

ELSI Workshop June 24, 2019

# *All of Us* Program Data

Where we are and  
what is to come

**All of Us**  
RESEARCH PROGRAM

The  
Future of  
Health Begins  
With You



Jennifer Ayala, PhD  
*All of Us* Data and Research Center  
Vanderbilt University Medical Center

## Outline

---

- *All of Us* Scientific Framework
- Current data types
- Data Browser walkthrough
- Future data types
- Data lifecycle
- Researcher Workbench preview

**All of Us SCIENTIFIC FRAMEWORK = ENABLE RESEARCH THAT WILL:**

**I. Increase wellness and resilience, and promote healthy living**

**II. Reduce health disparities and improve health equity in underrepresented in biomedical research (UBR) populations**

**III. Develop improved risk assessment and prevention strategies to preempt disease**

**IV. Provide earlier and more accurate diagnosis to decrease illness burden**

**V. Improve health outcomes and reduce disease/illness burden through improved treatment and development of precision interventions**

## Current *All of Us* Data Types

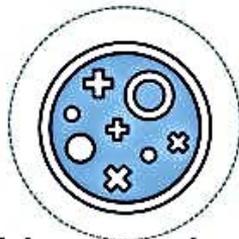
---



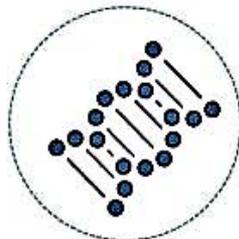
electronic health  
records



surveys



bio-samples



genomics



physical  
measurements



digital health

# Participant Provided Information

---

- **Current surveys**

- The Basics (*demographics, employment, income, contact information*)
- Overall Health (*general health, daily activities, and women's health*)
- Lifestyle (*tobacco, alcohol and recreational drug use*)
- Health Care Access and Utilization (*access to and use of health care*)
- Family Health History (*medical history of immediate biological family members*)
- Personal Medical History (*current and past medical conditions, approximate age of diagnosis*)



- **Upcoming surveys**

- Mental Health (*personality, mental health condition diagnosis, depression, suicidality, adversity/trauma, assault, specific disorders*)
- Social Determinants of Health (*optimism, social and emotional support, loneliness, perceived stress, discrimination, social status, food insecurity, neighborhood disorder, religion/spirituality*)
- Diet (*food/eating behavior*)

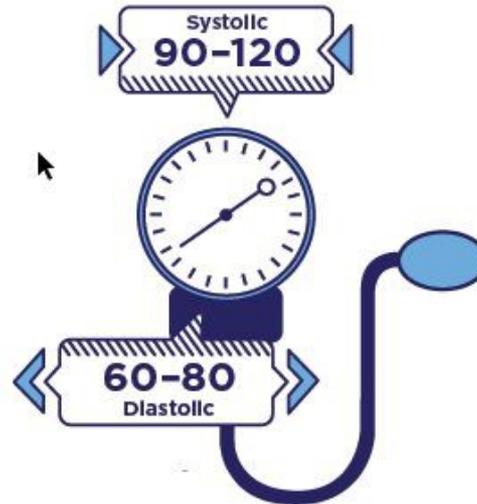
# Participant Provided Information

- **Physical measurements**

- Height
- Weight
- BMI
- Waist circumference
- Hip circumference
- Blood pressure
- Heart rate
- Heart rhythm status
- Wheelchair use
- Pregnancy



(These apply to everyone.)



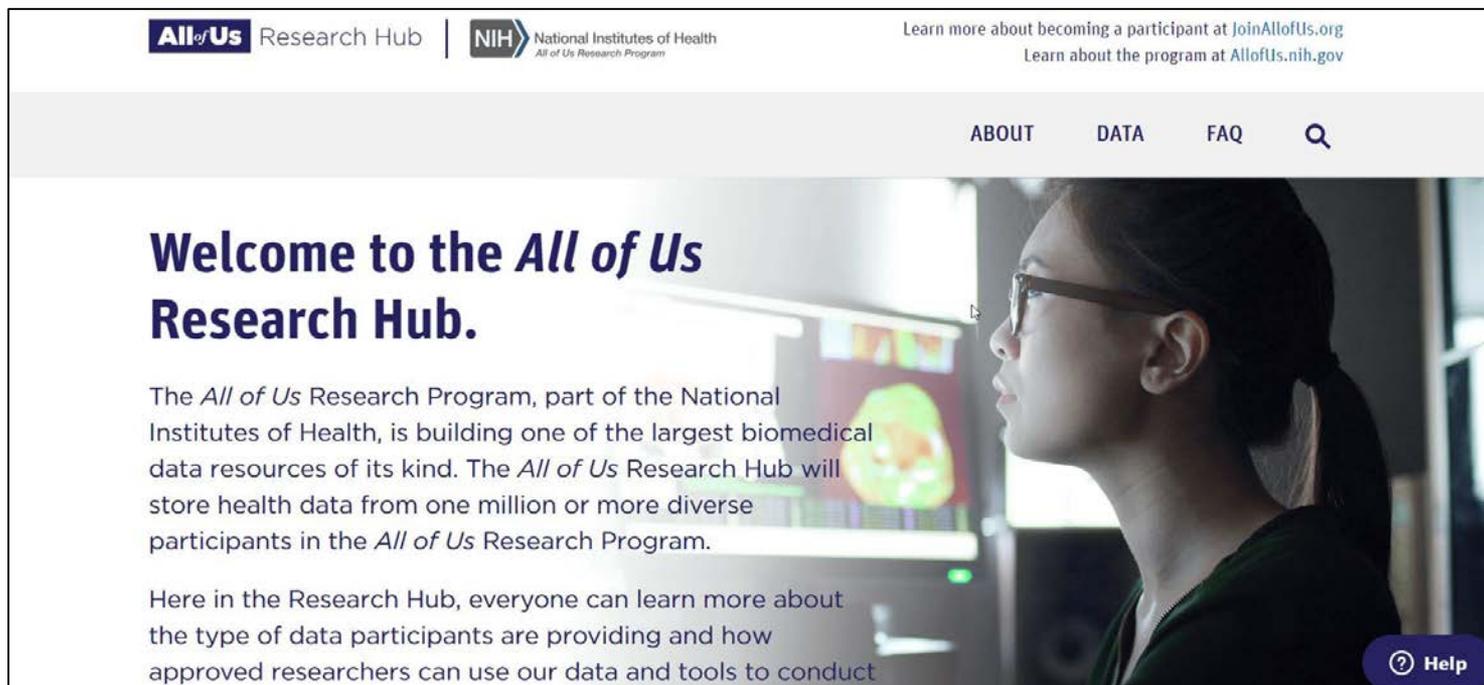
# Electronic Health Records

---

- **Current**
  - Labs
  - Measurements
  - Conditions
  - Procedures
  - Drugs/medications
  - Visit type
- **Future**
  - Notes
  - Images *(CT, MRI, etc.)*



# www.researchallofus.org



The screenshot shows the homepage of the All of Us Research Hub. At the top left, there are logos for "All of Us Research Hub" and "NIH National Institutes of Health All of Us Research Program". To the right, there are two links: "Learn more about becoming a participant at [JoinAllofus.org](https://www.joinallofus.org)" and "Learn about the program at [Allofus.nih.gov](https://www.allofus.nih.gov)". Below the logos is a navigation bar with links for "ABOUT", "DATA", "FAQ", and a search icon. The main content area features a large heading "Welcome to the *All of Us* Research Hub." followed by a paragraph: "The *All of Us* Research Program, part of the National Institutes of Health, is building one of the largest biomedical data resources of its kind. The *All of Us* Research Hub will store health data from one million or more diverse participants in the *All of Us* Research Program." Below this is another paragraph: "Here in the Research Hub, everyone can learn more about the type of data participants are providing and how approved researchers can use our data and tools to conduct". The background of the main content area is a photograph of a woman with glasses looking at a computer monitor displaying a colorful 3D model of a brain. A "Help" button with a question mark icon is visible in the bottom right corner of the screenshot.

**All of Us** Research Hub | NIH National Institutes of Health  
*All of Us Research Program*

Learn more about becoming a participant at [JoinAllofus.org](https://www.joinallofus.org)  
Learn about the program at [Allofus.nih.gov](https://www.allofus.nih.gov)

ABOUT DATA FAQ 🔍

## Welcome to the *All of Us* Research Hub.

The *All of Us* Research Program, part of the National Institutes of Health, is building one of the largest biomedical data resources of its kind. The *All of Us* Research Hub will store health data from one million or more diverse participants in the *All of Us* Research Program.

Here in the Research Hub, everyone can learn more about the type of data participants are providing and how approved researchers can use our data and tools to conduct

Help

# Other Data Types

---

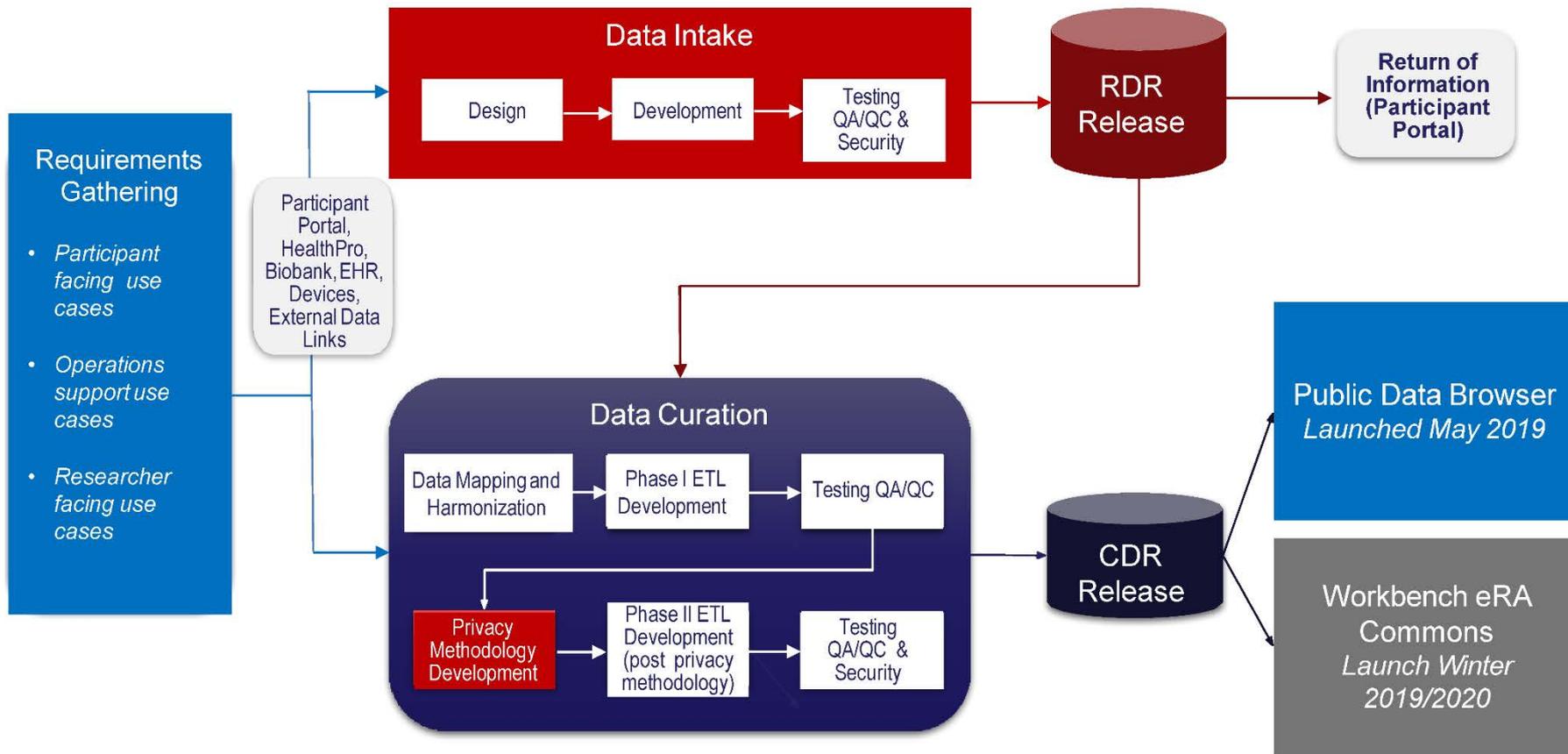
- **Bring Your Own Device (BYOD)**
  - Fitbit (*>2500 users, data has been collected but not available to researchers yet*)
- **Apple Health Kit** (*on hold*)
- **External Data**
  - Mortality data (National Death Index)
  - Cancer registry data (NAACCR)

# Biospecimens

---

- **Genomics – *All of Us* Array (*in progress*)**
  - Captures more than 1.8 million sites
  - Supports association studies using common and rare variations across the genome
  - Includes clinically actionable variation in the genome (ClinVar)
  - Includes variants influencing drug transport, metabolism, elimination
- **Salivary Pilot (*in progress*)**
  - Direct Volunteers
  - Samples processed at Biobank and stored until ready for genotyping
- **Assays (*proposed*)**
  - HbA1c
  - CBC
  - Heavy metals
  - Pesticides + creatinine

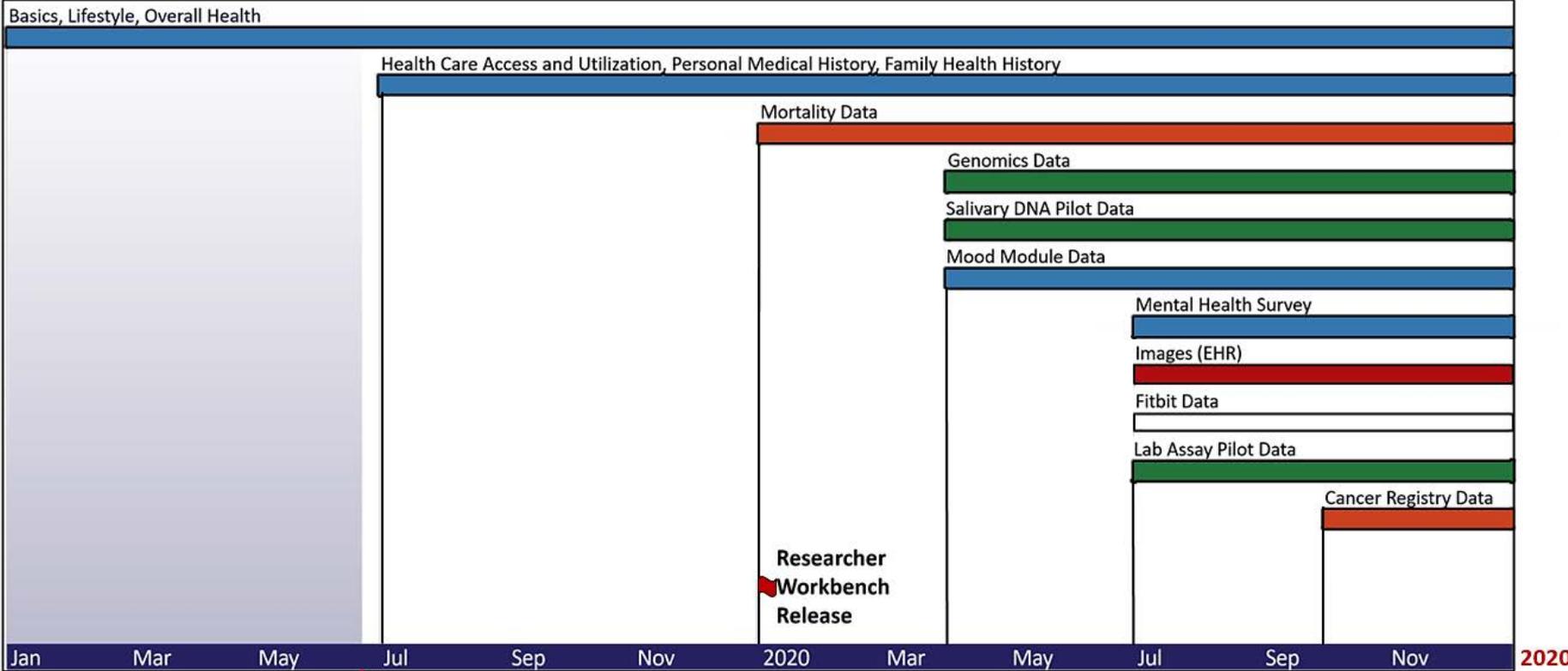
# New Data Types Lifecycle



## **All of Us Research Data (expected to grow by launch or shortly after)**

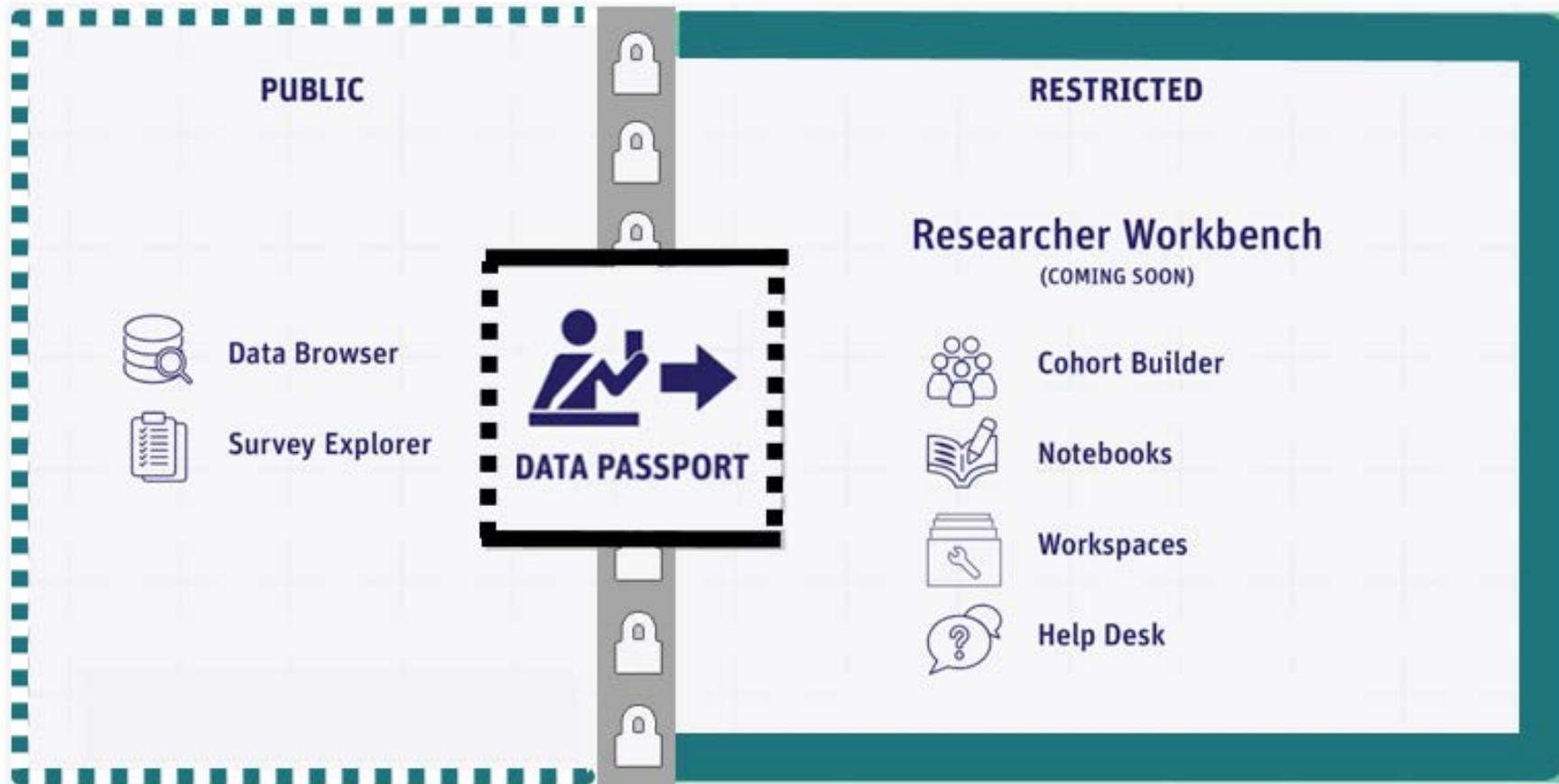
<b>Data Type</b>	<b>Method</b>	<b>Specification</b>	<b>Expected N at launch</b>
Blood pressure, heart rate, height, weight, BMI, hip and waist circumference, wheelchair use and pregnancy status	Physical Measurements	Baseline	>160,000
Sociodemographics, self-assessment	PPI Survey	Baseline	>200,000
Lifestyle, self-assessment	PPI Survey	Baseline	>190,000
Overall health, self-assessment	PPI Survey	Baseline	>190,000
Personal medical history, self-assessment	PPI Survey	Baseline	>30,000
Family medical history, self-assessment	PPI Survey	Baseline	>30,000
Health care access and utilization, self-assessment	PPI Survey	Baseline	>30,000
Electronic Health Record Data (diagnoses, drug exposures, lab measurements, and/or procedures)	EHR	Quarterly	>110,000

# Research Dataset Timeline *(incorporation into the CDR)*



All dates and order of new data types are tentative and subject to change

## Coming This Winter: Powerful Tools for Collaborative, Reproducible Science



# Custom Tools for Exploring and Selecting Subsets of AoU Data

The screenshot displays the 'All of Us Researcher Workbench' interface. At the top, there is a navigation bar with 'ABOUT', 'DATA', 'ANALYSIS', and 'PUBLISH' tabs. The 'DATA' tab is active. The user is logged in as 'Jeremy Chen'. The main content area is titled 'Data' and contains three informational cards: 'Cohorts', 'Concept Sets', and 'Datasets'. Below these is a 'Recent data' section showing four items: 'Diabetes Meds' (Dataset), 'All AoU Participants' (Cohort), 'Diabetes Case 1' (Cohort), and 'Diabetes Case 2' (Cohort). At the bottom, a table lists various data types and their details.

**All of Us**  
RESEARCHER WORKBENCH  
NON-PROFIT/PHILANTHROPIC DATA

Workspaces  
Workspace A

Curated Dataset v. 1

### Data

The Data Tab is the gateway to all Workbench tools and All of Us Research data that will help you complete your research project. Here, users can build a cohorts of participants, select concept sets of interest and build analysis-ready tables from the two called datasets.

#### Cohorts +

A "Cohort" is a group of participants that researchers are interested in. The cohort builder allows you to create and review cohorts and annotate participants in a researcher's study group.

#### Concept Sets +

Concepts describe information in a patient's medical record, such as a condition they have, a prescription they are taking or their physical measurements. Subject areas such as conditions, drugs, measurements etc. are called "domains". Users can search for and save collections of concepts from a particular domain as a "Concept set" and then use concept sets and cohorts to create a dataset, which can be used for analysis.

#### Datasets +

Datasets are analysis-ready tables that can be exported to analysis tools such as Notebooks. Users can build and preview a dataset for one or more cohorts by selecting the desired concept sets and values for the cohorts.

#### Recent data

**Diabetes Meds** (DATASET)  
Research purpose lorem ipsum dolor sit amet consectetur iscing velit...  
Last changed: 12:15 PM

**All AoU Participants** (COHORT)  
Research purpose lorem ipsum dolor sit amet consectetur iscing velit...  
Last changed: Mar 24, 2019

**Diabetes Case 1** (COHORT)  
Research purpose lorem ipsum dolor sit amet consectetur iscing velit...  
Last changed: Mar 22, 2019

**Diabetes Case 2** (COHORT)  
Research purpose lorem ipsum dolor sit amet consectetur iscing velit...  
Last changed: Mar 21, 2019

Data Type	Name	Owner	Last Changed
COHORT	Diabetes Control Group	Keri Wolf	Mar 19, 2019
CONCEPT SET	Demographics	Keri Wolf	Mar 1, 2019
CONCEPT SET	Physical Measurements	Karthik M	Feb 6, 2019
CONCEPT SET	Measurements	Helen Sullivan	Feb 1, 2019
CONCEPT SET	Vitals	Keri Wolf	Dec 24, 2018

# Intuitive, Advanced Tool for Building Cohorts of Participants

DATA ANALYSIS ABOUT PUBLISH Curated Dataset v.1

## Cohorts

Description area here littera gothica quam nunc putamus parum claram anteposuerit litterarum formas humanitatis per.

CREATE COHORT

### Include Only

**Group 1**

- Contains Demographics | 45562

OR

ADD CRITERIA

Temporal **Group Count: 45562**

AND

**Group 2**

ANY MENTION of

- Contains ICD9 code in group CL... | 12540

OR

- Contains ICD9 code in group 143... | 8595

OR

- Contains ICD10 code in group CL... | 11656

ADD CRITERIA

WITHIN X DAYS OF

- Contains Drugs Code | 2150

### Exceptions

**Group 4**

ADD CRITERIA

### Total Count: 8751

#### Results by Gender

Gender	Total Count
Female	~600
Male	~750

#### Results by Gender, Age, and Race

Gender	Age Group	American Indian	Asian	Black or African American	Hispanic	White
Female	15-44	~2%	~1%	~1%	~1%	~95%
Female	45-64	~1%	~1%	~1%	~1%	~95%
Female	> 65	~1%	~1%	~1%	~1%	~95%
Female	0-14	~1%	~1%	~1%	~1%	~95%
Male	> 65	~1%	~1%	~1%	~1%	~95%
Male	15-44	~1%	~1%	~1%	~1%	~95%
Male	< 5-14	~1%	~1%	~1%	~1%	~95%
Male	0-14	~1%	~1%	~1%	~1%	~95%

# Search Interface for Identifying Medical Concepts of Interest

The screenshot displays a web application interface for searching medical concepts. The top navigation bar includes tabs for 'ABOUT', 'DATA', 'ANALYSIS', and 'PUBLISH', along with a 'Curated Dataset v. 1' dropdown and a help icon. The main content area is titled 'Concept Sets' and contains a search bar with 'Metformin' and a filter for 'Standard concepts only'. Below the search bar, there are tabs for different domains: 'Conditions' (10,859), 'Drugs' (4,504), 'Measurements' (2,911), 'Procedures' (3,549), 'Demographics' (1,203), and 'Visits' (7,350). A table shows the top 100 of 4,504 drugs, with 'Metformin' highlighted in green. An 'ADD TO SET' button is visible at the bottom right of the table.

**Concept Sets**  
Nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat ut wisi enim ad.

Search: Metformin  Standard concepts only

Learn more about OMOP domains

Conditions: 10,859 | **Drugs: 4,504** | Measurements: 2,911 | Procedures: 3,549 | Demographics: 1,203 | Visits: 7,350 | Domain Descriptions

Showing top 100 of 4,504 Drugs

Name	Count
<input type="checkbox"/> Drugs used in diabetes	102,456
<input type="checkbox"/> Insulin and analogues	54,785
<input type="checkbox"/> Blood glucose lowering drugs, excl. insulins	21,012
<input type="checkbox"/> Biguanides <b>Metformin</b>	51,045
<input type="checkbox"/> Sulfonylureas	12,874
<input type="checkbox"/> Alpha glucose inhibitors	11,986
<input type="checkbox"/> Thiazolidinediones	6,789

ADD TO SET +

# Point and Click Interface for Creating Analysis-ready Datasets

**ABOUT** DATA ANALYSIS PUBLISH Curated Dataset v.1

## 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.

### Select Cohorts

**Cohorts** +

- All Aou Participants
- Diabetes Case 1
- Diabetes Case 2
- Diabetes Control Group

### Select Concept Sets

**Concept Sets** +

- Demographics
- Physical Measurements
- Measurements
- Vitals
- Drugs

### Values

Select All

- Race
- Ethnicity
- Drugs**
- DRUG\_CONCEPT\_ID
- DRUG\_CONCEPT\_NAME
- DRUG\_CONCEPT\_CODE
- DRUG\_VOCABULARY\_ID
- DRUG\_EXPOSURE\_START\_DATE

### Preview Dataset

A visualization of your data table based on the variable and value you selected above. **SAVE DATASET**

**Demographics** **Drugs**

Participant ID	DRUG_CONCEPT_ID	DRUG_CONCEPT_NAME	DRUG_CONCEPT_CODE	DRUG_VOCABULARY_ID
12345	201235	Acetohexamide	5645415	Vocab value
12346	88451	Chlorpropamide	545856	Vocab value
12347	117456	gliclazide	650210	Vocab value

# Jupyter Notebooks for Powerful, Interactive Data Analysis

All of Us RESEARCH PROGRAM RESEARCHER WORKBENCH

Workspaces > Phenotype Notebooks > Notebooks > Major Depressive Disorder

Joseph DiPaolo

jupyter Major Depressive Disorder Last Checkpoint Last Thursday at 4:36 PM (unsaved changes)

File Edit View Insert Cell Kernel Navigate Widgets Help Snippets

Contents

- Major Depressive Phenotype
  - 1. Find Concept IDs
    - Condition Concept IDs
    - Drug Concept IDs
      - Simple drug pull
      - Complex drug pull
    - Procedure Concept IDs
  - 2. Select participants with condition
  - 3. Select participants with drug exposure
    - Simple drug pull
    - Complex drug pull
  - 4. Select participants with procedure
  - 5. Select participants with PPI data
  - 6. Final results
  - 7. Cohort Breakdown
    - Gender Breakdown
    - Race Breakdown
    - Hispanic Breakdown
    - Age Breakdown

```
WHEN DATE_DIFF(CURRENT_DATE(),DATE(CDR_DF.birth_datetime), DAY)/365 < 86 THEN '76-85'
ELSE '86+' END) AS Age_Bracket,
CAST(SUM(Simple_DF.Simple) AS NUMERIC) AS Simple,
CAST(SUM(Complex_DF.Complex) AS NUMERIC) AS Complex,
CAST(SUM(CDR_DF.CDR) AS NUMERIC) AS CDR

FROM
  CDR_DF
LEFT JOIN Simple_DF ON CDR_DF.person_id = Simple_DF.person_id
LEFT JOIN Complex_DF ON CDR_DF.person_id = Complex_DF.person_id

GROUP BY Age_Bracket

ORDER BY Age_Bracket ASC

***.format(simple_cohort_PIDs_subquery, complex_cohort_PIDs_subquery),
dialect = "standard")
```

In [124]:

```
display(age_table)
print("\n\n")
sns.distplot(age['Age'])
```

Age_Bracket	Simple	Complex	CDR	
0	0-17	None	None	2
1	18-25	11	11	12182
2	26-35	27	27	24268
3	36-45	35	35	22301
4	46-55	49	48	26219
5	56-65	75	74	32051
6	66-75	55	54	24364
7	76-85	18	18	8788
8	86+	None	None	904

Out[124]: <matplotlib.axes.\_subplots.AxesSubplot at 0x7f22fd3cc50>

## Tools and Services for Research Support

---

- **Help Desk**
  - General & technical support
  - Community forums
  - Knowledge base
- **Training Materials**
  - Tool user guides
  - Instructional videos
  - Data model tutorials
  - Workspaces w/ example analyses
  - Reusable “code snippets”
- **Data Dictionary**
  - PDF as part of detailed data documentation
  - Integrated directly within Workbench tools



**Thank you**

---