An Introduction to the All of Us Researcher Program

MAXIMILIAN WEGENER, MPH

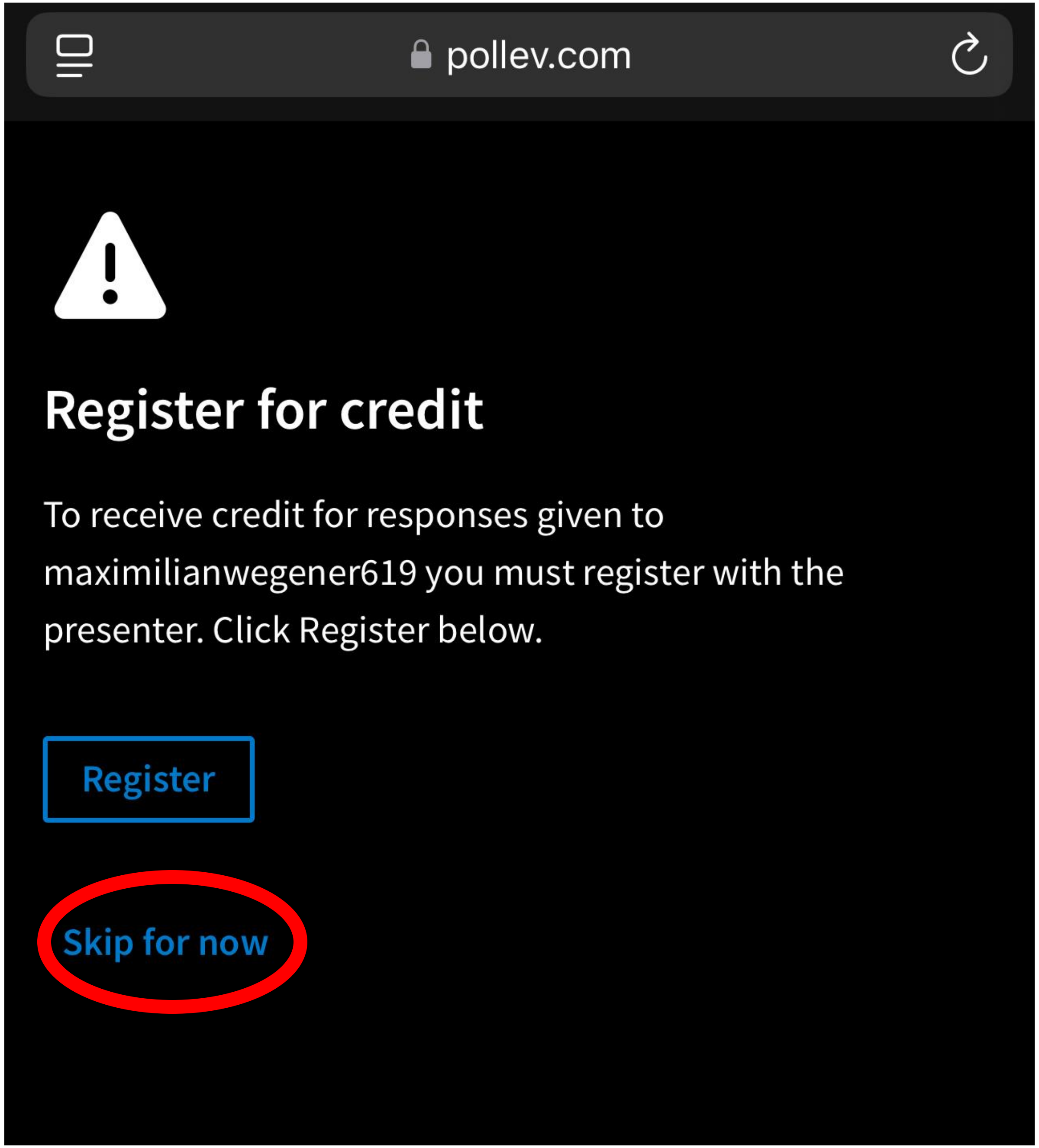
Biomedical Informatics Librarian, Cushing/Whitney Medical Library

[Click here to find a recording of the class]

Yale SCHOOL OF MEDICINE
Biomedical Informatics and Data Science

Yale *Harvey Cushing/John Hay Whitney Medical Library*

Interactive poll questions



Register for credit

To receive credit for responses given to maximilianwegener619 you must register with the presenter. Click Register below.

Agenda

- Learning objectives
- About All of US
- Data sources
- Accessing data
- Analyzing data
- Create an All of Us account and Workspace

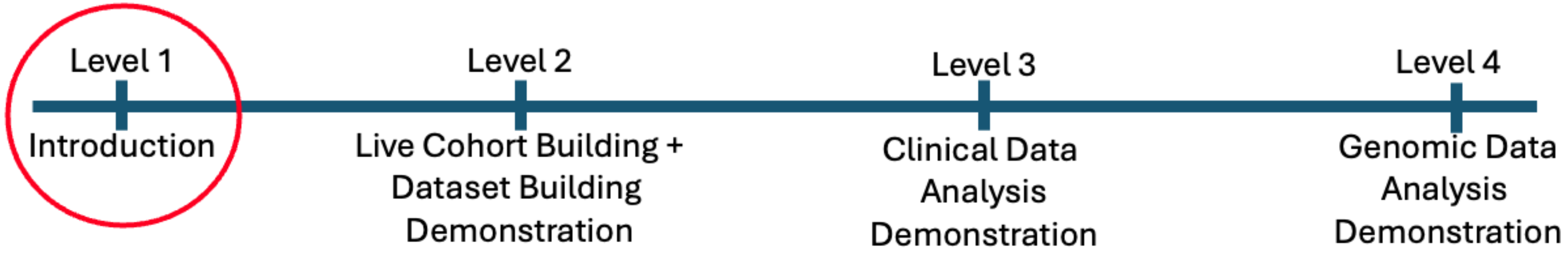
Learning Objectives

After this training, you should be able to:

1. Explain the All of Us program
2. Identify key data sources
3. Understand how data are stored
4. Create datasets using the cohort builder + dataset builder
5. Explain some of the analytic methods available for All of Us data
6. Create a user account and workspace

Course Progression

Current Level and Future Directions



Level 1

Introduction

Level 2

Live Cohort Building + Dataset Building Demonstration

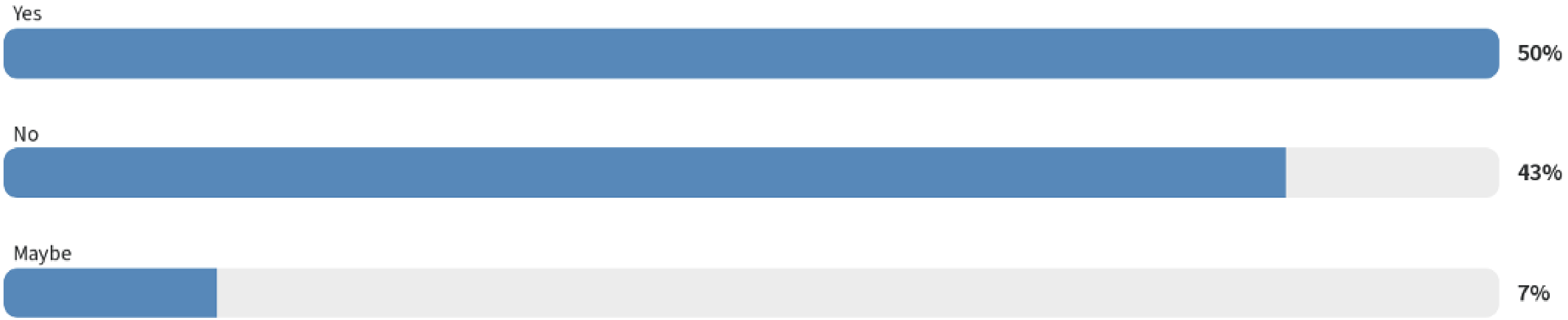
Level 3

Clinical Data Analysis Demonstration

Level 4

Genomic Data Analysis Demonstration

Have you ever heard of All of Us before signing up for this class?



Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

About the All of Us Research Program

Background

- A national initiative by the NIH
- Launched in 2018
- To gather health data from over 1 million people
- Eligible participants join voluntarily
  - signing up via JoinAllofUs.org
  - participating health care provider

About the All of Us Research Program

Vision and Goals

- Enable precision medicine using data on lifestyle, environment, and genetics.
- Shift from one-size-fits-all to personalized healthcare.
- Boost health research and medical breakthroughs.
- Tailor and make healthcare equitable.
- Support diverse studies reflecting all populations.

About the All of Us Research Program

What makes it unique

- Participant-Centered
  - Participants view and access their data
- Open science
  - De-identified data available globally
- Research opportunities
  - Health disparities; environmental impacts; disease prevention and treatment
  - Phenotypic and genotypic data
- Scalability & Innovation
  - Complex analyses; AI and machine learning

Data Sources

Surveys

EHR

Wearables

Physical measurements

Genomic

Lifestyles, medical history, healthcare access, etc.

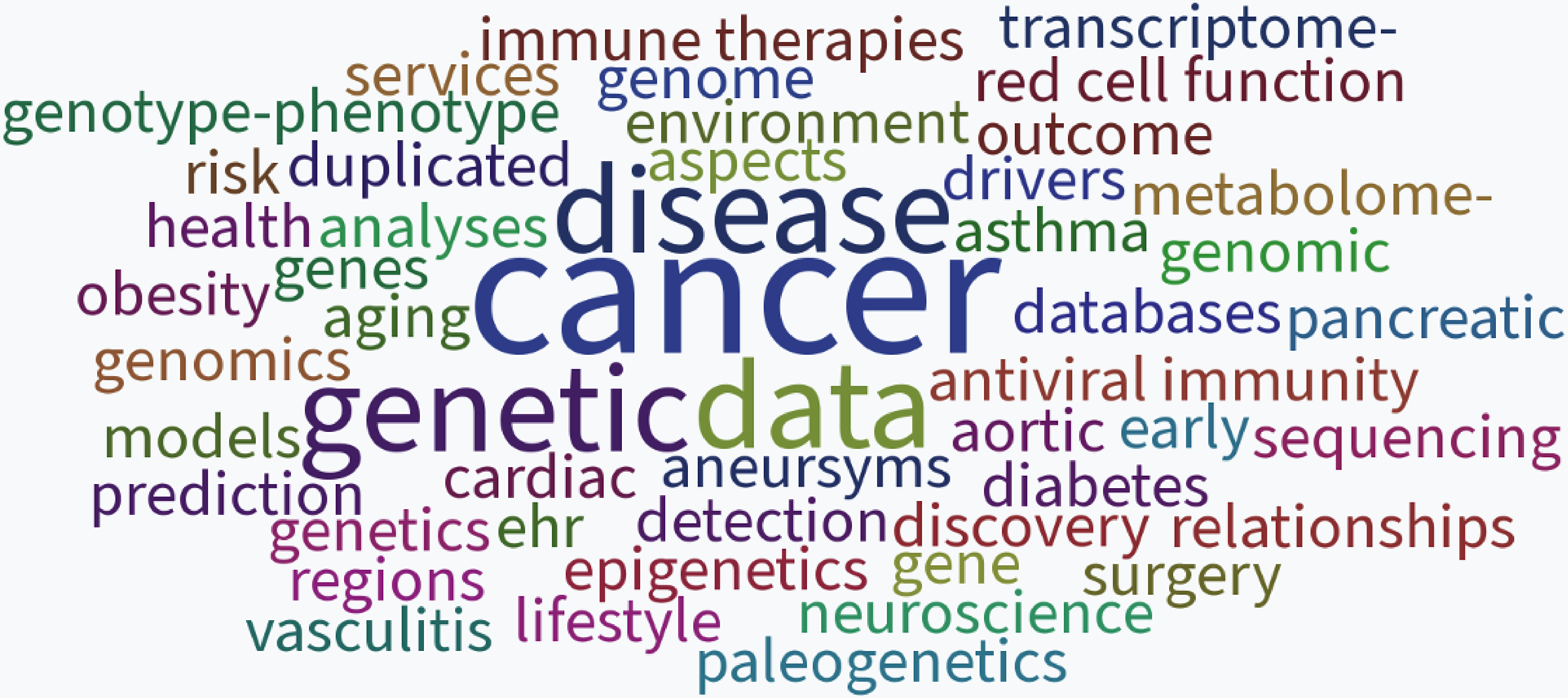
Standardized using Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)

Heart rate, activity, or sleep data from devices like Fitbit

Sources: EHRs, self-reported height and weight, in-person visits.

Whole genome sequencing and genotyping from blood, saliva, & urine

What are your current research interests? [list as many as you want]



Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Accessing Data

Background

- Organized in a curated data repository (CDR)
  - De-identified
  - Cleaned, validated, standardized

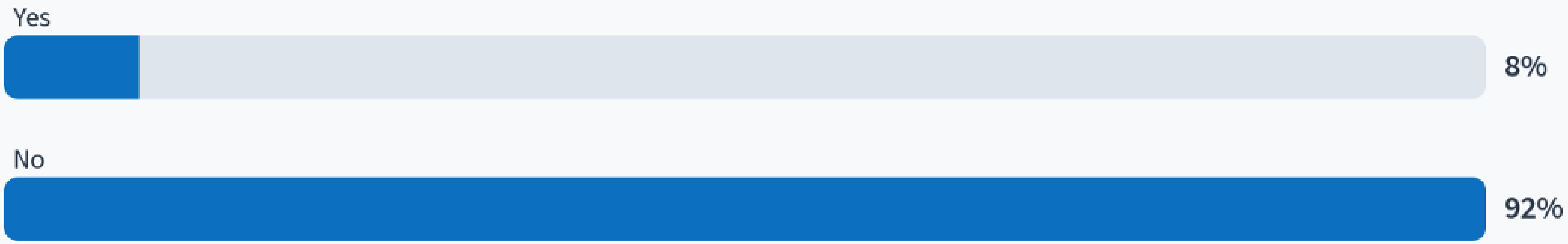
- Hosted on Google Cloud Platform (GCP)
  - scalability, security, and computational power
  - Data accessed from anywhere
  - Large-scale data processing and advanced analytics

Accessing Data

Background

- Google BigQuery
  - Database for the CDR
  - Uses SQL-like queries to retrieve data
  - Jupyter Notebooks: R or Python
  - Billed for queries executed and data processed
  - large genomic datasets can be costly
- $300 credit for Google Cloud usage
- The credit will eventually run out
- Set up a billing account to continue your work
  - Navigate here for more information
  - helpful video here covering billing in more detail

Do you already have an All of Us account?



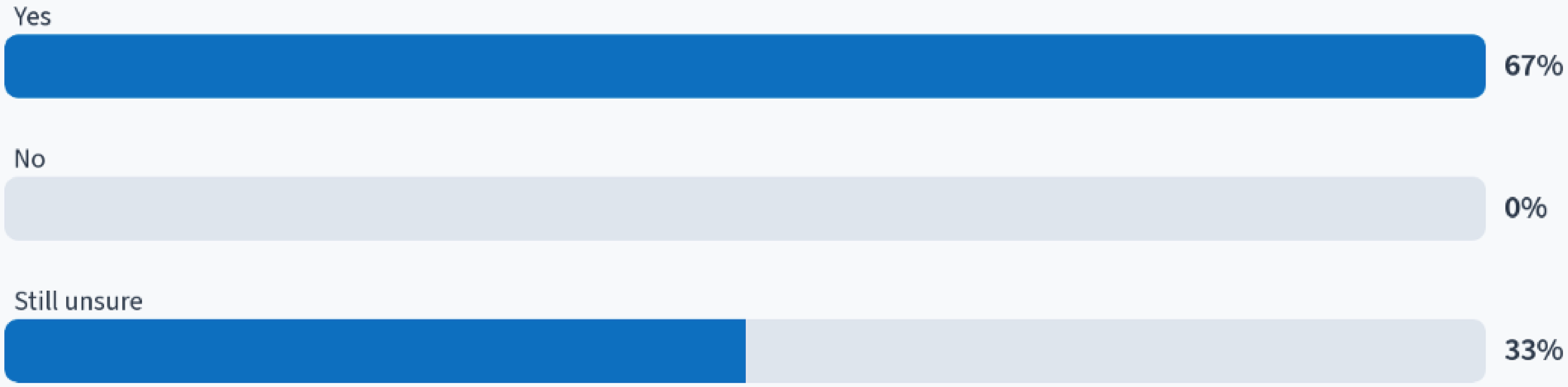
Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Accessing Data

Create a workspace

- Create a Research Hub account
  - Navigate to All of Us Research Hub and choose create an account and follow the prompts
- Complete data access registration
  - ID verification; mandatory trainings; code of conduct
- Create an all of us workspace
  - Research Use Statement Questions
  - All information is publicly available
  - More than one workspace
  - Shareable
  - Where you interact with data

Are you interested in making an All of Us account for your research?



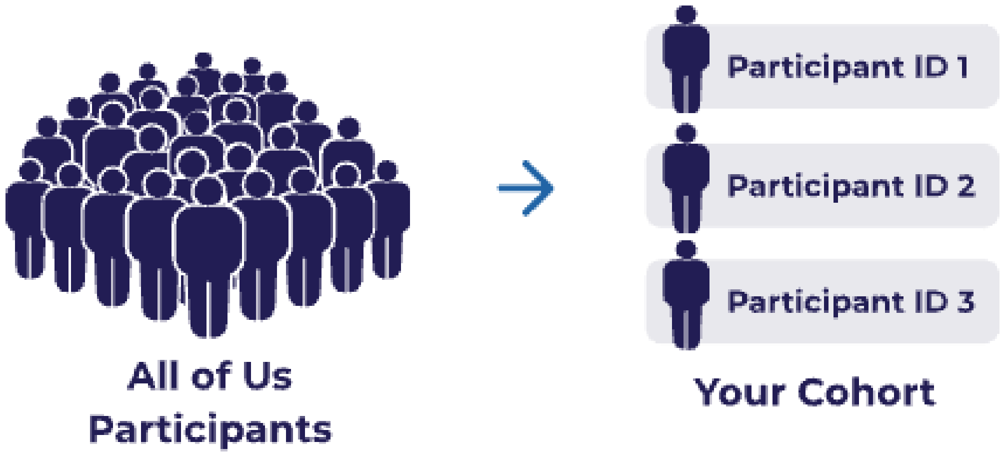
Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Accessing Data

Two options

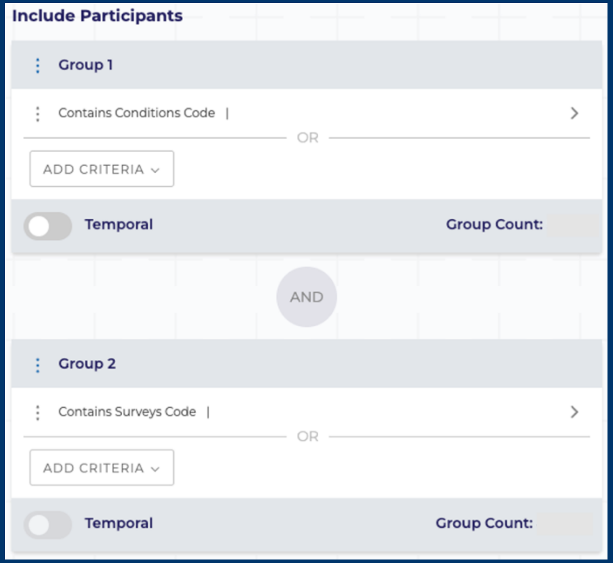
1. Build your own SQL query within Jupyter Notebooks

2. **Cohort Builder + Dataset Builder**
   - Click and choose
   - Better for non-SQL users
   - Less processing
   - Better for limiting data to what you need



Accessing Data

Cohort builder



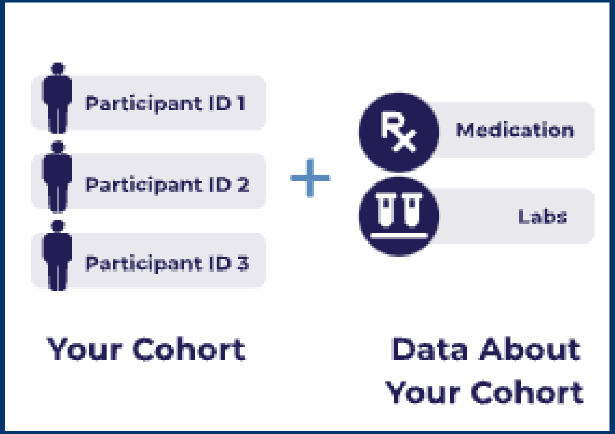
- Subset your study population from all participants based on specified criteria using AND/OR operators
  - Demographics, conditions, measurements, medications, procedures, and surveys
- Easily select inclusion and exclusion criteria
- Add temporal events
  - Illness occurs N days before medication was prescribed

Example criteria:

- Race -> **White**
- Sex **-> Woman**
- Age **-> 65 or older**
- Condition - Diagnosed with **breast cancer**
- Medication -Taking **tamoxifen**

Accessing Data

Concepts

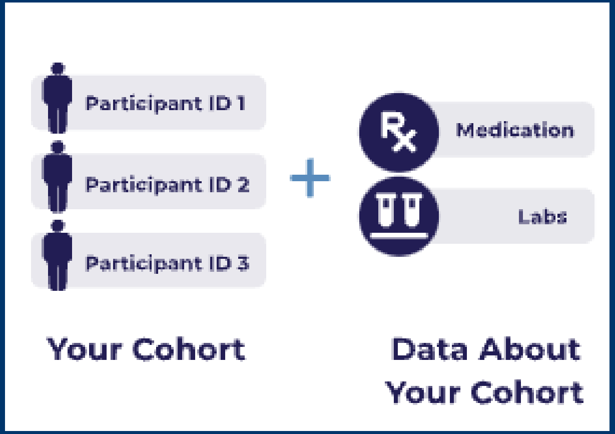


- Variables to include in your analysis
- Describe information from medical record
  - Demographics, conditions, prescription, physical measurements, etc.
- Organized into *concept domains* which are the subject areas for each concept

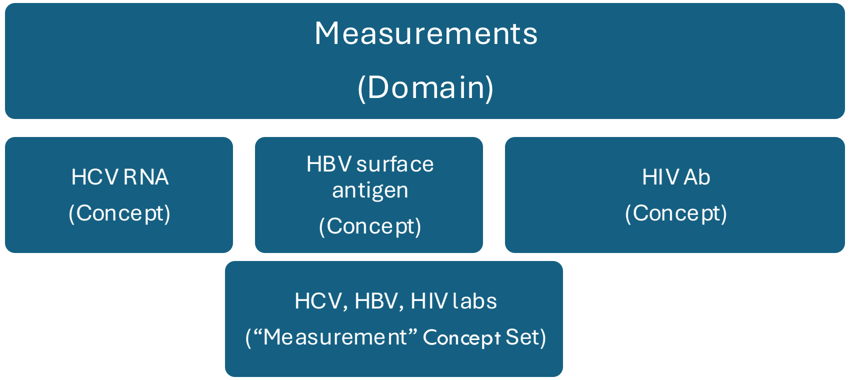
Domain	**Description**
Conditions	Listed by ICD9 or ICD10 or SNOMED standard codes.
Procedures	Listed by ICD9, ICD10, CPT, or SNOMED standard codes.
Drugs or Medications	Listed by ingredient and organized by therapeutic uses.
Measurements	**Laboratory tests and vital signs, organized in the LOINC code hierarchy.**
Visits	Type of facility where medical care was received (ED, OP, IP).
Surveys	Questions and associated response options for participant-completed surveys.
Physical Measurements	At time of participant enrollment: BP, HR, height, weight, BMI, pregnancy, etc.
Demographics	Age, gender, race, ethnicity, and deceased status.

Accessing Data

Concepts



- *Concept sets* are **one or more** concepts from a particular domain used to create the dataset for your analysis



Measurements (Domain)

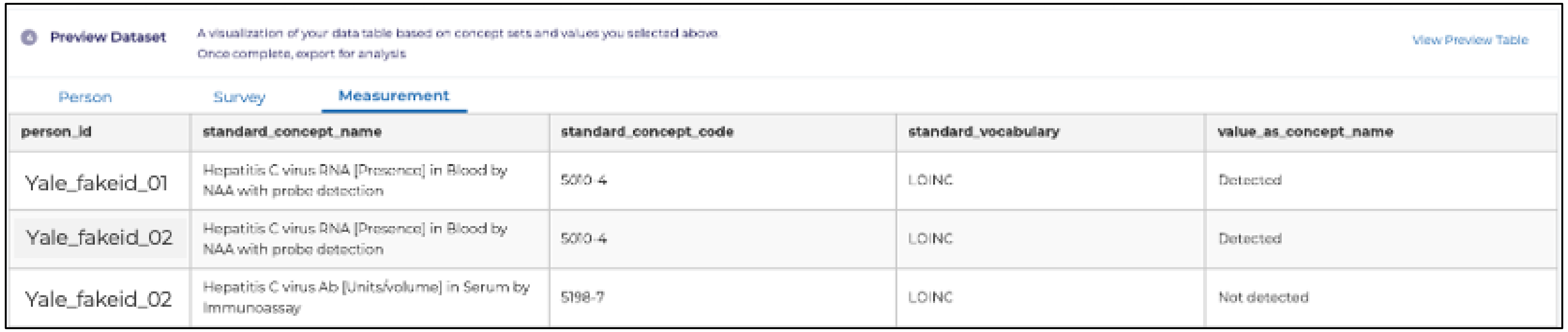
HCV RNA (Concept)

HBV surface antigen (Concept)

HIV Ab (Concept)

HCV, HBV, HIV labs (“Measurement” Concept Set)

- The values from each concept set becomes its own table under the domain name in the analysis phase



Preview Dataset

A visualization of your data table based on concept sets and values you selected above

Once complete, export for analysis

View Preview Table

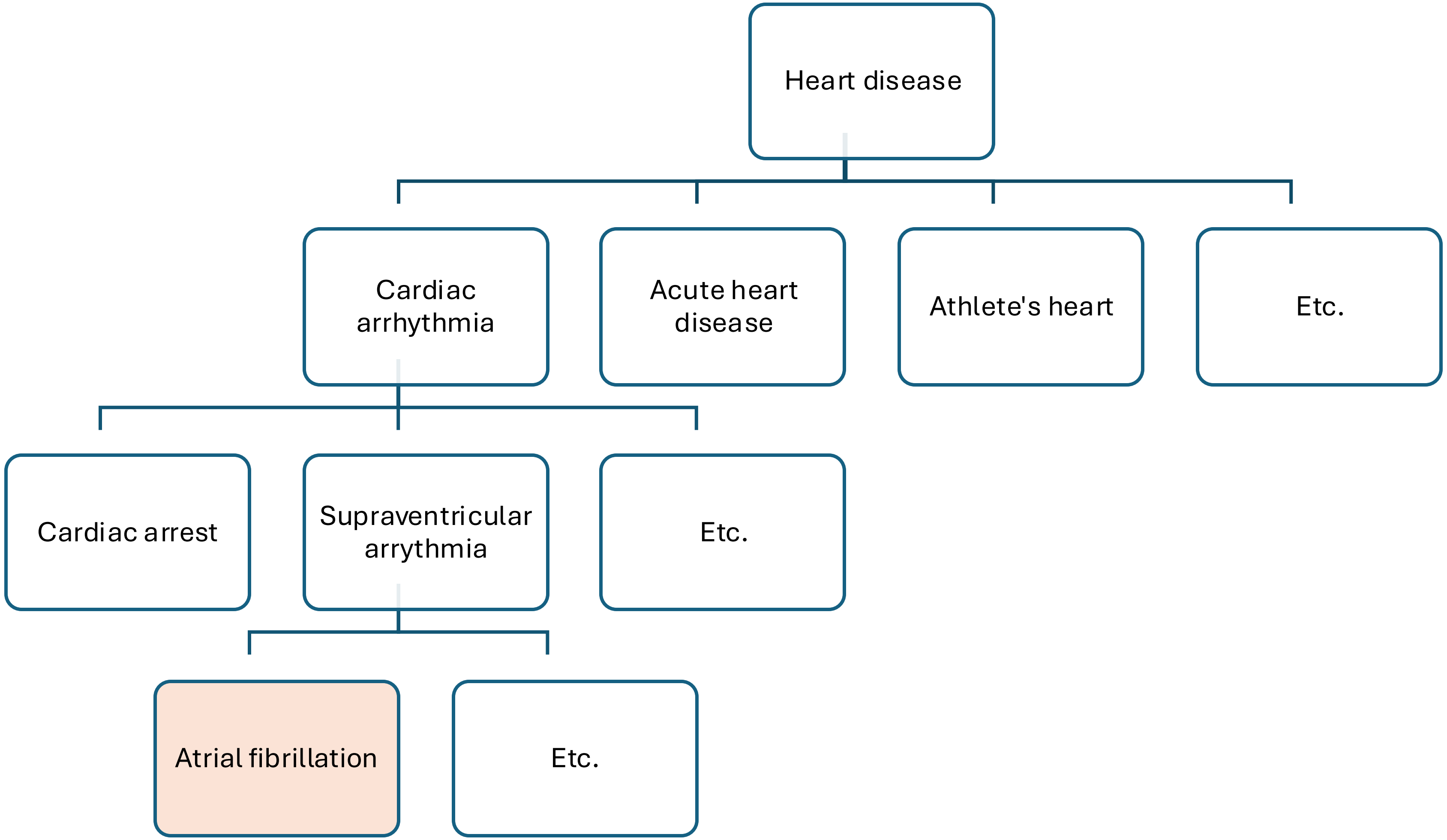
Person | Survey | **Measurement**

person_id	standard_concept_name	standard_concept_code	standard_vocabulary	value_as_concept_name
Yale_fakeid_01	Hepatitis C virus RNA [Presence] in Blood by NAA with probe detection	5010-4	LOINC	Detected
Yale_fakeid_02	Hepatitis C virus RNA [Presence] in Blood by NAA with probe detection	5010-4	LOINC	Detected
Yale_fakeid_02	Hepatitis C virus Ab [Units/volume] in Serum by Immunoassay	5198-7	LOINC	Not detected

- Remember to include the same concepts from your cohort builder

- After choosing your concept sets, select which values (columns) you’d like to keep

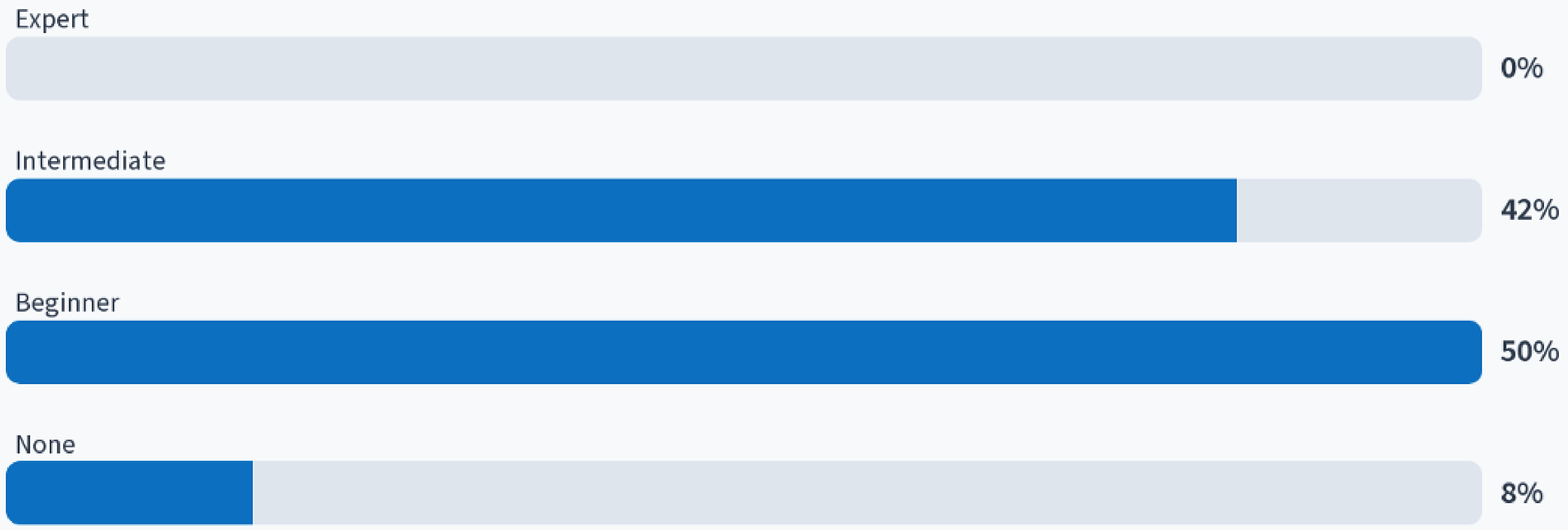
Concept Hierarchy for Condition "Atrial fibrillation"



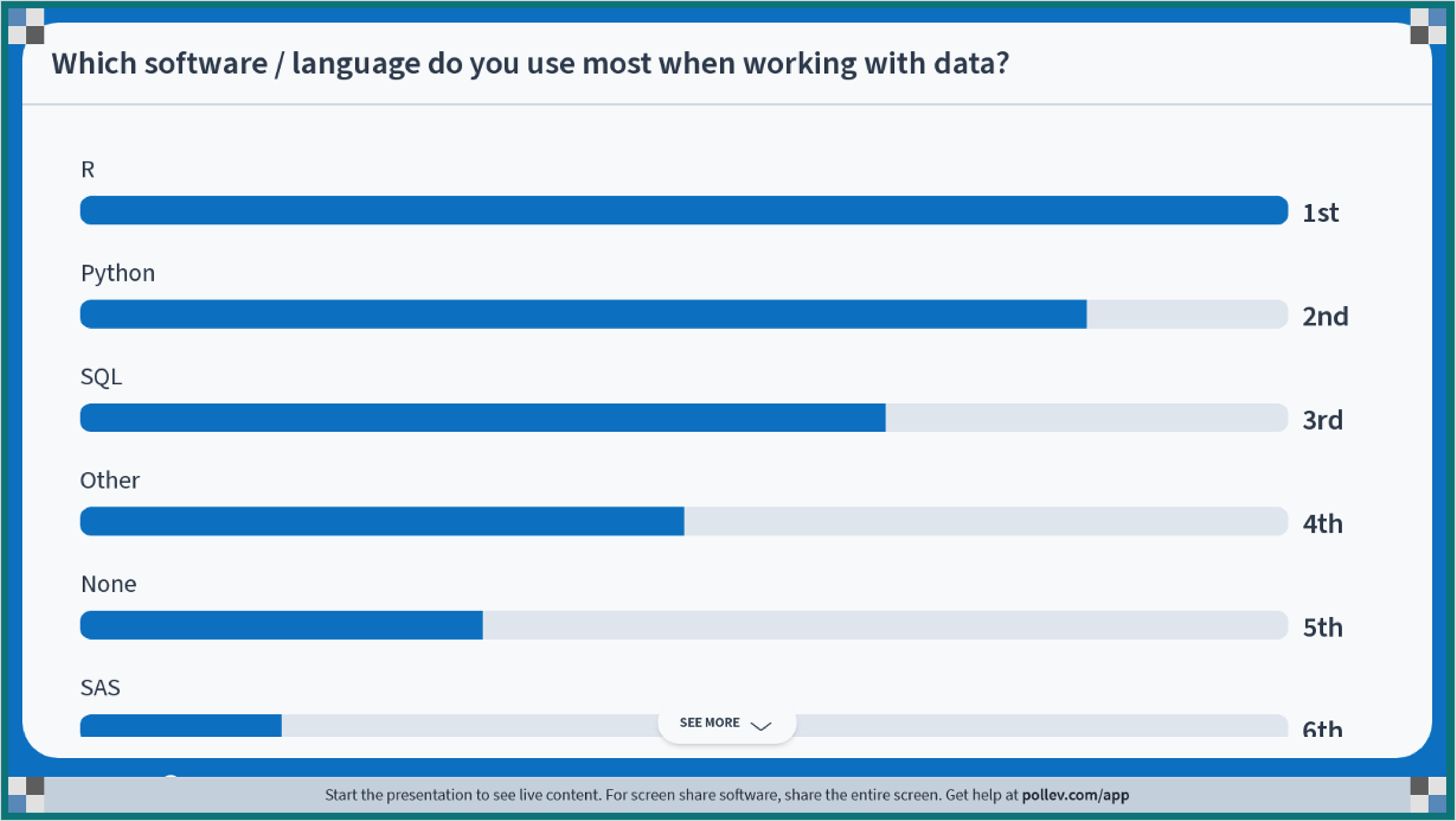
General:
a larger pool of patients

Specific:
smaller pool of patients

What is your data analysis experience level?



Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**



Which software / language do you use most when working with data?

R

1st

Python

2nd

SQL

3rd

Other

4th

None

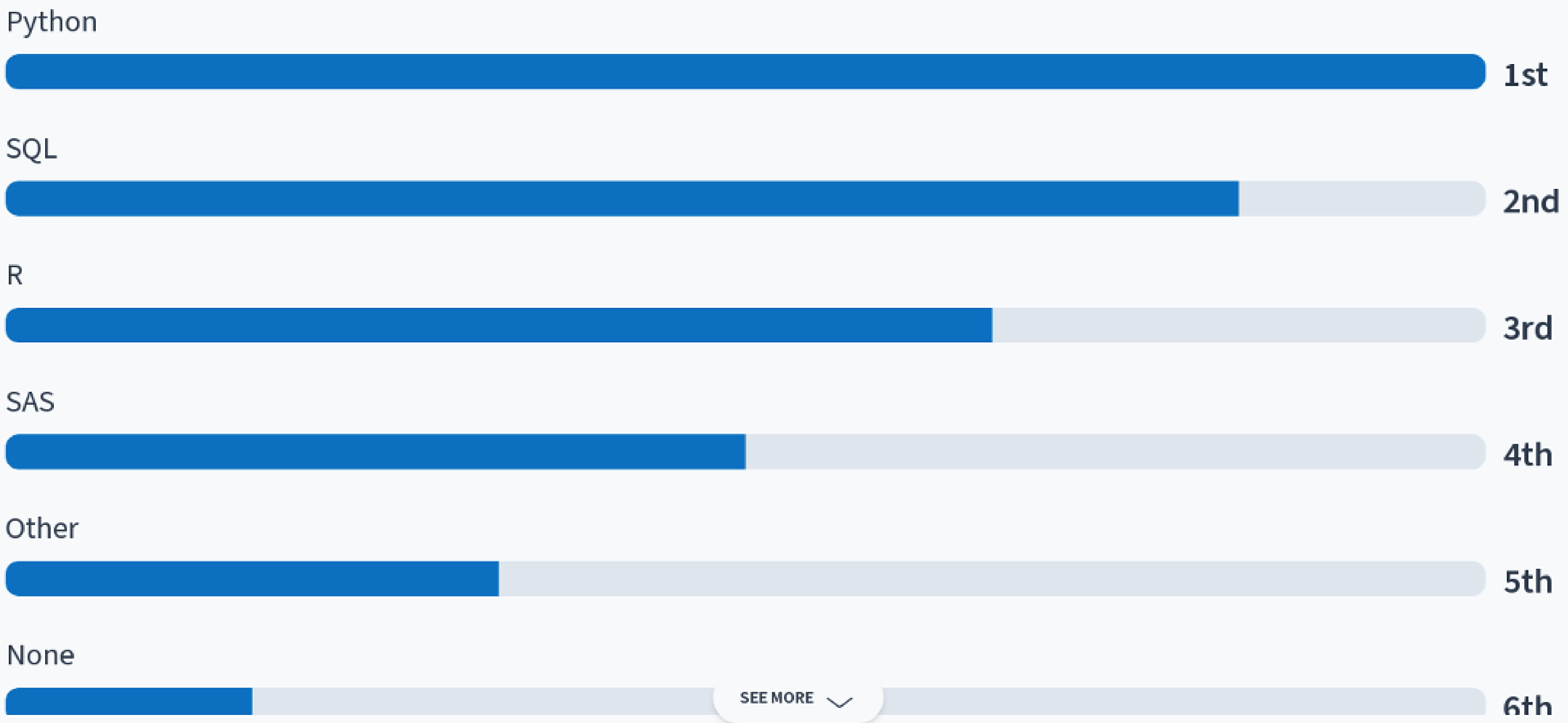
5th

SAS

6th

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Which software / language do you want to learn more about?

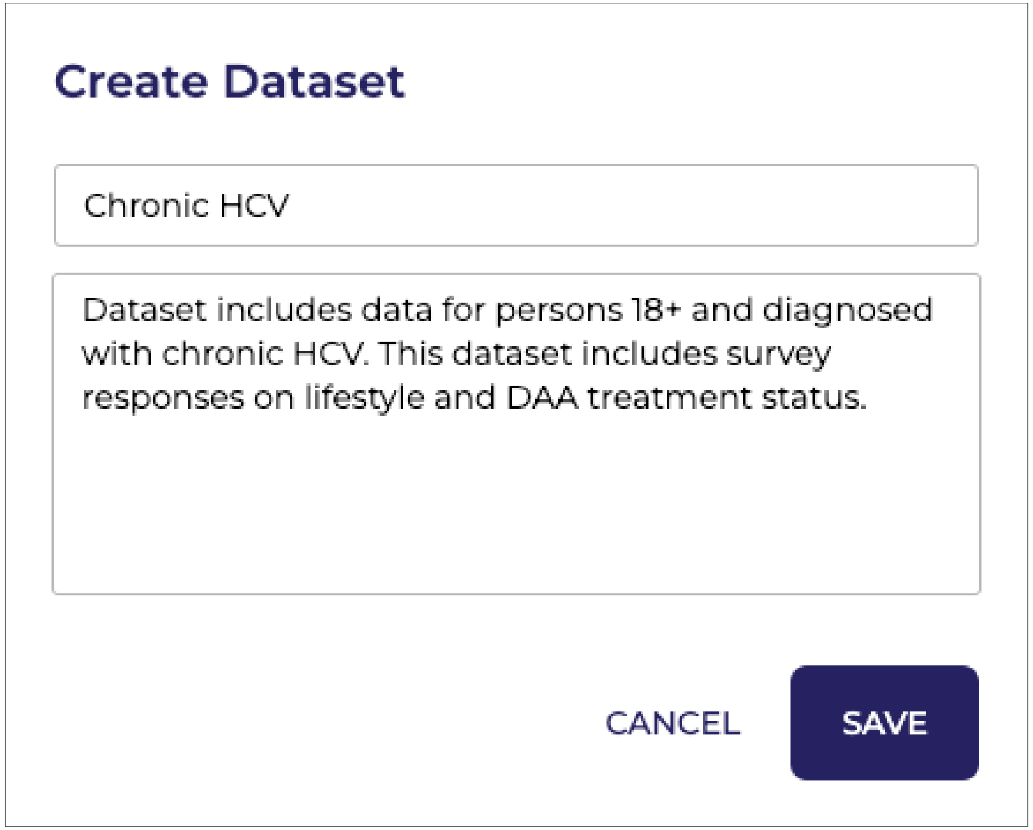


Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Accessing Data

Dataset

- Once you select your cohort, concept sets, and concept values, you can create your dataset



Create Dataset

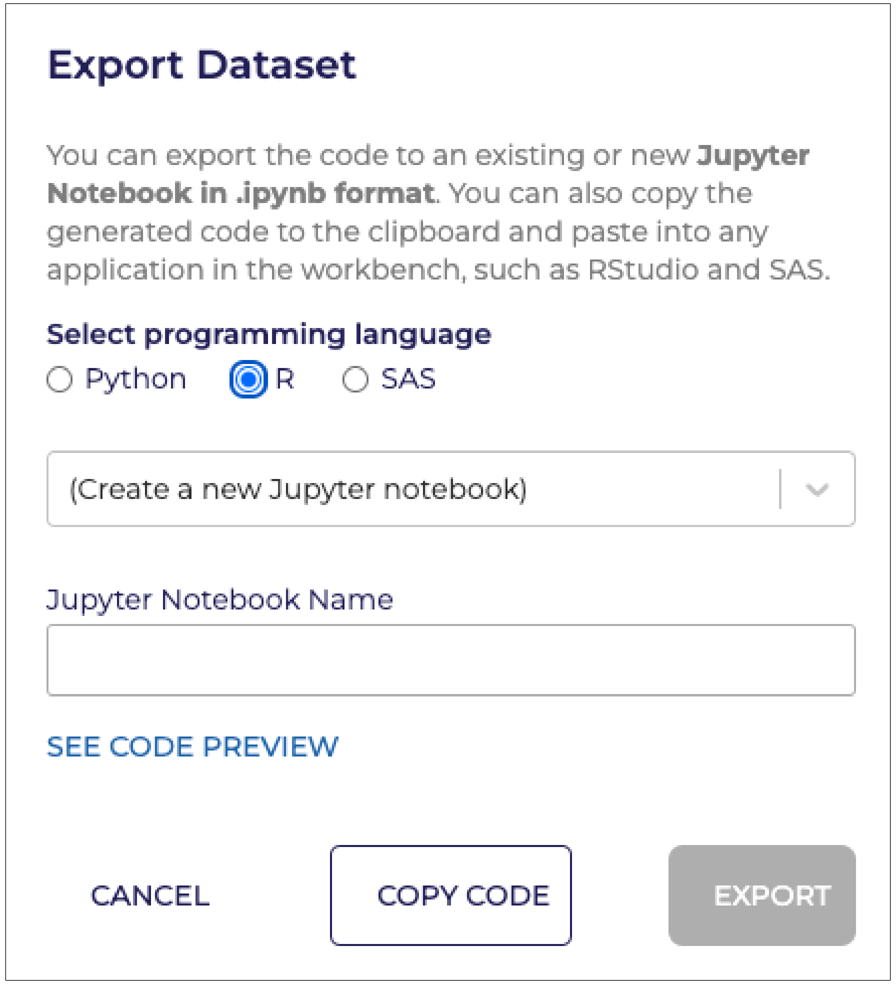
Chronic HCV

Dataset includes data for persons 18+ and diagnosed with chronic HCV. This dataset includes survey responses on lifestyle and DAA treatment status.

Accessing Data

Analysis

- Then choose *Analyze*, and select the your preferred programming language



Export Dataset

You can export the code to an existing or new **Jupyter Notebook in .ipynb format**. You can also copy the generated code to the clipboard and paste into any application in the workbench, such as RStudio and SAS.

Select programming language

○ Python ● R ○ SAS

(Create a new Jupyter notebook)

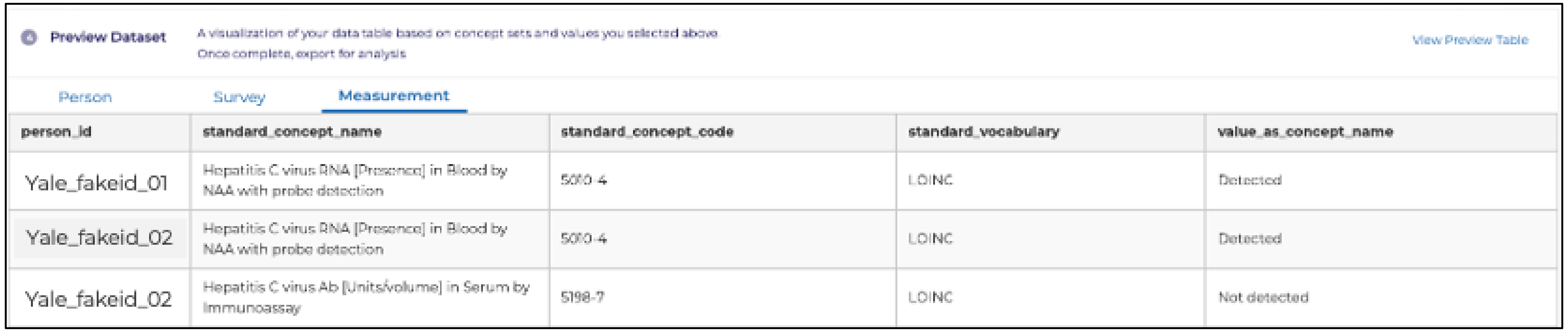
Jupyter Notebook Name

Accessing Data

Analysis

- Once you choose *Analyze*, a Jupyter notebook will be generated

- The SQL queries will already be written
  - One query per concept set domain



Preview Dataset A visualization of your data table based on concept sets and values you selected above. Once complete, export for analysis

person_id	standard_concept_name	standard_concept_code	standard_vocabulary	value_as_concept_name
Yale_fakeid_01	Hepatitis C virus RNA [Presence] in Blood by NAA with probe detection	5010-4	LOINC	Detected
Yale_fakeid_02	Hepatitis C virus RNA [Presence] in Blood by NAA with probe detection	5010-4	LOINC	Detected
Yale_fakeid_02	Hepatitis C virus Ab [Units/volume] in Serum by Immunoassay	5198-7	LOINC	Not detected

- Datasets from queries will be saved in *Workspace Buckets* on Google cloud

- You can update/manipulate code to meet your needs
  - i.e. save datasets so you don't have to rerun the queries every time
  - Handle duplicates

- There are also helpful tools included
  - Code snippets

Helpful Resources

- All of Us Publications

- All of Us YouTube videos
  - Billing in the researcher workbench
  - Cohort builder and dataset builder
  - Using the concept set selector in the workbench

- Setting up your Billing Account

- Getting started (dictionaries, data types, OMOP, etc.)

Future Directions

- Future CWML/BIDS courses aimed to help you work more effectively with All of Us including:
  - R
  - Python
  - SQL
  - Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)
  - Open to suggestions

- Let us know if you are interested in presenting your All of Us work

- Check the library training calendar for current course

Questions