on 17th Jan | 19 comments

**AU RESEARCH PRIORITIES USE CASES**

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

Alcohol effects 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 Edelberg (Challenger) on 13th Jan | 19 comments

**AU 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 death in the U.S. are diet-related and many other chronic diseases are influenced by diet and lifestyle. A number of dietary patterns have been described by researchers and are of interest to the public and to policy makers. More detailed research is needed to understand the relationship between dietary patterns and disease risk, and to understand the role of behavioral science principles in promoting healthy dietary habits. This could be done through research, pilot projects, and public health programs. In addition, the impact of diet on health outcomes (including obesity, diabetes, and other chronic diseases) should be examined with a focus on low-income populations.

Submitted by Christopher Lynch (MelroseWakefield) on 15th Dec 2017 | 17 comments

**AU 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 (the reaction)? Is it likely that we will both environmental factors and underlying genetics. How can we understand the transition to disease? Are these modifiable? In which patients is the risk ...more...

Submitted by Robert Guter (Challenger) on 18th Oct 2017 | 5 comments
Data and Research Center

Help build vibrant research community

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

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

Provide research support and analysis tools

Acquire, organize and provide the largest, richest biomedical dataset securely
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
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

Transformation of Source Data to OMOP Common Data Model

Data Quality Checks

All of Us
OMOP CDM

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

- Checks
  - File names
  - Column names and order
  - Column type
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 $2.54\text{cm} = 1\text{in}$.
- Percent difference of the two, using PM Height as the standard:
  $$\frac{\text{<PM Height> - <EHR Height>}}{\text{<PM Height>}}$$
- 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

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

Additional data types added + 2 tiers data

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

EHR Curation Product Manager
Development Team
Data Quality Dashboards Team

Quality Assurance Product Manager
Quality Assurance Methodology Teams
Dev Teams
Quality Control
Data Validation Teams

Privacy Product Manager
Privacy Methodology Teams
Dev Teams

Next 18 Months

Next 36 Months

Sys Admin / Sys Ops Mgr

EHR Curation Product Manager
Development Team

Quality Assurance Product Manager
Dev Teams

Quality Assurance Methodology Teams
Dev Teams

Privacy Product Manager
Dev Teams

Quality Assurance Methodology Teams
Dev Teams

Privacy Product Manager
Dev Teams

QC
Data Validation
Data Quality Dashboards Team

Teams by Data Types
Teams by Data Types
Teams by Data Types
Teams by Data Types

= new data types

Current

In support of EHR, PPI, PM data only, 1 data tier
Tools
Data Browser (Public)

https://databrowser.researchallofus.org/

Search Across Data Types

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
ICD10 CODES

Type 2 diabetes mellitus

Selected Criteria

ICD9
- Group 250 Diabetes mellitus

ICD10
- OR Group E10 Type 1 diabetes mellitus
- OR Group E11 Type 2 diabetes mellitus
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.

**STEP 2**
Complete Online Training
Complete mandatory compliance training courses on how data should be used and handled.

**STEP 3**
Login to eRA Commons
Link to your eRA Commons account to the workbench to gain full access to data and tools.

Quick Tour & Videos

How to Use the All of Us Researcher Workbench
A1c comparison

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 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.

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

```python
In [9]: plot=aics.boxplot(by='status').set_title("Maximum Hemoglobin A1C")
plt.suptitle"
Out[9]: Text(0.5, 0.98, '')
```
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