All of Us RESEARCH PROGRAM | Researcher Academy

Researchers' Guide to Using the Workbench

Self-directed Learning: RTI-TTT-3

Last updated March 16, 2026

Contents

Tables

Figures

1. Introduction

The National Institutes of Health's ***All of Us* Research Program** is a historic effort to collect and study data from 1 million or more people living in the United States.

The goal of *All of Us* is to speed up health research discoveries, enabling new kinds of individualized health care.

The *All of Us* Research Program is a nationwide, participant-centered initiative to build one of the largest and most diverse health databases in U.S. history by enrolling 1 million or more participants who share information such as surveys, electronic health records (EHRs), physical measurements, biospecimens, genomic data, and digital health data. Designed to advance precision medicine, the program emphasizes inclusion and participant partnership—offering clear choices about what to share, options to receive certain validated health and genomic results, and strong privacy and security protections. De-identified data are made available to qualified researchers through controlled access tools to enable studies of how genetics, environment, and lifestyle influence health and to help improve prevention, diagnosis, and treatment across all communities.

The *All of Us* Researcher Workbench is a secure, cloud-based analysis platform that gives controlled access to the program's de-identified participant data to credentialed and approved researchers so they can build cohorts, run reproducible analyses, and share results. Designed to support precision medicine research, the Workbench provides tools such as Cohort and Dataset Builders, preloaded curated data sets, integrated analytic environments (e.g., Jupyter Notebook and RStudio), and scalable cloud compute within isolated workspaces, while enforcing data governance through tiered access (Registered and Controlled Tiers), required researcher training and data use agreements, and strict privacy and security protections to protect participant confidentiality.

The purpose of this training manual is to equip you with practical knowledge, step-by-step procedures, and ready-to-use resources so that you can train others to use the Workbench effectively and responsibly. It consolidates technical guidance on Workbench features (Cohort Builder, Dataset Builder, workspaces, analysis tools, billing), data orientation (Data Browser, Survey Explorer, Curated Data Repository [CDR]v8, Observational Medical Outcomes Partnership [OMOP]), and policy/compliance essentials (Data Use and Registration Agreement [DURA], Data User Code of Conduct [DUCC], tiered access).

A Note about Researcher Workbench 2.0

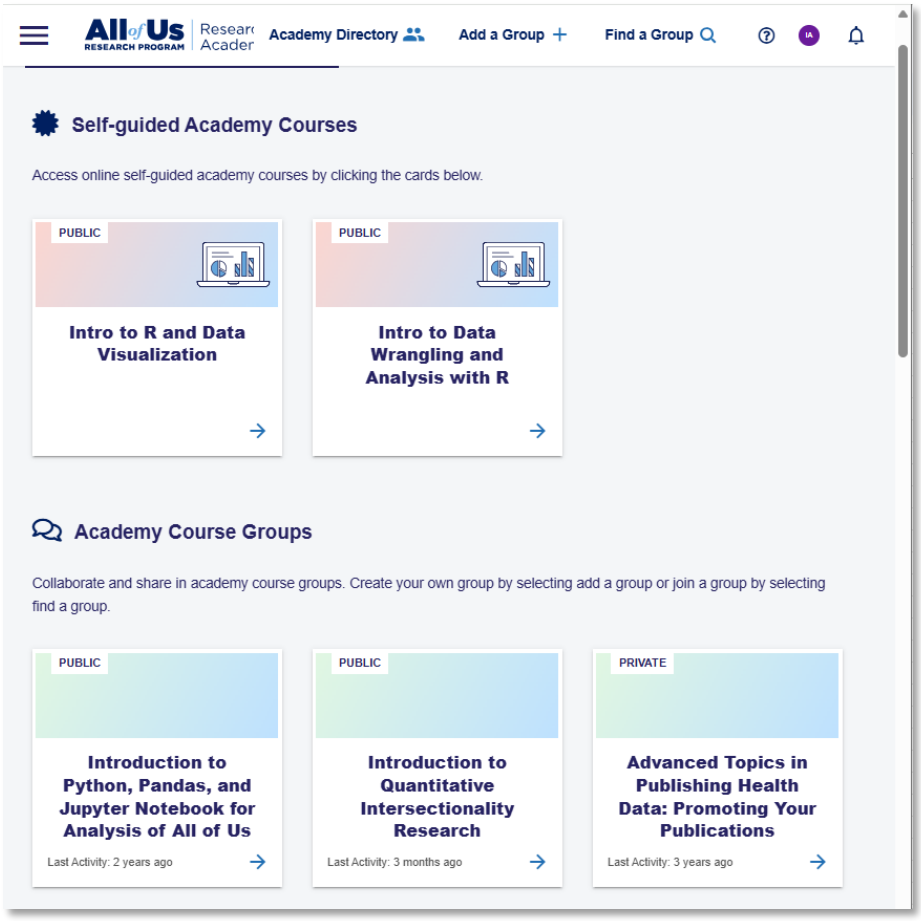
This guide was created for Researcher Workbench 1.0. While Workbench 2.0 will look different and have added features, the underlying logic and most of the information in these training materials will remain applicable.

2. Registering for Amaze and Accessing Courses on Amaze

2.1 Overview of Amaze

Amaze is a cloud-based collaboration and learning platform designed and developed by RTI International. The platform contains courses (**Figure 2-1**) that anyone with an account can take asynchronously to improve their skills on the Workbench as well as a directory for networking to facilitate connections.

Figure 2-1. Academy Course Groups



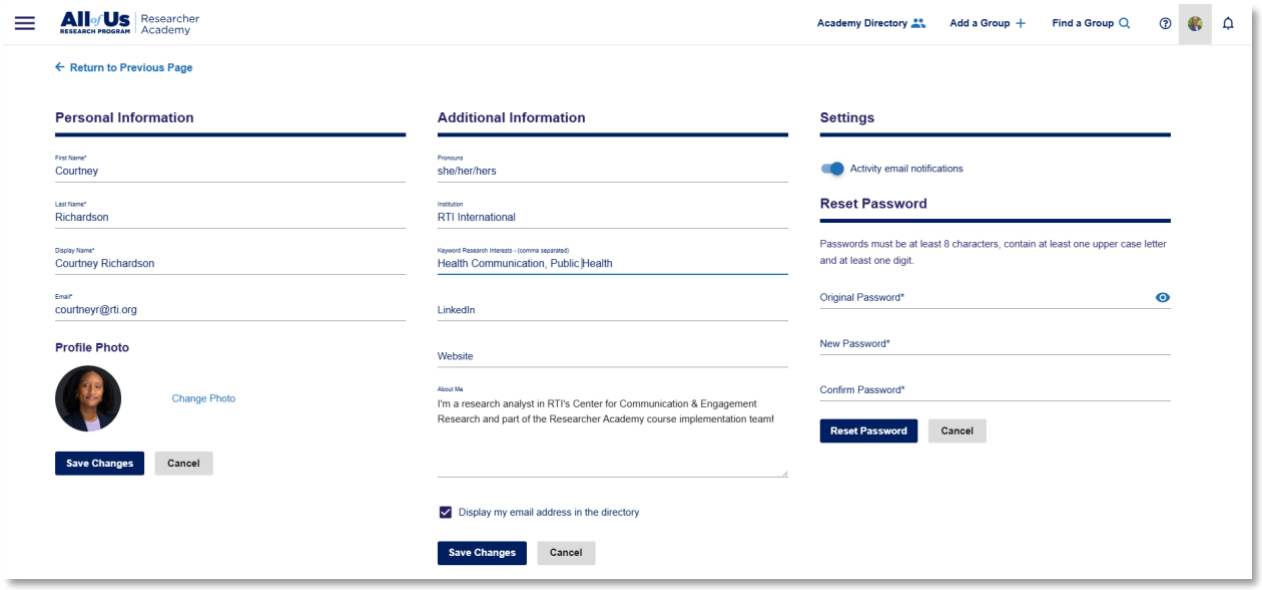
2.2 Creating an Amaze Account and Bio

To create an account on Amaze, please complete the Researcher Academy Amaze Account Request Form.

1. Once you have created an account, log in to your account at https://researcheracademy.rti.org/.
2. Click your profile icon at the top right of the webpage.

3. Select **Edit**.
4. At the bottom of your profile, enter your bio, key research interests, and any other information you wish to share (**Figure 2-2**). This will be visible to other Amaze users.
5. Select **Save Changes** when you are finished.

Figure 2-2. Creating a Bio in Amaze



2.3 *All of Us* Researcher Academy Course Catalog

The *All of Us* Researcher Academy course materials are freely available online for users through Amaze. The *All of Us* Researcher Academy Course Catalog organizes course descriptions, instructor bios, and course materials for users to review and access. In addition to courses on the Workbench and programming, we offer courses on grant writing, publishing your research, statistical techniques, literature searching, and other topics. Other resources include data use cases and presentations on topics like time management and developing an elevator pitch for conference poster sessions.

We strongly encourage new users to take the courses listed below to get started on the Workbench:

- **Introduction to the Researcher Workbench**
- **Researcher Workbench Notebooks 101**

Access the course catalog for more information about each course.

In addition, if you are not already proficient in either R or Python, we strongly encourage you to take the course(s) for the programming language you plan to use to conduct analyses on the Workbench:

- **Introduction to R and Data Visualization**
- **Introduction to Data Wrangling and Analysis Using R**
- **Introduction to Python, Pandas, and Jupyter Notebook for Analysis of *All of Us* Data**.

If you have limited statistical analysis skills or simply need a refresher, you may want to take our course on statistical techniques. It is an intermediate course with a comprehensive overview of essential statistical methods and their practical applications in research:

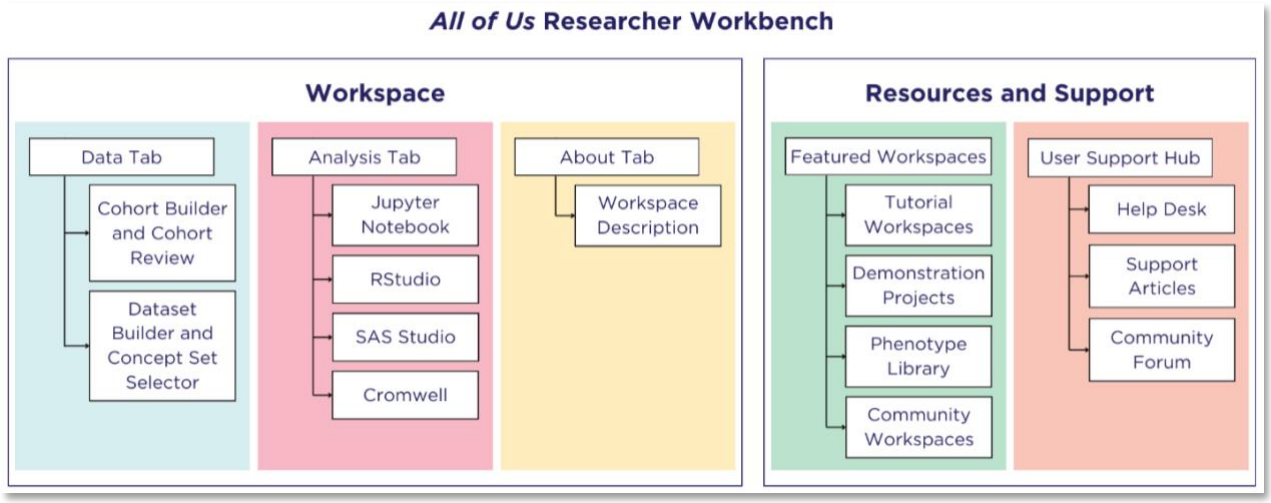
- **Statistical Techniques and Machine Learning for Research Analysis**

3. Using the *All of Us* Researcher Workbench

3.1 Introduction to the *All of Us* Researcher Workbench

The *All of Us* Researcher Workbench is a secure, cloud-based platform where users can access and analyze the *All of Us* dataset. The Workbench contains multiple tools to support you in querying, wrangling, and analyzing the data (**Figure 3-1**).

Figure 3-1. *All of Us* Researcher Workbench Components



Source: Introduction to the *All of Us* Researcher Workbench. (2025). User Support. https://support.researchallofus.org/hc/en-us/articles/360039175232-Intro-to-the-All-of-Us-Researcher-Workbenchmanaging

As you familiarize yourself with the Workbench and its features, Table 3-1 outlines key resources to help users become familiar with the Workbench and serve as a reference as you navigate the platform and conduct your research.

Table 3-1. Resources to Get Started with the *All of Us* Researcher Workbench

Resource	Description
New User Orientation (Video)	This video provides a comprehensive introduction designed for new users of the Workbench. It provides a structured overview of the platform, detailing essential features, tools, and resources that facilitate effective research, as well as a step-by-step walkthrough of how to start a project on the Workbench. Developed by the *All of Us* Support Team, this orientation is tailored to support users in mastering the functionalities of the Workbench, ensuring a smooth onboarding experience. The resource is intended for researchers and students who seek to engage in collaborative research and maximize the potential of the *All of Us* platform.
Starting Your Research (Video Playlist)	This video playlist provides a step-by-step guide to help you navigate the foundational aspects of the Workbench. Covering key topics such as writing a meaningful workspace description, building a cohort, and setting up a billing account, these videos offer practical advice and best practices to set you up for success. Whether you are new to the platform or refining your approach, this playlist equips you with essential tools to start strong and stay on track.

Resource	Description
A Researcher's Guide to the *All of Us* Research Program – User Support	This guide provides background information on the *All of Us* program's experimental design and execution, participant eligibility, enrollment procedures, and more. Researchers should read the guide to understand the strengths, weaknesses, and nuances of the *All of Us* dataset before developing methods to answer their scientific questions and interpret their results. Understanding this information will help avoid errors in methodology and data use.

3.1.1 Workbench Registration

To register for the Workbench, please visit researchallofus.org/academy_sm. The registration process consists of four key steps as shown in **Figure 3-2**:

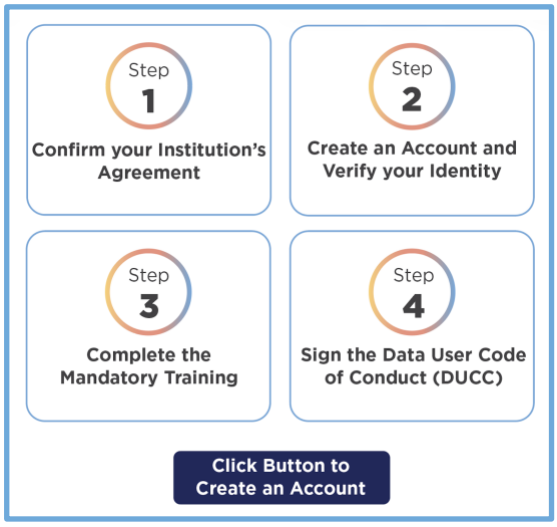
1. Confirm your institution's *All of Us* DURA.
   - To check whether your institution has an existing DURA, visit the DURA page and search for your institution. If your institution does not have a DURA, refer to How to Obtain a DURA with *All of Us* for detailed instructions on how to request one.
   - Two tiers are available:
     - The **Registered Tier** dataset contains individual-level data, available only to registered researchers on the Workbench. The Registered Tier includes data from EHRs, wearables, surveys, and physical measurements.
     - The **Controlled Tier** dataset contains genomic data in the form of short-read whole genome sequences (WGS), long-read WGS, structural variants, and genotyping arrays; previously suppressed demographic data fields from EHRs and surveys; and unshifted dates of events.

       To learn more about particular data fields available within each tier, visit the Registered and Controlled Tier Data Dictionaries.
2. Create an account and verify your identity through Login.gov or ID.me.
3. Complete the mandatory *All of Us* Workbench Training.
4. Sign the Data Use Code of Conduct (DUCC).

IMPORTANT: Please follow the instructions below when answering the survey questions to register for the Workbench.

- How did you learn about the *All of Us* Researcher Workbench?
  - Select: **Other**; then type **RTI-TTT-3** in the box.
- Did you learn about the *All of Us* Researcher Workbench from any of these partners?
  - Select: ***All of Us* Researcher Academy/RTI International**.

Figure 3-2. How to Register for the *All of Us* Researcher Workbench



When setting up your account, select these responses to the following questions:

3.1.2 Data User Code of Conduct

All Workbench users must follow the Data User Code of Conduct (DUCC). Below you will see sections of interest from the Code of Conduct that articulate what ***not*** to do when using the Workbench.

Sections of Interest from the Code of Conduct

As an Authorized Data User of the *All of Us* Research Program data, I will

- **NOT** upload data or files containing personally identifiable information (PII), protected health information (PHI), or identifiable private information (IPI);
- **NOT** take screenshots or attempt in any way to copy, download, or otherwise remove any participant-level data from the *All of Us* Workbench;
- **NOT** publish or otherwise distribute any participant-level data from the *All of Us* Research Program database; and
- **NOT** publish or otherwise distribute any data or aggregate statistics corresponding to fewer than 20 participants unless expressly permitted under the terms of the *All of Us* Data and Statistics Dissemination Policy.

I acknowledge that failure to comply with the requirements outlined in this DUCC may result in termination of my *All of Us* Research Program account and/or other sanctions, including, but not limited to, the posting of my name and affiliation on a publicly accessible list of violators, and notification of the National Institutes of Health or other federal agencies as to my actions.

I understand that failure to comply with these requirements may also carry financial or legal repercussions. Any misuse of the *All of Us* Research Hub, Workbench, or data resources is taken very seriously, and other sanctions may be sought.

It is required that you adhere to the following requirements to ensure compliance with the DUCC and communicate this information when training others to use the Workbench:

- Do ***not*** take screenshots of any PHI or PII information (date of birth, ZIP code, EHR data, etc.).
- Do ***not*** make video recordings of the database of any kind.
- Do ***not*** let anyone who does ***not*** have Controlled Tier access to *All of Us* database view any sensitive information.
- Do ***not*** install any AI bots or malware on Google Cloud, this includes training sessions involving the Workbench database.
- Do ***not*** download or remove any data that is not summarized or has the potential to re-identify a participant.

4. Getting Familiar with *All of Us* Data

Before starting your *All of Us* project workspace, it is important to familiarize yourself with the data and thoughtfully address the following questions. The referenced subsections will guide you through *All of Us* resources to help answer the following four questions:

1. Is your specific research question feasible to answer using the *All of Us* dataset? (Section 4.1.)
2. Have you explored current *All of Us* publications and research projects to verify that your question has not already been answered with *All of Us* data? (Section 4.2)
3. Have you explored the *All of Us* surveys and data to refine your research question and plan your analysis? (Section 4.3)
4. Have you clearly defined your research purpose, scientific approach, and how findings will contribute to the field? (Section 4.4)

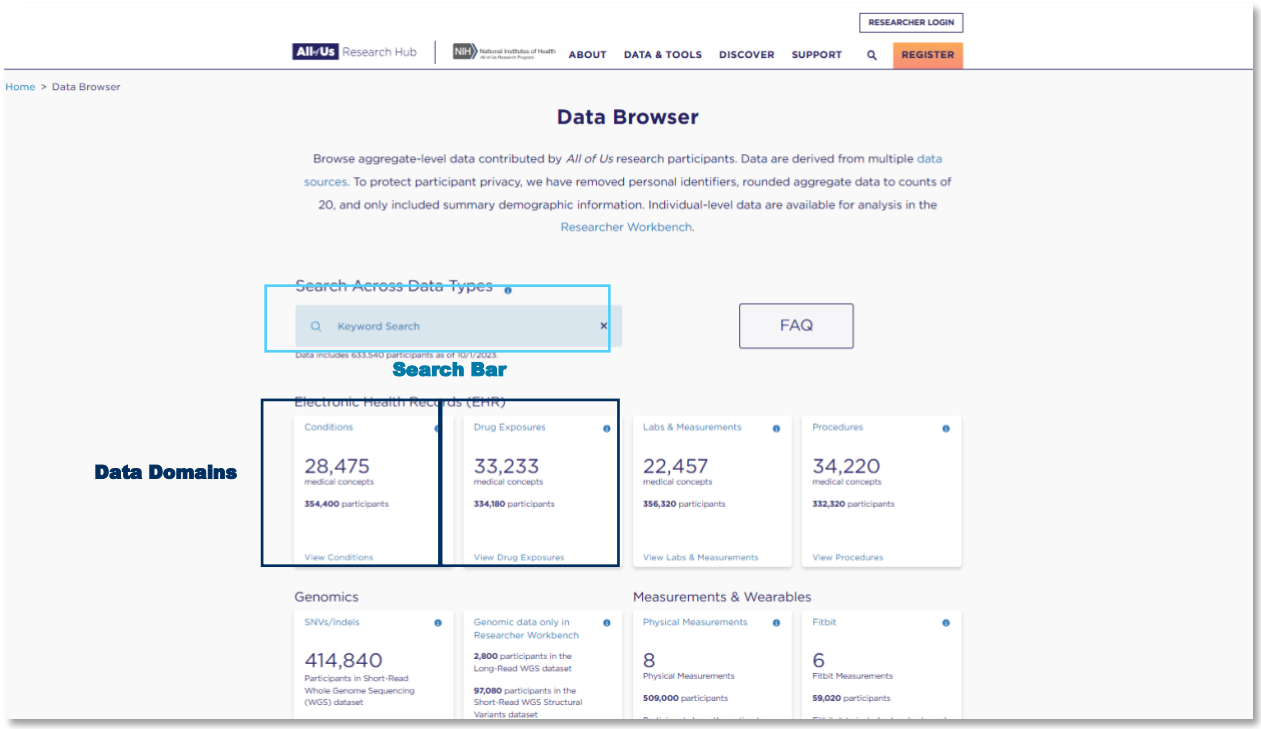
4.1 The Data Browser

To answer Question 1 (Is your specific research question feasible to answer using the *All of Us* dataset?), use the *All of Us* Data Browser.

4.1.1 What Is the Data Browser?

The **Data Browser** is a publicly accessible tool designed to explore variables measured in the *All of Us* dataset and review high-level descriptive statistics (**Figure 4-1**).

Figure 4-1. *All of Us* Data Browser



Additional resources, like **Getting Started with the Data Browser (video)** and **FAQ guides**, provide detailed information about how to navigate the Data Browser.

4.1.2 What Can You Learn from the Data Browser?

The *All of Us* dataset is a valuable resource for research. The nature of the data nonetheless limits the types of questions researchers can explore and answer, and so it is important to consider your research question when considering using the dataset for analysis. Version 8 is the most current version of the *All of Us* dataset and is the version we will use during the training.

Once you have identified a research question, you can use the Data Browser to see what variables are available. The Data Browser can help you to

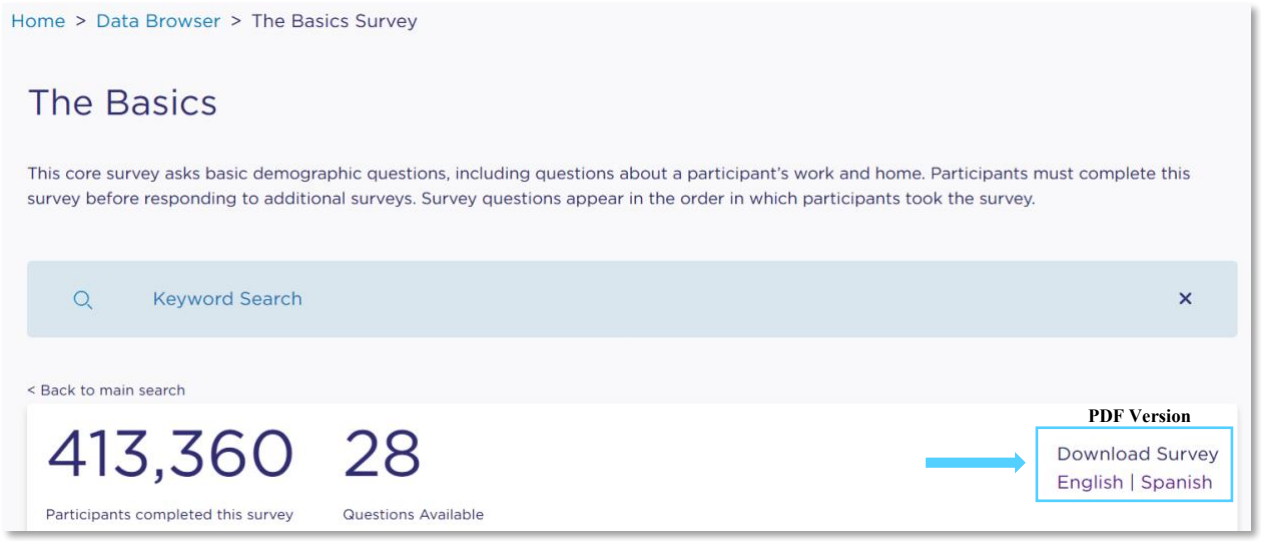
- identify which variables are included in the dataset,
- see a high-level snapshot of sample sizes available for each variable, and
- discover where scales and measures are derived from.

4.1.3 How is the Data Browser Organized?

In the Data Browser, you can easily search for variables or keywords of interest by typing them directly into the **search bar** (see Figure 4-1).

If your search does not yield the desired results, consider exploring the different **data domains** manually (also highlighted in **Figure 4-2**). Each data domain corresponds to a specific survey or set of measurements and includes all variables collected within that category.

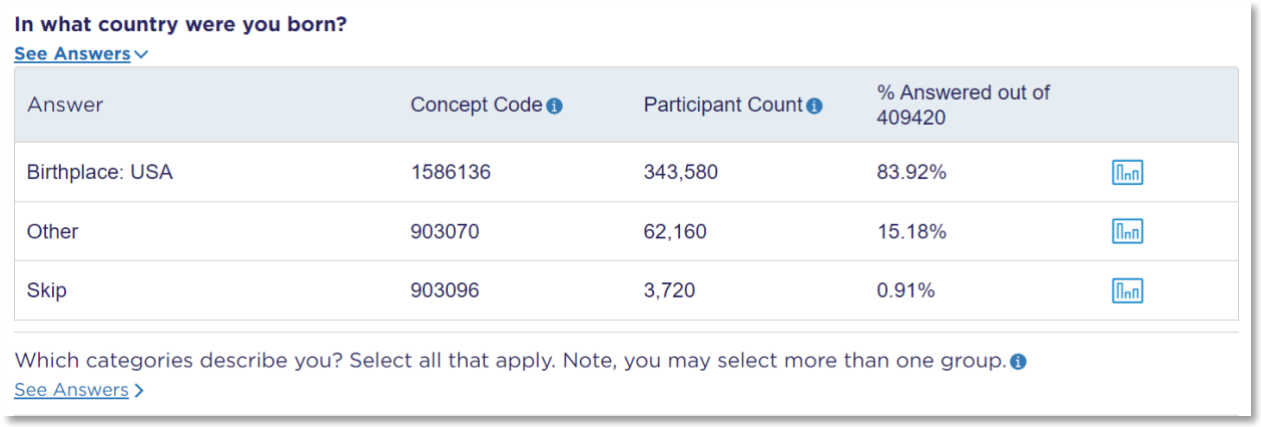
Figure 4-2. Data Domain Description and Measures



Each data domain provides a high-level summary of the variables and participant information, along with detailed variable data. Researchers can also access and download PDF versions of the surveys in English and Spanish for further reference.

For example, **Figure 4-3** presents a breakdown of participant counts and percentages for responses or outcomes to specific questions. The concept code shown corresponds to the unique identifier for each question or response in the dataset. By clicking on the icon, you can access a bar chart view that displays the selected variable options aggregated by age, for example.

Figure 4-3. Preview of High-Level Data for an Individual Variable within a Data Domain



4.1.4 Where Can Each Data Source Be Found in the Workbench?

While summaries of data contributed by *All of Us* participants can be found in the Data Browser, registered researchers can access individual-level data from various sources through the Workbench. As you transition from developing your research focus to conducting your research, please refer to **Table 4-1** to locate your data source(s) of interest within the Workbench.

Table 4-1. Data Sources and Location Matrix

Data Source	Description	Access Tier	Location
Surveys	Self-reported participant responses covering the following Surveys: The Basics; Overall Health; Lifestyle; Health Care Access and Utilization; Personal/Family History; Social Factors of Health; COVID-19 Participant Experience (COPE) (responses closed); Minute Survey on COVID-19 Vaccines (responses closed)	Registered Tier (individual-level survey data are available on the Workbench)	Access via the Cohort Builder and Dataset Builder For full text versions of each survey, please refer to the Survey Explorer
Electronic Health Records (EHRs)	Longitudinal clinical records capturing diagnoses, medications, lab results, procedures, and other details from healthcare encounters	Registered Tier (less detailed and granular individual-level EHR data are available to approved researchers) and Controlled Tier (more detailed and granular EHR data are available to approved researchers)	Access via the Cohort Builder and Dataset Builder. For more detailed instructions, please refer to Intro to Electronic Health Record (EHR) Data (Video)

Data Source	Description	Access Tier	Location
ZIP Code Data	Three-digit ZIP code–level data are the most granular geolocation data available within the Workbench. In cases where fewer than 20,000 participants are reported to reside in a three-digit ZIP code area, those participants are aggregated into another nearby three-digit ZIP code area	Controlled Tier (restricted due to privacy concerns)	Access via the Dataset Builder and select the pre-packaged concept set **"ZIP Code Socioeconomic Status Data"**
Genomic Data (Biosamples and Bioassays)	Biological specimens (e.g., blood and saliva) collected from participants that are used for genomic analysis and biomarker research, including short-read whole genome sequencing data (srWGS); structural variant data; All by All tables (genome and phenome-wide analysis); long-read whole genome sequencing data (lrWGS); and more	Controlled Tier (genomic data derived from biosamples are available only to researchers with Controlled Tier access)	Access via the Cohort Builder and Dataset Builder For more detailed instructions, please refer to Intro to Genomic Data (Video)
Physical Measurements	Direct measurements (e.g., vital signs, anthropometrics, lab test results) gathered during study visits or clinical encounters	Registered Tier (individual-level physical measurement data are provided through the Workbench)	Access via the Cohort Builder and Dataset Builder
Wearable Devices and Digital Health (Fitbit)	Data from wearable devices (e.g., Fitbit) that capture activity, heart rate, sleep patterns, and other digital health metrics	Registered Tier (wearables data are integrated into the Workbench as non-OMOP tables)	Access via the Cohort Builder and Dataset Builder For more detailed instruction, please refer to Intro to Fitbit Data (Video)
(NEW) Country of Origin Data	Researchers can now integrate participants' country of origin in their research and analysis plan	Controlled Tier (restricted due to privacy concerns)	Access via the Dataset Builder by adding the following question into a concept set: "In what country were you born?" (1586135)

Data Source	Description	Access Tier	Location
(NEW) Emotional Health History and Well-Being (Survey)	The Emotional Health History and Well-Being Survey asks information about mental health functioning (e.g., depression symptoms, trauma experiences) and well-being	Registered Tier	Access to this survey data is NOT yet available within the Cohort Builder For the full text version of this survey, please refer to the **Survey Explorer** For more details on how to access please visit the following Jupyter Notebook, **Emotional Health History and Well-Being (EHHWB) & Behavioral Health and Personality (BHP) (R)**
(NEW) Behavioral Health and Personality (Survey)	The Behavioral Health and Personality Survey asks information about mental health functioning (e.g., ADHD symptoms, experiences of psychosis) and personality	Registered Tier	Access to this survey data is NOT yet available within the Cohort Builder For the full text version of this survey, please refer to the **Survey Explorer** For more details on how to access please visit the following Jupyter Notebook, **Emotional Health History and Well-Being (EHHWB) & Behavioral Health and Personality (BHP) (R)**

For more information on these data sources and how they are curated within the Workbench, please visit **Data Sources**.

4.1.5 Current Data Version: Curated Data Repository Version 8

The **Curated Data Repository (CDR)** is the cornerstone of the *All of Us* Research Program, serving as the central hub for the vast health data contributed by participants. This resource brings together information from various sources, including survey responses, EHRs, physical measurements, genomic data, and wearable device data. The release of **CDR Version 8 (CDRv8)** greatly increased the breadth and depth of available information to researchers. The following are some of the key updates that were made in CDRv8:

- **Data from 633,000 participants**: This represents a 53% increase.
- **69% more genomic data**: CDRv8 includes WGS from more than 414,000 participants, with over 1.2 billion genetic variants identified, more than 90,000 participants with short-read WGS structural variants data available, and 2,000+ participants with long-read WGS data.
- **A massive expansion of wearable data**: CDRv8 has quadruple the number of participants with records now at 59,000, an outcome of the Wearables Enhancing *All of Us* Research (WEAR). This has greatly expanded upon what was already the largest public dataset of Fitbit information available.
- **Expansion of population descriptors:** Researchers in the Controlled Tier can now access self-reported sub-categories of population descriptors to support research.
- **American Indian and/or Alaska Native participant data:** The inclusion of data contributed by more than 26,000 participants who self-identify as American Indian and/or Alaska Native follows an extensive engagement process, including Tribal consultation and information sessions and the creation of **new guidance and updated policies** to help researchers use the data responsibly.
- **Cognitive and mental health and well-being data:** New cognitive task data from the Exploring the Mind study supported by the National Institute of Mental Health are available from more than 36,000 participants. Additionally, over 110,000 participant responses to the Mental Health and Well-Being surveys are now available.

For more detailed information on CDRv8 and earlier versions of *All of Us* data, visit **Curated Data Repository (CDR) Version 8 Release Notes** and **Curated Data Repository Release Notes**. Information on the most recent changes to the Workbench, including features, updates, bug fixes, and new data releases, are provided in the Workbench **Release Notes**.

Key Takeaways: Using the *All of Us* Data Browser

- Ensure that your research question aligns with the dataset's capabilities, including the way(s) that variables and concepts were measured and the limits of the data.
- After defining your research question, review the dataset thoroughly to confirm it contains the necessary variables, sufficient sample size, and appropriate measures for your analysis.
- Stay informed about new CDR versions and data updates, as these enhancements can significantly expand research possibilities and potentially influence or refine your research questions. The Announcements and News section of the *All of Us* website is an excellent resource to stay up to date on the program and the dataset.

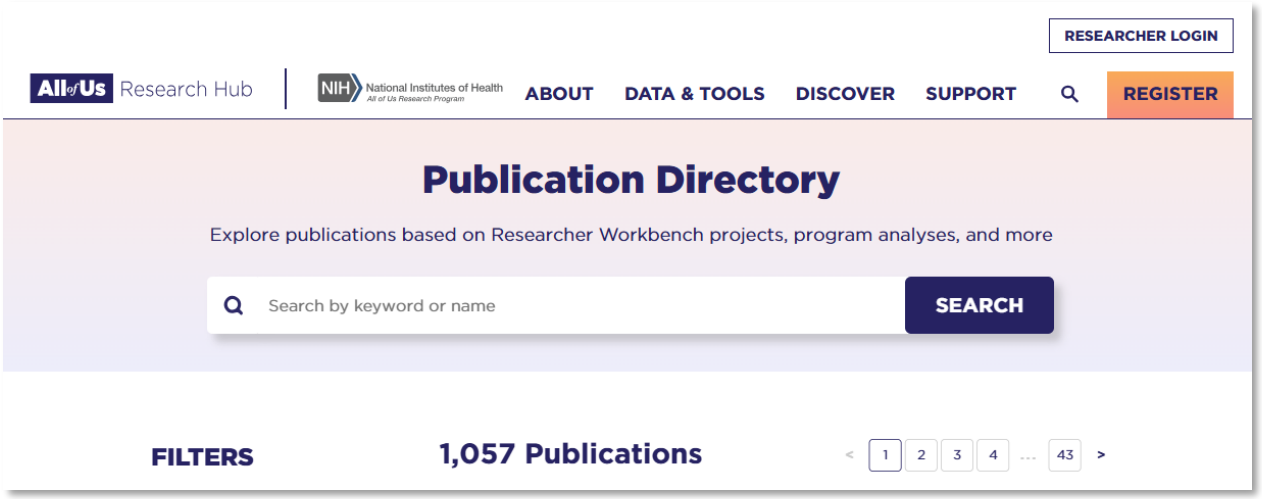
4.2 The Publication and Research Projects Directories

To answer Question 2 (Have you explored current *All of Us* publications and research projects to verify that your question has not already been answered with *All of Us* data?), there are two key resources to leverage. First, to see if the answer to the research question has already been published, use the *All of Us* Publication Directory. However, a key advantage to conducting research in *All of Us* is that you also have visibility into what other research teams are doing. To see if another research team is already analyzing data on your research question, you can use the *All of Us* Research Projects Directory.

4.2.1 What Is the Publication Directory?

The *All of Us* Publication Directory is an online catalog of research articles, reports, and other scholarly works that use or reference the *All of Us* Research Program. You can search and browse publications by author, date, topic, or keyword to find citations, links to full texts, and brief details about each item. It is a useful resource for researchers looking to see if their research question has already been answered in *All of Us*, but is also great for finding examples of analyses, methods, and findings that showcase how the *All of Us* dataset has been used.

Figure 4-4 *All of Us* Publication Directory

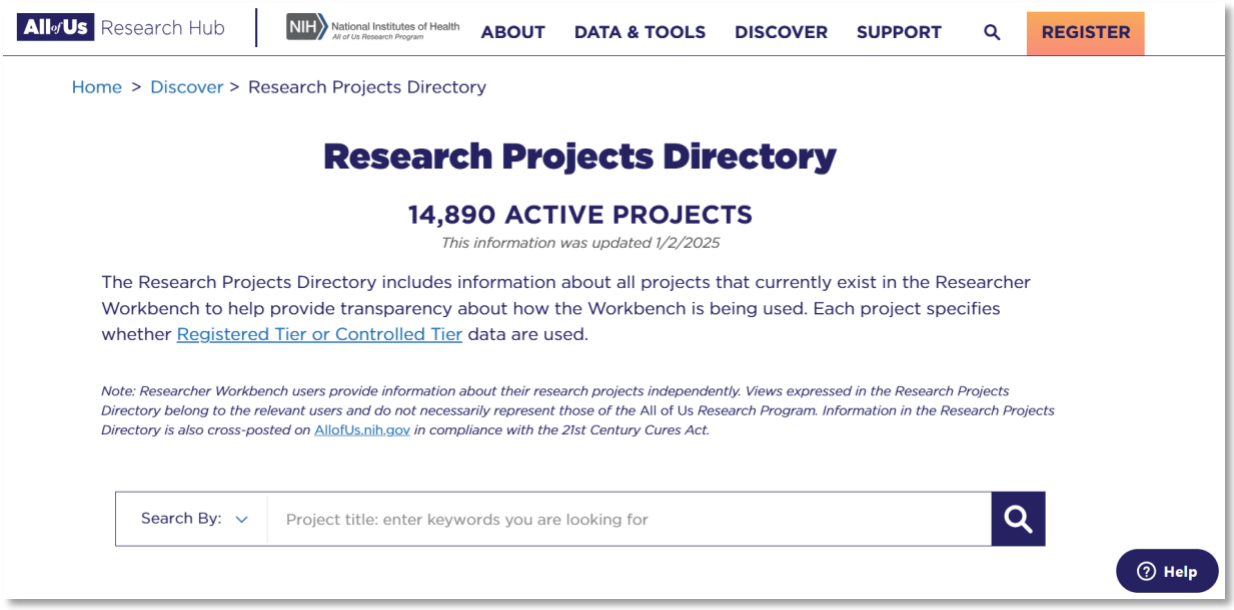


4.2.2 What Is the Research Projects Directory?

The Research Projects Directory provides information about all ongoing projects within the Workbench, offering transparency into how the platform is being used. Each project specifies whether Registered Tier or Controlled Tier data are used.

You can search for active projects within the Workbench by using the Project Directory "Search By" feature, as shown in **Figure 4-5**. This allows you to find projects based on their title, owner, or the scientific questions being studied.

Figure 4-5. *All of Us* Research Projects Directory



4.2.3 How Can You Identify a Research Gap to Explore Using These Resources?

The *All of Us* dataset provides researchers with access to a wide range of health-related data, including healthcare access, health behaviors and outcomes, EHR information, genomic data, and more. Before starting an analysis, one of the first steps should be identifying a research gap that can be effectively addressed using the available data. The first step is to see what research has already been published using the *All of Us* data on your topic of interest or research question. The best place to look for these answers is the **Publication Directory.**

Whether or not work has been published on your topic or research question, you may also wish to look into the **Research Projects Directory.** Here, you can find detailed information about both published and unpublished analyses, including the

- research questions being asked and the anticipated findings,
- scientific approaches being applied,
- types of variables on which researchers are focusing,
- researchers collaborating on each project, and
- data tier being used for the project.

Key Takeaways: Using the Publication and Research Project Directories

- The *All of Us* Publication Directory is an online catalog of research articles, reports, and other scholarly works that use or reference the *All of Us* Research Program
- The Research Projects Directory is a valuable resource for understanding the types of research questions that can be explored using *All of Us* data.
- Before starting your analysis, use both resources to check whether your research question has already been addressed by existing studies.

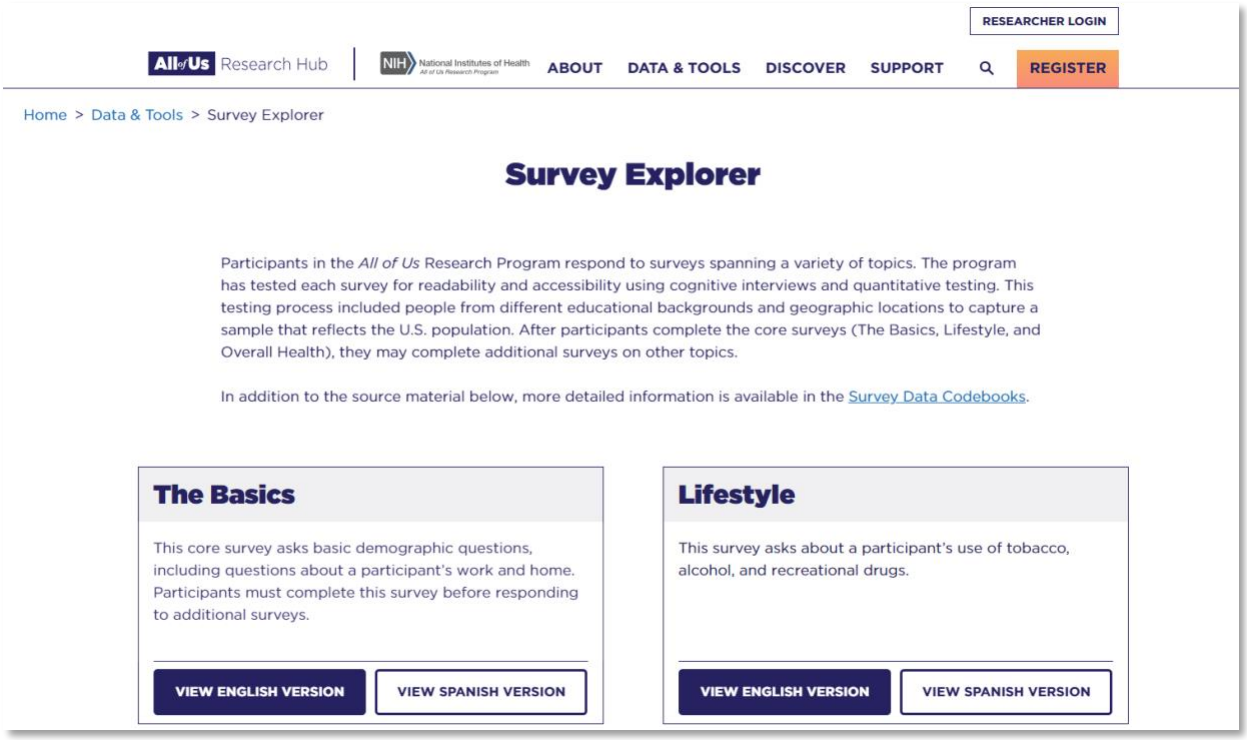
4.3 The Survey Explorer

To answer Question 3 (Have you explored the *All of Us* surveys and data to refine your research question and plan your analysis?), use the *All of Us* Survey Explorer.

4.3.1 What Is the Survey Explorer?

The **Survey Explorer** (**Figure 4-6**) is a tool that allows you to browse the questions in the *All of Us* Program surveys and to see the source information for each question.

Figure 4-6. *All of Us* Survey Explorer



Clicking on the English or Spanish version options in each survey will direct you to a PDF file of the corresponding survey. Additional resources, like the FAQ guides, provide detailed information about how to navigate the Survey Explorer.

Key Takeaways: Using the Survey Explorer

- Use the Survey Explorer to identify the origins of survey questions and scales.
- If you plan to use survey data for your research project, be sure to use the Survey Explorer to refine and develop a focused, specific research question that can make a meaningful contribution to the research literature.

4.4 Finalizing Research Objectives

To answer Question 4 (Have you clearly defined your research purpose, scientific approach, and how findings will contribute to the field?), you will need to create a workspace on the Workbench.

As outlined in Section 4.2, every project on the Workbench must include a description for inclusion in the Research Projects Directory. **Section 6** provides a step-by-step guide to creating a workspace on the Workbench and detailed instructions on how and where to input your research project description.

5. The Observational Medical Outcomes Partnership Common Data Model for EHR Data

The OMOP CDM is a standardized framework used in healthcare research to organize and structure clinical data from different sources—such as EHRs, insurance claims, and registries—into a consistent format. It was created by the Observational Health Data Sciences and Informatics (OHDSI) collaborative to enable large-scale, reproducible, and reliable observational research.

This section serves only as a brief introduction to the OMOP CDM as it is used in the Workbench. The Book of OHDSI is the gold-standard resource if you want to learn more about OHDSI, OMOP, and the OMOP CDM. Chapter 4 explains the CDM, and Chapter 5 delves deeper into the Standardized Vocabularies.

The EHR data available for analysis in the Workbench is an artifact created by providers when they are treating patients for the purpose of documentation and billing. Because different providers and healthcare systems have different ways of documenting and billing for the same services, these data are often stored in incompatible formats, use different coding systems, and follow site-specific conventions. These differences make it difficult to combine data or compare results across institutions. The OMOP CDM solves this by providing

- a common structure so datasets look and function the same across institutions,
- standardized vocabularies to harmonize clinical terms, and
- a shared analytics ecosystem where tools and methods work on any OMOP-formatted data.

All of Us uses the OMOP CDM to standardize participant health information across diverse providers, formats, and data types. By transforming all clinical data into OMOP, *All of Us* ensures consistency, enables scalable analysis, and allows researchers to use a shared set of tools and methods to generate reliable, comparable findings.

5.1 OMOP CDM Structure and Tables in *All of Us*

The *All of Us* EHR data are built as a collection of interrelated OMOP CDM tables that center on individual participants. Most clinical event tables link back to the PERSON table via person_id so that, together with event dates, you can reconstruct each participant's longitudinal healthcare timeline. In addition to those event tables (conditions, visits, drugs, measurements, procedures, observations, devices, specimens, deaths, etc.), the CDM includes metadata and vocabulary tables (concept, concept_relationship, vocabulary, cdm_source) and *All of Us* extension tables that record provenance- and program-specific fields. Note that a few standardized health-system tables are modeled differently and are associated directly with domain-specific events rather than routed solely through PERSON. **Table 5-1** provides short, descriptive summaries of the purpose and contents of each table to help researchers find the fields they need.

Table 5-1. Descriptive Summaries of the OMOP CDM Tables as Used in *All of Us*

Table	Summary
person	One row per participant: unique identifier, demographic attributes (sex/gender, birth year/month/day, race, ethnicity), residence/location and care links, and source codes/*All of Us* custom demographic fields
condition_occurrence	Individual condition/diagnosis records with start/end dates and datetimes, mapped standard and source codes, provenance/type, provider and visit links, status and stop-reason information
visit_occurrence	Records of healthcare encounters with start/end dates and datetimes, visit type and source codes, provider and care site links, admitting/discharge details, and links to preceding visits
visit_detail	More granular encounter records (sub-visits) with start/end datetimes, type and source codes, provider/care site links, parent/preceding relationships and visit_occurrence linkage
drug_exposure	Drug prescription/administration events with start/end dates and datetimes, mapped drug and source codes, type/modifier, dosage/quantity/refills/days supply, route/lot, and provider and visit links
measurement	Clinical measurement or lab result rows recording date/time, standard and source measurement concepts, numeric or coded values, units, normal ranges, operator, and links to provider/visit
procedure_occurrence	Procedure events documenting date/time, standard and source procedure codes, type/modifier, quantity, provider and visit linkage, and qualifier/source details
observation	Recorded observations (including personal protected information [PPI] answers) with date/time, standard and source concept mapping, value as number/string/concept, units and qualifiers, and provider/visit links
observation_period	Time windows of available data for a person: observation period start/end dates and the provenance/type of that period
device_exposure	Device or supply use events with start/end dates and datetimes, standard and source device codes, unique device identifiers, quantity, and provider/visit linkage
death	Death records with person linkage, death date/datetime, death type and cause concepts, and source cause coding
fact_relationship	Linking table that records relationships between two facts (domain/table and row identifiers) and the relationship concept describing that link
specimen	Biospecimen metadata including specimen identifier, person link, specimen type and anatomic site, collection date/time, quantity and unit, disease status, and source identifiers/values

Table	Summary
condition_occurrence_ext	*All of Us* metadata row linking condition_occurrence rows to data provenance (src_id) and the original condition_occurrence identifier
device_exposure_ext	Metadata row linking device_exposure records to data provenance (src_id) and the original device_exposure identifier
drug_exposure_ext	Metadata row linking drug_exposure records to data provenance (src_id) and the original drug_exposure identifier
measurement_ext	Metadata row linking measurement records to data provenance (src_id) and the original measurement identifier
observation_ext	Metadata row linking observation records to data provenance (src_id) and the original observation identifier
procedure_occurrence_ext	Metadata row linking procedure_occurrence records to data provenance (src_id) and the original procedure_occurrence identifier
visit_occurrence_ext	Metadata row linking visit_occurrence records to data provenance (src_id) and the original visit_occurrence identifier
cdm_source	CDM instance metadata: source name/abbreviation/holder, descriptive summary, documentation/ETL references, source and CDM release dates, and CDM/vocabulary versions
attribute_definition	Definitions of custom attributes: identifier, short name and full description, attribute type concept, and syntax to operationalize the attribute
concept	Vocabulary concept registry rows: unique concept id and name, domain, source vocabulary and class, standard-concept flag, source concept code, validity dates and invalidation reason
concept_ancestor	Hierarchy relationships between concepts listing ancestor and descendant concept ids and minimum/maximum levels of separation
concept_class	Concept class definitions with id, display name, and linked concept identifier describing the class
concept_relationship	Concept-to-concept mappings: source and target concept ids, relationship type id, validity dates, and invalidation reason
concept_synonym	Alternate names for concepts with the concept id, synonym text, and language identifier
domain	Definitions of OMOP domains with domain id, human-readable domain name, and linked domain concept id

Table	Summary
drug_strength	Drug formulation and strength records linking drug and ingredient concepts with absolute and concentration amounts, units, box size, and validity metadata
relationship	Relationship vocabulary entries: relationship id and name, flags for hierarchy/ancestry, reverse relationship id, and linked relationship concept identifier
vocabulary	Vocabulary registry rows: vocabulary id and name, external reference/documentation, version, and linked vocabulary concept id
person_ext	*All of Us* custom person fields such as generalized state-of-residence concept and source value (with suppression rules for small counts)
survey_conduct	Survey response metadata: unique survey response id, person link, survey concept/version, start/end date and datetimes, respondent/timing/collection method and related source values, provider and visit links, and validation status
survey_conduct_ext	Survey metadata linking survey_conduct rows to provenance (src_id), language of completion, and the survey_conduct identifier
aou_death	*All of Us*–specific death table: GUID death record id(s) per person, person link, death date/datetime (nullable), death type/cause codes and source fields, provenance (src_id), and primary_death_record flag

For a full description of the tables and specific variables that appear in each table, please consult the **OMOP-Compatible Tables sheet of the *All of Us* Data dictionary**. The "Additional Notes" column in the OMOP-Compatible Tables sheet describes cases in which *All of Us* made changes to the standard OMOP model. However, a few key differences from the standard OMOP CDM should be noted:

- **Provider identifiers and provenance**: In *All of Us,* the provider_id that normally links many OMOP rows to a specific clinician is masked, so direct provider-level linking is not available. Instead, provenance is indicated by src_id values (the recruiting site or data source), which are included in the *All of Us* extension tables (e.g., condition_occurrence_ext) when source attribution is required.
- **Additional *All of Us* tables and useful UI tool tables**: Beyond the standard OMOP tables, *All of Us* includes several extension tables (the *_ext tables) that surface src_id and other provenance metadata, plus a set of Workbench "UI tool" tables that can be helpful for analysis. For example, cb_search_person contains flags and summary fields about participants that are not present in the base PERSON table. See the ***All of Us* data dictionary tab** listing these tables for details and recommended uses.

- **Program-specific observations inserted into OMOP:** *All of Us* adds certain program-specific records into the OBSERVATION table (e.g., consent dates, EHR consent status, survey completion timestamps, and enrollment dates). These items use observation concept codes that may not be obvious without consulting the documentation, so consult the referenced resources (**EHR consent**, **enrollment date**) and the vocabulary mapping tables when searching for those values.

5.2 OMOP CDM Concepts and Source Values

In the CDM, individual data values are standardized by encoding them as concepts. The tables store these concept_id keys that point to a central concept table, which serves as the general reference table. The concept table contains metadata about each term (e.g., its name, domain, and class), while auxiliary tables, such as concept_relationship and concept_ancestor, capture the relationships and hierarchies needed by the standardized vocabularies.

Many OMOP tables store the same clinical information in three complementary forms: the original source value, a source-mapped concept, and a standardized concept. These serve different purposes:

- **Source values** are the raw codes or text as they appeared in the originating system (e.g., ICD9CM, NDC, CPT4, or local site codes, or even short free-text entries like "F" or "M." They appear in fields named [EVENT]_SOURCE_VALUE and are useful for provenance checks and QA because they preserve exactly what the source provided.
- **Source concepts** are the concept records that represent those original codes when the code system is known; these are recorded in [EVENT]_SOURCE_CONCEPT_ID and are used only for codes that exist in source vocabularies.
- **Standard concepts** are the normalized concept identifiers chosen as the canonical representation of a clinical meaning across all CDM instances; these are stored in [EVENT]_CONCEPT_ID. When a non-standard concept has an equivalent standard concept, the standardized vocabulary contains a mapping from the source (non-standard) concept to the corresponding standard concept.

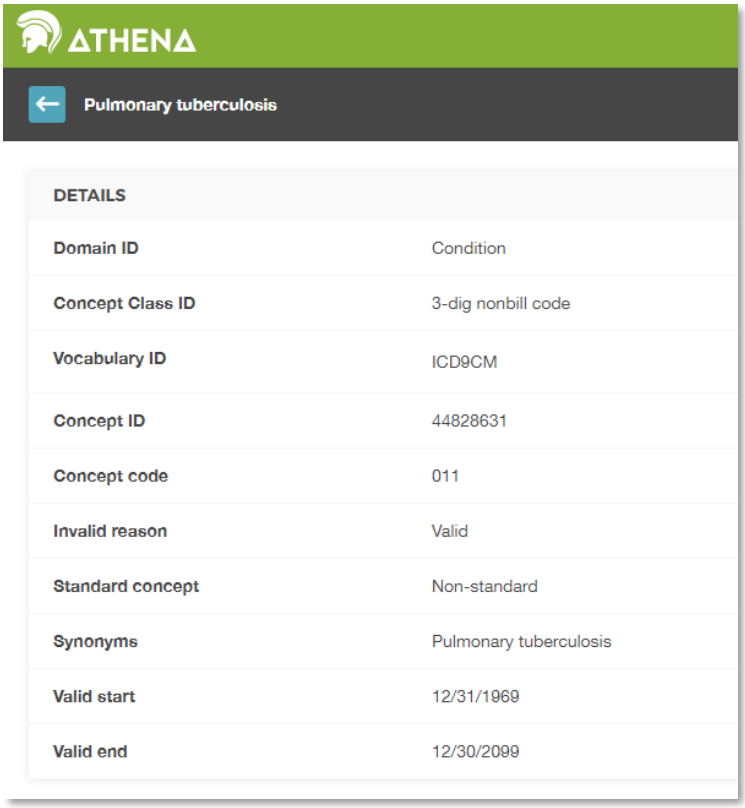
Providing source values helps in retaining traceability and validation. In *All of Us,* using standard concepts is highly recommended. This method enforces normalization and is what enables interoperability across the EHR from the many different healthcare systems that contribute data to *All of Us*. For example, the raw ICD9CM code "011" without context could mean very different things in different vocabularies, but its mapped CONCEPT_ID points to the correct ICD9CM entry, and that source concept is linked via a "non-standard → standard" mapping to a SNOMED standard concept for pulmonary tuberculosis. Because such mappings exist for ICD, Read, MeSH, and other code sets, querying by the SNOMED standard concept will reliably capture all equivalent source codes.

To illustrate, **Figure 5-1** shows the ICD9CM code for Pulmonary Tuberculosis, and **Figure 5-2** shows the SNOMED code for Pulmonary Tuberculosis. Alone, the concept code "011" (Figure 5-1) is ambiguous. In different vocabularies, it can mean very different things (e.g., a UB04 entry for "Hospital Inpatient" or a DRG entry for a nervous-system neoplasm). That ambiguity is resolved by using concept identifiers: the ICD9CM code 011 is represented by CONCEPT_ID 44828631, which ties that source code to its correct source vocabulary and distinguishes it from

similarly numbered codes in UB04 or DRG. That ICD9CM source concept is then linked via an OMOP "non-standard → standard" mapping to the SNOMED standard concept 253954 for pulmonary tuberculosis. Equivalent mappings exist for other code systems (e.g., Read, ICD-10, CIEL, MeSH), so querying the SNOMED standard concept reliably retrieves all corresponding source codes.

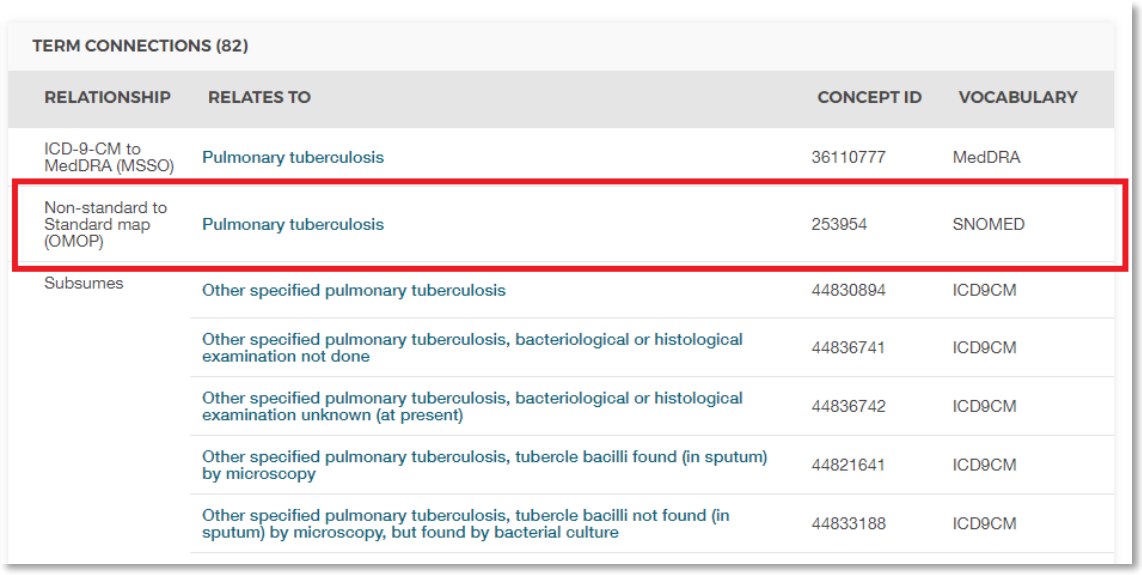
To illustrate how you might use these tables in practice, please see The Book of ODHSI Chapter 4.3 and the example of a single patient's experience with endometriosis.

Figure 5-1. ICD9CM Code for Pulmonary Tuberculosis



Source: Observational Health Data Sciences and Informatics (OHDSI) collaborative. (2021). *The Book of OHDSI.* https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html

Figure 5-2. SNOMED Code for Pulmonary Tuberculosis



TERM CONNECTIONS (82)

RELATIONSHIP	RELATES TO	CONCEPT ID	VOCABULARY
ICD-9-CM to MedDRA (MSSO)	Pulmonary tuberculosis	36110777	MedDRA
Non-standard to Standard map (OMOP)	Pulmonary tuberculosis	253954	SNOMED
Subsumes	Other specified pulmonary tuberculosis	44830894	ICD9CM
	Other specified pulmonary tuberculosis, bacteriological or histological examination not done	44836741	ICD9CM
	Other specified pulmonary tuberculosis, bacteriological or histological examination unknown (at present)	44836742	ICD9CM
	Other specified pulmonary tuberculosis, tubercle bacilli found (in sputum) by microscopy	44821641	ICD9CM
	Other specified pulmonary tuberculosis, tubercle bacilli not found (in sputum) by microscopy, but found by bacterial culture	44833188	ICD9CM

Source: Observational Health Data Sciences and Informatics (OHDSI) collaborative. (2021). *The Book of OHDSI*. https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html

6. Workspaces

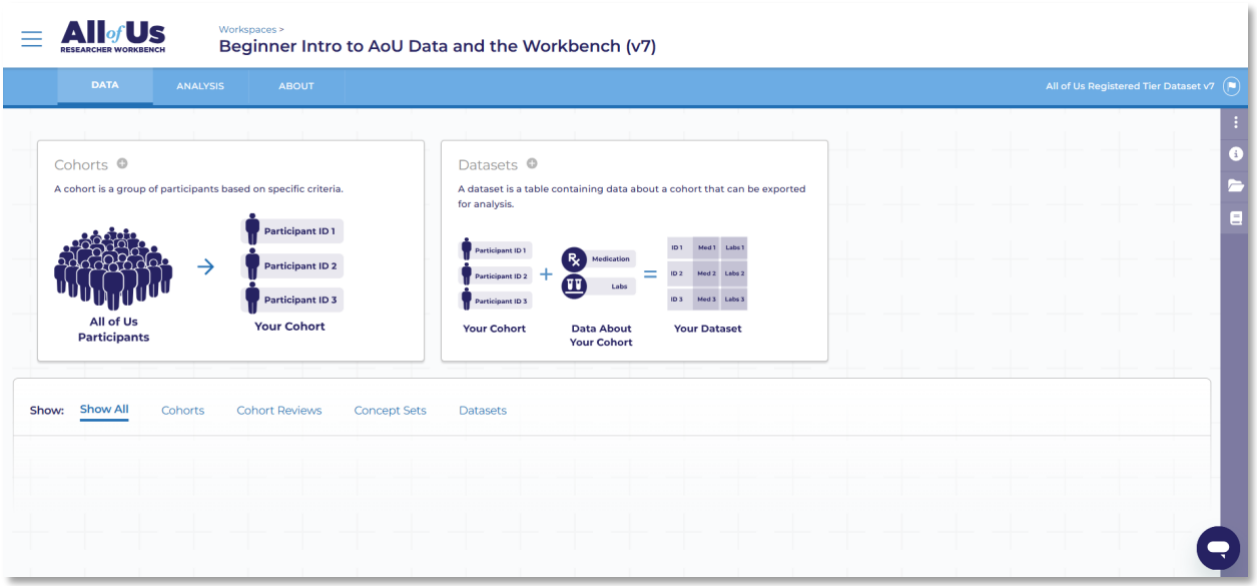
6.1 What Is a Workspace?

All analyses on the Workbench happen in a **workspace**, which is your place to store, build, and analyze your dataset. For projects with multiple researchers, workspaces serve as your collaboration space for team science.

As shown in **Figure 6-1**, each workspace has three main tabs:

- Under the **Data** tab, you can use the Cohort Builder to select participants based on specific criteria and use the Dataset Builder to select variables of interest (concepts) that you are interested in researching in your cohort.
- Under the **Analysis** tab, you can perform queries and analyses using integrated cloud-based analysis tools, including Jupyter Notebook, RStudio, and SAS Studio with programming languages Python, R, and SAS.
- Under the **About** tab, you can view your workspace description, which provides an overview of your research purpose, scientific approach, anticipated findings, and more.

Figure 6-1. Featured Workspace on the Workbench



6.1.1 Creating a Workspace

For detailed steps on how to create a workspace, please visit **Creating a Workspace**. **Figure 6-2** highlights the starting point to create your workspace, and **Table 6-1** outlines aspects to consider when creating a workspace.

Figure 6-2. Workspace Creation Button on the Workbench

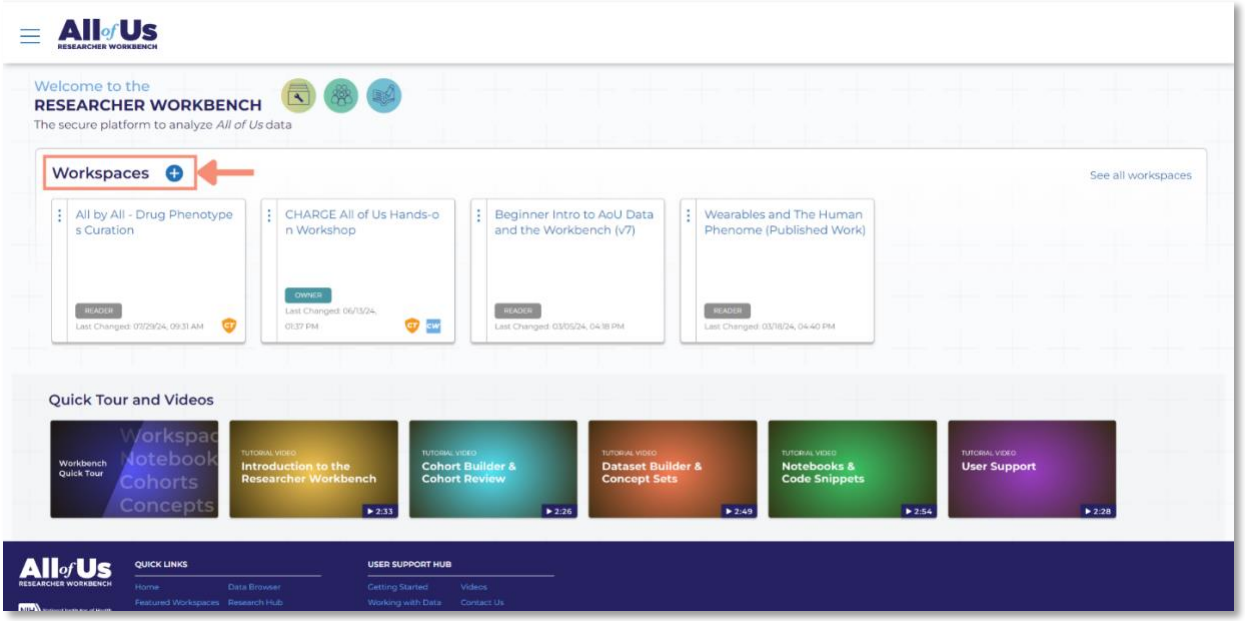


Table 6-1. Workspace Considerations

Aspect	Key Considerations
Data Access Tier Selection	Choose between the **Registered Tier** and **Controlled Tier** based on data sensitivity. Once selected, the tier cannot be changed—a new workspace is required if an incorrect tier is chosen. For details on data within each tier, refer to Data Access Tiers.
Dataset Version	Select the most recent **Curated Data Repository (CDR)** version unless replicating a study that used an earlier dataset. For details on how data are curated, refer to Data Methods and CDR Version 8 Release Notes (the most recent CDR version).
Initial Credits and Billing Account Setup	Each Workbench account begins with **$300 in initial credits** for research analysis activities. **Initial credits will expire 365 days** after registering for the Workbench. Researchers can check how much of the initial credit balance has been spent at any time by visiting their Workbench profile page. For detailed instructions, visit Using *All of Us* Initial Credits – User Support**.** An active billing account is needed to cover any computational and storage costs beyond the initial $300 in credits. You also have access to $300 in promotional credits from Google, if you have not already used the credits from another Google Cloud Platform (GCP) product. This User Support Hub article summarizes how to set up a GCP account. The GCP credits last for **90 days** once they have been activated. Watch Billing in the Workbench and Paying for Your Research to learn more about billing in the Workbench.
Workspace Management	Familiarize yourself with **sharing**, **duplicating**, **editing**, and **deleting** workspaces to maintain efficient workflows. For detailed instructions, visit Managing Workspaces.

Aspect	Key Considerations
Collaboration Access Levels	If working with a research team, set access levels for your workspace carefully: ▪ **Readers** can view, but not edit, notebooks. Readers cannot delete or share the workspace. ▪ **Writers** can view and edit notebooks, as well as delete files in the workspace bucket, the permanent storage area within the Google Cloud Platform (GCP). Writers cannot delete or share the workspace. ▪ **Owners** have full permissions and can view and edit notebooks, as well as delete files in the workspace bucket. Owners can delete and share workspaces, so this level of access should be granted with caution.
Collaboration Requirements	Researchers must have the same access tier (Registered or Controlled) as the workspace to collaborate. If access is insufficient, researchers will receive a prompt to complete additional requirements.
Credit Exhaustion Workaround	If a workspace owner exhausts their $300 initial credits, a collaborator with remaining credits can **duplicate** the workspace, assume ownership, re-share it with the team, and continue the project, delaying the need to set up a billing account.

Key Takeaways: Using Workspaces

- Select the correct data tier (Registered or Controlled) for your workspace up front, as it cannot be changed later.
- Collaborators can duplicate the workspace and take ownership to use their remaining credits, delaying the need to set up a billing account.

6.2 Writing Your Workspace Description

As outlined in Section 4.2, every project on the Workbench must include a description for inclusion in the Research Projects Directory. For detailed guidance on writing your workspace description, please visit **Writing Your Workspace Description** or access the **Writing a Meaningful Workspace Description (Video)**.

Additional context and tools for crafting your workspace description can be found in Section 4 of this manual, Getting Familiar with *All of Us* Data.

6.2.1 Actively Updating Your Workspace Description

You do not need to finalize your workspace description at the time of submission. As your research progresses, you may need to adjust your approach, including adding or removing variables, refining research questions, or modifying methods. Because your workspace description will be publicly visible on the *All of Us* Research Projects Directory (see Section 4.2), it is important to provide a clear and general overview of the research you plan to conduct using *All of Us* data.

You can edit your workspace description at any point during your research. Keeping your workspace description accurate and up to date is highly recommended to ensure that it reflects your current research objectives and methods. For instructions on how to edit your workspace description, please visit **Managing Workspaces**.

After writing your workspace description and clicking "Create Workspace," you will see the Data Tab in your new workspace, which is where you will select participants using the Cohort Builder.

Key Takeaways: Writing Your Workspace Description

- Your workspace description should provide a clear overview of your planned research, even if details evolve later.
- Edit your workspace description as needed to keep it aligned with your current research as it will be publicly available on the *All of Us* Research Projects Directory.

6.3 Working with Others in Shared Workspaces

When collaborating in a shared workspace, there are a few key issues and best practices to keep in mind.

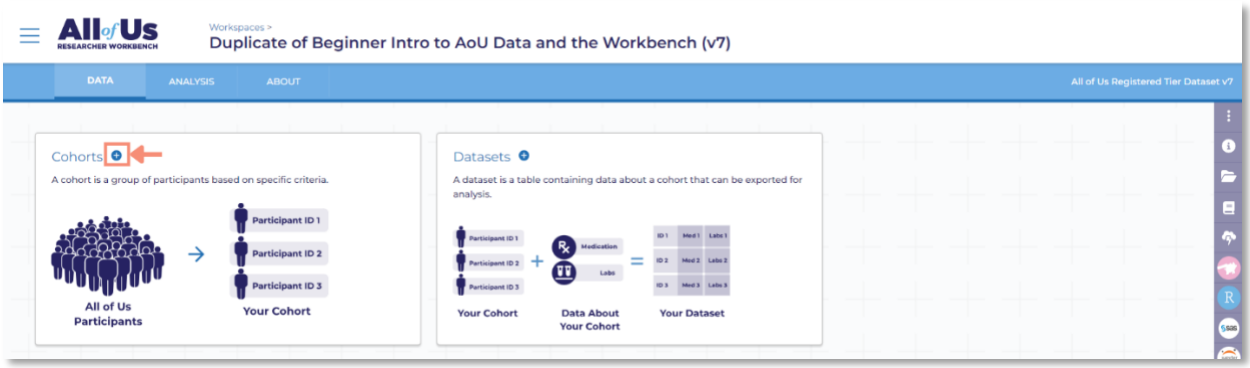
- Establish clear access, roles, and stewardship up front: confirm every collaborator has the same data access tier and completed required trainings and agreements before adding them, assign reader, writer, and owner roles deliberately (grant owner only when necessary), and apply the principle of least privilege so users have only the access they need.
- Coordinate billing responsibilities and starter-credit use with the team and avoid leaving compute environments running under another person's account. If credits are exhausted, duplicate and transfer ownership to a team member with remaining credits rather than continuing work on someone else's billing.
- Adopt consistent naming conventions and maintain a concise README in the About tab (purpose, key cohorts and concept sets, active analyses, owner/contact, and billing status) so teammates can find and resume work easily.
- Creating an output (e.g., dataset) in analysis code, and later referencing output, will not work for other users unless they run the same code to produce the output themselves or the output is saved to the workspace bucket. Move important outputs to the workspace bucket while deleting persistent disks after sessions to reduce costs and ensure team access to outputs created by certain sets of code that are needed in other sets of code.
- For major changes or ownership transfers, create a workspace snapshot by duplicating the workspace to preserve reproducibility and provide a clean handoff.

7. Cohort Builder

In the Data tab of your workspace, you have access to the **Cohort Builder**, the first point-and-click tool on the Workbench to choose data for your research project.

With the Cohort Builder (**Figure 7-1**), you can include and exclude participants to build a cohort for your research project. You can add criteria based on the program data available in the selected data tier (Registered or Controlled) for your workspace.

Figure 7-1. Cohort Builder Creation Button in a Workspace



7.1 Creating a Cohort

For more detailed instructions on selecting participants in the Cohort Builder, please visit **Selecting Participants Using the Cohort Builder** or access the **Introduction to the Cohort Builder and Dataset Builder (Video and PowerPoint)**. **Table 7-1** provides video quick links for different aspects of the Cohort Builder.

Table 7-1. Cohort Builder Video Quick Links

Aspect	Description	Timestamp*
Data Overview	Understand the types of data available in the Workbench, including EHR, survey data, physical measurements, wearable devices, and genomic data.	2:49
Cloud-Based Repository	Data reside in a secure Google Cloud Platform (GCP), the CDR, accessible through BigQuery and point-and-click tools.	3:19
Pulling Data Options	Use point-and-click tools within the Cohort Builder or SQL queries to pull data. Point-and-click simplifies the process for non-technical users, while SQL queries offer flexibility.	3:51
Cohort Builder	Define inclusion and exclusion criteria to select participants based on research goals.	6:18
Temporal Feature	Define relationships between clinical events over time, such as one event occurring before or after another.	7:25
Apply Modifiers	Refine cohorts by applying demographic filters, event occurrences, or specific conditions like lab values or visit types	9:00
Variant Search	Filter participants based on specific genetic variants, including gene names, consequences, or ClinVar significance.	9:37

Aspect	Description	Timestamp*
Cohort Review	Validate and review up to 10,000 participant records to ensure alignment with your inclusion/exclusion criteria before exporting.	10:13
Live Demo	This live demonstration is a walkthrough for creating a cohort using the Cohort Builder.	15:23

*Link to full video: **Introduction to the Cohort Builder and Dataset Builder (Video and PowerPoint)**

Table 7-2 outlines aspects to consider when using the Cohort Builder.

Table 7-2. Cohort Builder Considerations

Aspect	Key Considerations
Logical Operators	Use **AND** to narrow the focus by meeting multiple criteria or **OR** to broaden inclusion criteria.
Group vs. Total Counts	Understand that **group** counts reflect individual criteria, whereas **total** counts represent all criteria combined.
Concept Relationships	Use the **Hierarchy Viewer** to distinguish between broader (parent) and specific (child) concepts.
Criteria Overlap	Avoid unintentional overlap in **inclusion** and **exclusion** criteria, which may reduce participant counts.
Cohort Documentation	Assign clear names and descriptions to cohorts for easier tracking and collaboration.
Exclusion Criteria	Prioritize exclusion criteria carefully, as they take precedence and may drastically reduce counts.
Modifiers Impact	Test the effect of demographic and temporal modifiers on cohort size iteratively.
Reusability	Save and edit cohorts for future use without starting from scratch.

After selecting your participants with the Cohort Builder, you can create another cohort or create a cohort review set (optional). For more detailed instructions on reviewing participants in your cohort, please visit **Reviewing Participants with the Cohort Review**. Once your cohorts are finalized, you can move forward and create a dataset.

7.2 Using the Temporal Feature

The **Temporal Feature** in the Cohort Builder enables you to define inclusion or exclusion criteria based on the timing of clinical events—for example, whether they occur before, after, within, or during the same time range as another event. It also allows you to account for when a clinical event appears in the data, such as any mention, the first mention, or the last mention. **Table 7-3** outlines aspects to consider when using the Temporal Feature in the Cohort Builder.

For more detailed instructions on using the Temporal Feature, please visit **Using the Temporal Feature within the Cohort Builder** or access the **Temporal Feature in the Cohort Builder (video)**.

Table 7-3. Temporal Feature Considerations

Aspect	Considerations
Purpose of the Temporal Feature	The Temporal Feature is used to define relationships between clinical events based on their timing (e.g., before, during, or after).

Aspect	Considerations
Applicable Data	Temporal criteria are primarily applicable to EHR data.
Defining Clinical Events	Use to define temporal relationships between two or more clinical events, such as diagnoses, procedures, or treatments.
Event Timing Options	Include before, during, or after as timing options to establish the temporal relationship.
Practical Use Cases	Example: Identify participants with a specific diagnosis before receiving a treatment within a defined time frame.
Modifiers with Temporal Feature	Use modifiers, such as occurrence counts or specific demographic filters, to refine temporal criteria further.
Outcome of Use	The Temporal Feature identifies participant IDs that meet the specified criteria but does not pull raw data or detailed event records.
Limitations	The Temporal Feature does not directly provide data for event details and requires additional wrangling in the dataset or analysis environment.

7.3 Using the Variant Search Feature

The **Variant Search Feature** in the Cohort Builder allows you to filter *All of Us* genomic data by gene, consequence, ClinVar significance filter, allele count, allele number, and allele frequency. You can also sort by participant count and add all participants with structural variants to your inclusion criteria. **Table 7-4** outlines aspects to consider when using the Variant Search Feature in the Cohort Builder.

Table 7-4. Variant Search Feature Considerations

Aspect	Considerations
Purpose of the Variant Search	Use to filter participants based on specific genetic variants within their genomic data.
Applicable Data	The Variant Search Feature is primarily applicable to genomic data, including short-read genomic sequences and structural variants.
Search Filters	Use to filter by specific criteria, such as gene name, variant consequence, allele frequency, or ClinVar significance.
Gene-Specific Filtering	Select participants with specific gene variants of interest, aiding targeted genomic studies.
Data Preview	Preview participant counts meeting the variant search criteria for real-time feedback.
Integration with Cohort Builder	Combine results from the variant search with other inclusion/exclusion criteria to refine participant selection.
Practical Use Cases	Example: Identify participants with a rare variant in a gene linked to a disease or condition of interest.
Limitations	Variant search focuses on high-level filtering; additional data analysis is required to extract specific variant details.

For more detailed instructions on using the Variant Search Feature, please visit **Using the Variant Search Feature within the Cohort Builder** or access the **Variant Search Function within the Cohort Builder (Video)**.

Key Takeaways: Using the Cohort Builder

- Use logical operators, modifiers, the Temporal Feature, and the Variant Search Feature within the Cohort Builder for precise cohort creation tailored to research needs.
- Use the Cohort Review Feature to validate cohorts with up to 10,000 participant reviews and assign clear names/descriptions for easy tracking and reuse.
- The point-and-click nature of the Cohort Builder simplifies access for all users, while SQL and the Hierarchy Viewer enable advanced customization.

8. Dataset Builder

In the Data tab of your workspace, you have access to the **Dataset Builder**, the second point-and-click tool on the Workbench to choose data for your research project.

Using the Dataset Builder, you can choose **concept sets** of interest and related **values** for your cohort of participants to create and finalize a dataset. The following sections review all of these features in detail.

8.1 Concept Sets and Values

A **concept set** describes information in a patient's medical record, such as a condition, a prescription they are taking, or their vital signs. Subject areas, such as conditions, medications, and physical measurements, 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. Concept sets are valuable for organizing and viewing key variables, such as conditions, medications, and physical measurements within a dataset.

Values are the specific information related to your selected concept sets that you would want to include in your dataset. Examples of values include a survey question, the participant's answer to a survey question, the date the participant took the survey, the participant's identification number, and more. **Table 8-1** outlines aspects to consider when using concept sets.

For a general overview on selecting concept sets and values within the Dataset Builder, please refer to **Dataset Builder and Concept Sets**.

Table 8-1. Concept Sets Considerations

Aspect	Considerations
Purpose of Concept Sets	Use to define the variables and values (e.g., conditions, labs) to be included in your dataset.
Prepackaged Options	Use prepackaged concept sets for quick access to frequently used data points.
Custom Selection	Use the concept set selector to manually search for and add concepts tailored to your research goals.
Domain-Specific Searches	Search concepts within specific data domains (e.g., EHR, survey, genomic) to refine your selections.
Hierarchical Relationships	Understand parent-child relationships between concepts to include the appropriate level of detail.
Concept Combinations	Combine multiple concept sets (e.g., conditions and lab measurements) for comprehensive datasets.
Editing and Reuse	Save and edit concept sets for reuse in future projects, saving time on subsequent analyses.

Two Quick Cautions about Survey Data

- Survey data will always be in "long" format; each observation represents a survey response for a particular respondent. These data will likely need to be transformed to "wide" format before they can be merged with other data types.
- Before using survey variables in analysis, you may want to read the original question wording, response options, and any skip/branching logic using the Survey Explorer. This will help you understand exactly what each value represents. Special attention should be paid to how a zero value is differentiated from a missing value. Don't assume a numeric "0" is equivalent to "no" or that a blank/null cell means the same thing as a coded "missing" value. In your analysis and recoding phase, always check frequency tables and the value labels to distinguish true zeros (a measured zero outcome) from nonresponse or inapplicability, document any recoding rules you apply, and be cautious about imputing or treating missing values as zeros without a defensible, documented rationale.

8.2 Creating a Dataset

Building your **dataset** involves three key steps. Follow this three-step process to create a dataset:

1. Select your cohort (your participants) as created in the Cohort Builder.
2. Select your concept set (your rows) (e.g., conditions, measurements).
3. Select your values (your columns) for analysis.

For more detailed instructions on selecting participants in the Cohort Builder, please visit **Building a Dataset with the Dataset Builder** or access **Introduction to the Cohort Builder and Dataset Builder (Video and PowerPoint)**.

These three steps are easily identified and represented as columns within the Dataset Builder (see **Figure 8-1**). **Table 8-2** provides video quick links for different aspects of the Dataset Builder, and **Table 8-3** outlines aspects to consider when using the Dataset Builder.

Figure 8-1. The Dataset Builder

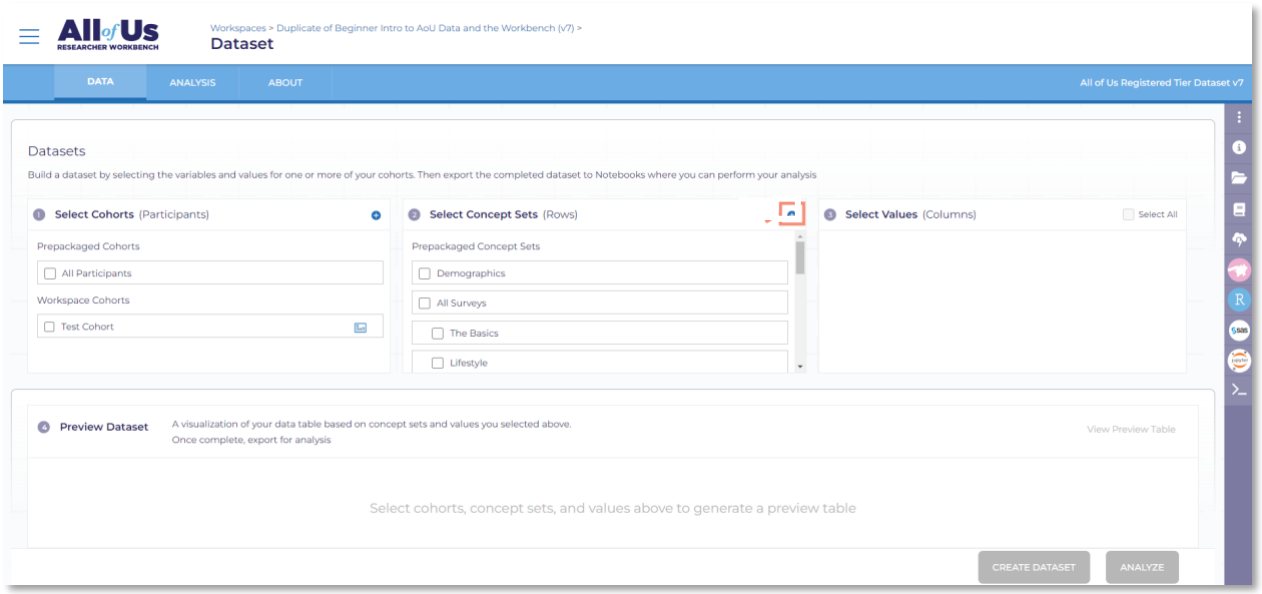


Table 8-2. Dataset Builder Video Quick Links

Aspect	Description	Timestamp*
Purpose of the Dataset Builder	Use this to select variables and values of interest (e.g., lab results, demographic data) for analysis, aligned with cohorts created in the Cohort Builder.	11:13
Three-Step Process	Follow the three-step process to create database: (1) select cohorts created in the Cohort Builder, (2) select relevant concept sets (e.g., conditions, measurements), and (3) select values for analysis.	11:57
Concept Selection	Use prepackaged concept sets or create custom concept sets to match research needs.	12:48
Data Preview	Review included variables and adjust selections to exclude unnecessary fields before creating the dataset to optimize size and clarity.	10:13
Export Options	Export datasets to Jupyter Notebook or other environments (e.g., R Studio, SAS) for detailed analysis.	27:45
Efficiency Tips	Exclude redundant or unused fields to streamline the dataset and reduce processing time during analysis.	13:36
Support Resources	Access help via the User Support Hub and Help Desk for troubleshooting or guidance with the Dataset Builder.	29:29
Live Demo	View live demonstration and walkthrough creating a cohort using the Dataset Builder.	23:14

*Link to video: **Introduction to the Cohort Builder and Dataset Builder (video and PowerPoint)**

Table 8-3. Dataset Builder Considerations

Aspect	Considerations
Data Domains	Accessing data from various domains, including EHR, physical measurements, and genomic data, is dependent on the data access tier you selected for your workspace.
Data Preview	Use the preview tool to validate selected variables and exclude irrelevant fields before exporting the dataset.
Column Field Selection	Carefully review field descriptions to avoid unnecessary data and streamline the dataset.
Export Options	Export datasets in formats compatible with Jupyter Notebook, SAS, and R Studio for analysis.
Dataset Optimization	Exclude unused variables to reduce dataset size and enhance processing efficiency.

Key Takeaways: Using the Dataset Builder

- The Dataset Builder follows three key steps—selecting a cohort, selecting concept sets, and selecting values—ensuring a systematic approach to building research datasets.
- Concept sets group related medical information, while values provide specific data points, allowing researchers to organize and extract relevant insights effectively.
- Researchers can preview selections, exclude unnecessary fields, and export datasets in compatible formats (e.g., Jupyter Notebook, R Studio, SAS) for streamlined analysis.

9. Analysis and Workflow Tools

The Workbench supports a variety of cloud-based applications to analyze *All of Us* participant data (e.g., Jupyter Notebook, RStudio, and SAS Studio) and to manage workflows (Cromwell and Nextflow).

These applications incur costs because they use cloud environments, which are virtual spaces that allow users to access and manage computing resources over the internet. To learn more about costs associated with your analysis, please visit **What to Know About Costs**. Because this is such an important topic, Section 10 of this manual is devoted entirely to billing in the Workbench.

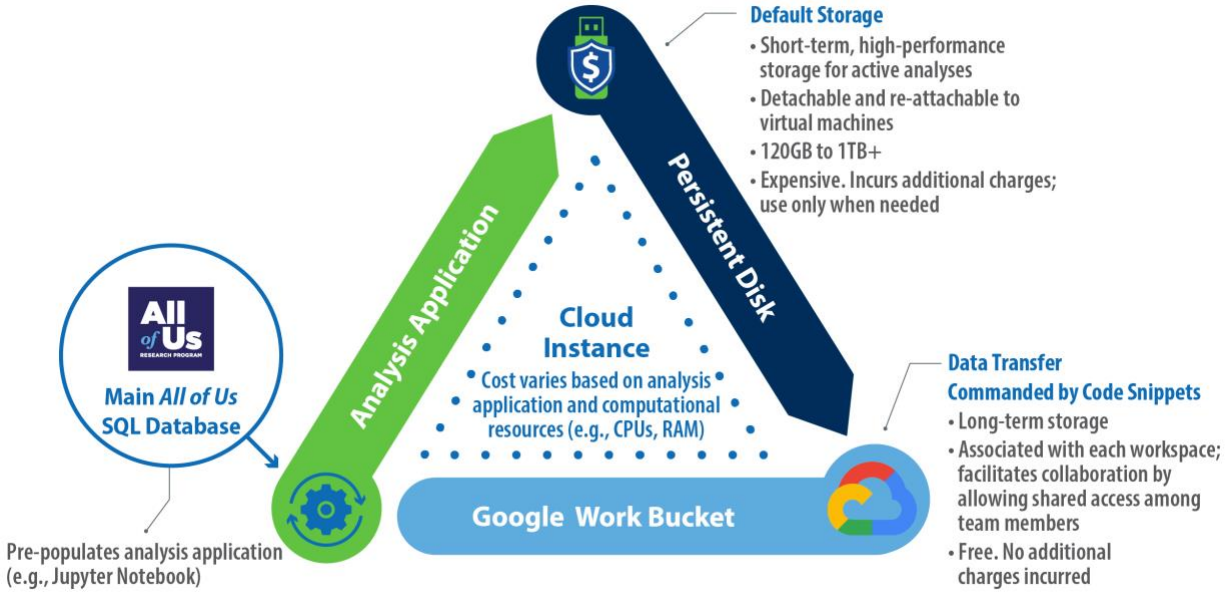
9.1 Workbench Data Flow

The dataflow from the main SQL database to being accessible to researchers can be confusing to follow and unclear, but this section helps to shed light on this process. **Figure 9-1** illustrates the standard flow from the SQL database to the user's analytical platform (e.g., Jupyter Notebook, RStudio). This is a two-part phase because the moment the data are downloaded to the platform, the data are actually downloaded to the persistent disk as well. The persistent disk is described in greater detail below. During the duration the platform is open, the persistent disk will also be open and in use, but when the platform and the instance is terminated, the persistent disk will still be present and available in the background with the downloaded and now edited data.

Key Takeaway: Cost and Storage Considerations

To reduce costs and prevent potential storage issues, be sure to move the data to the Google work bucket at the end of each cloud instance session and prior to logging out. This is not cold storage; therefore, your data can be easily accessible when you restart a new cloud compute instance.

Figure 9-1. Workbench Data Flow



9.2 Jupyter Notebook

Jupyter Notebook is a popular tool used by data scientists, researchers, and developers for interactive computing and data analysis. Jupyter Notebook provides an environment where you can write and execute code, visualize data, and document your work in a single document. Programming languages R or Python are available for Jupyter Notebook in the Workbench. **Table 9-1** summarizes the pros and cons of using Jupyter Notebook.

Table 9-1. Pros and Cons of Using Jupyter Notebook

Pros	Cons
Cost: Jupyter Notebook is currently the most cost-effective analysis tool on the Workbench.	**Performance:** Large datasets or complex computations can cause slower execution or crash the environment.
Multi-Language Support: Jupyter Notebook can run multiple programming languages, including Python and R.	**Truncation:** Depending on the dataset size, some datasets can be truncated making it challenging to observe the dataset fully.

The following resources are useful for analyzing data with Jupyter Notebook on the Workbench: Customizing Jupyter Notebook Environments and Code Snippets in Jupyter Notebooks.

9.3 RStudio

RStudio is an integrated development environment for researchers using programming language R for statistical computing and graphics. RStudio provides a user-friendly interface that makes it easier for users to write, debug, and execute R code. Programming language R is

available for RStudio in the Workbench. **Table 9-2** summarizes the pros and cons of using RStudio.

Table 9-2. Pros and Cons of Using RStudio

Pros	Cons
User-Friendly Interface: RStudio provides an interface with multiple panes for script editing, console output, environment variables, and file management.	**Memory Intensive:** R operates in memory; therefore, handling big data directly in RStudio can be a challenge if the machine is not equipped with enough computer power and storage.
Data Visualization: RStudio is excellent for data visualizations with integrated support from libraries to engage in real-time plotting, plot customization, and high-quality exports.	**Cost:** Using RStudio in the Workbench is more expensive than using the Jupyter Notebook.
Cross-Platform: RStudio works across different operating systems.	

Visit **Using RStudio on the Workbench** for more detailed guidance on using RStudio in the Workbench.

The following are additional useful resources for learning and analyzing data with R: **R Project Website**, **The Epidemiologist R Handbook**, and **CRAN: Manuals**.

9.4 SAS Studio

SAS is also available to researchers on the Workbench in the form of SAS Viya, known as **SAS Studio**. SAS Studio is a statistical software tool used by researchers for data analysis and statistical modeling. SAS Studio can handle large datasets efficiently with an extensive set of statistical techniques and a user-friendly interface. To access the Workbench data in SAS, you will need to create a dataset in the Dataset Builder and then copy the auto-generated SQL code into SAS Studio. **Table 9-3** summarizes the pros and cons of using SAS Studio.

Table 9-3. Pros and Cons of Using SAS Studio

Pros	Cons
Tools: The point-and-click interface, SAS Tasks, and Snippets make SAS more accessible to researchers with limited coding experience.	**Higher Cost:** SAS is a licensed software and is more expensive than open-source alternatives, like R or Python.
Support: SAS offers extensive documentation, tutorials, and support resources for users at all levels.	**Limited Flexibility:** Integration and customization options are more restrictive compared with open-source tools.

Visit **Exploring *All of Us* data using SAS Studio** for more detailed guidance on using SAS Studio in the Workbench.

9.5 Workflow Engines: Cromwell and Nextflow

The Workbench supports two workflow engines: **Cromwell** and **Nextflow**.

- **Cromwell** is a workflow management system that is designed to help researchers organize and execute complex computational workflows. Cromwell provides a platform for defining, running, and monitoring workflows, making it easier to manage and automate scientific analyses. The Cromwell application can be readily found in the workspace analysis panel.
- **Nextflow** is a workflow engine that uses a DSL (Groovy with workflow-specific extensions).

Table 9-4 summarizes the pros and cons of Cromwell and Nextflow.

Table 9-4. Pros and Cons of Cromwell and Nextflow

Pros	Cons
Organization: Cromwell and Nextflow are great tools for repetitive tasks and reproducibility for large and complex pipelines that will be used multiple times in the Workbench.	**Syntax Learning:** Cromwell and Nextflow have their own syntaxes and could be difficult to learn.
Reproducibility: Cromwell and Nextflow can be easily dockerized[a] and transferred to different workspaces, further increasing reproducibility.	

[a] Dockerizing an application means to create a Docker container for your application. This container will then encapsulate the application's code, dependencies, and runtime environment, making it easy to deploy, scale, and run on any platform that supports Docker.

- Visit **Workflows in the *All of Us* Workbench: Nextflow and Cromwell** for general information about how to use batch organizers and application programing interfaces.
- Visit **How to Use Cromwell in the *All of Us* Workbench** for more information on Cromwell.
- Visit **Using Docker Images in the Workbench** for information on using Docker images in the Workbench.

9.6 Cloud Environments and Storage Options

Figure 9-2 illustrates the flow of information within the Workbench. We describe each step below.

1. The main cloud server for the Workbench is owned by Google and runs under the Google Terra software. Given the sensitivity of the *All of Us* data, only researchers with access to the database through a server connection can view and analyze the data. Cloud environment access and usage are needed to interact with the data. The cloud storage options within the Workbench consist of short- and long-term storage.
2. For short-term storage and easy accessibility, a **persistent disk** is a reliable, high-performance block storage for virtual machine instances. Like a USB drive, a persistent disk can be detached from the virtual machine upon deletion and reattached to a new one, allowing for files to be stored permanently. The storage amount ranges from

120GBs to 1TB+. ***A persistent disk can be extremely expensive to have and leave open, which is why it is typically used for short-term storage.***

3. For all other storage needs, use a personalized Google cloud bucket designated to all Workbench users. The maximum storage amount for the bucket is not known but estimated to be 5 TBs (the maximum storage for an average Google bucket). There is no additional cost to use and store information in the Google bucket.

Visit **Cloud Environments and Storage Options** to learn more about cloud environments and storage options.

Figure 9-2. Flow of Information within the Workbench

(A)	(B)	(C)
The main cloud server runs under the Google Terra software. The application you choose, whether it is Rstudio or Jupyter Notebook, will need to be initially pulled from the main database in which the data are stored.	For short-term storage and easy accessibility, a persistent disk is a reliable, high-performance block storage for virtual machine instances. This can be expensive if not monitored.	For all other storage needs, use a Google cloud bucket designated to all Workbench users. There is no additional cost compared to the persistent disk.

10. Billing in the Workbench

Billing in the Workbench is based on three criteria: (1) storage type, (2) computational configuration of cloud instance, and (3) data amount. Among these three criteria, the cloud instance is directly associated with billing cost. The selection of the analysis application (e.g., Jupyter Notebook, R Studio, SAS Studio) used for cloud computing dictates the cost because of potential resource allocation. It is important to have a general idea of the necessary cost of each project and actively monitor the credit expense in real time throughout the project. This is beneficial for reassessing cost and determining how the expenses are calculated.

Each Workbench account **begins with $300 in initial credits** for research analysis activities. **Initial credits will expire 365 days** after registering for the Workbench. Researchers can check how much of the initial credit balance has been spent at any time by visiting their Workbench profile page. For detailed instructions, visit **Using *All of Us* Initial Credits –User Support**. An active billing account is needed to cover any computational and storage costs beyond the initial $300 in credits.

Trainees also have access to $300 in promotional credits from Google if they have not already used the credits from another Google Cloud Platform (GCP) product. This User Support Hub article summarizes how to set up a GCP account. The GCP credits last for 90 days once they have been activated.

Visit **Getting Started and What to Know About Costs**, **Billing in the Workbench**, and **Paying for Your Research** to learn more about billing in the Workbench.

10.1 Optimizing Billing and Budgeting in the *All of Us* Workbench (Terra)

Efficient cost management is critical when working with cloud-based platforms like Terra, which powers the Workbench. This section presents three common billing pitfalls and clear strategies to help researchers avoid unnecessary spending while maximizing the platform's capabilities.

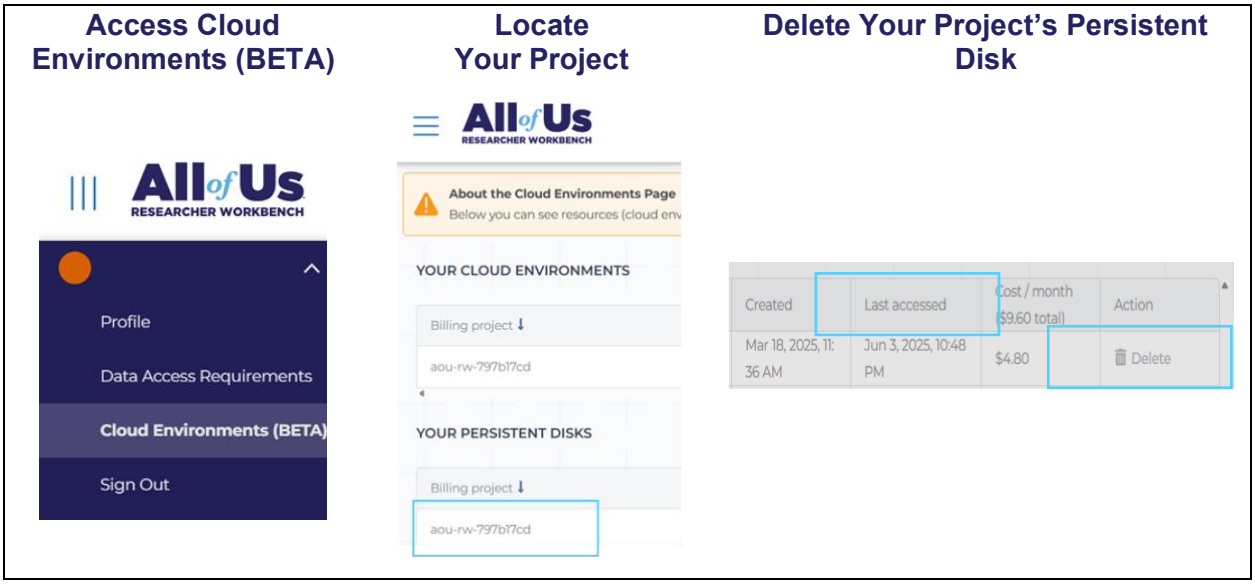
Scenario 1: Persistent Disk Over-Spending

Issue:
A researcher is storing all project data on their notebook's persistent disk, which accrues costs even when not actively used. Over time, this leads to bloated storage charges, especially across multiple projects or users.

Solution:
Take the following steps to reduce costs:

- Move all important data and output files to your **Google Cloud Storage bucket** within your Terra workspace.
- Once your work is saved, **delete your persistent disk** after each session (see **Figure 10-1**).
- Re-create a new persistent disk the next time you start a session. It is free to create and helps ensure you are only billed for storage when actively needed.

Figure 10-1. How to Delete Persistent Disk



For additional information on deleting your project's persist disk and alternative methods, visit **Persistent Disk: Managing and Deleting**.

Pro Tip

You can automate this step by adjusting your code so that at the end of each session your data are pushed to permanent storage. This will allow you to confidently delete the persistent disk after every use without fear of losing information and data.

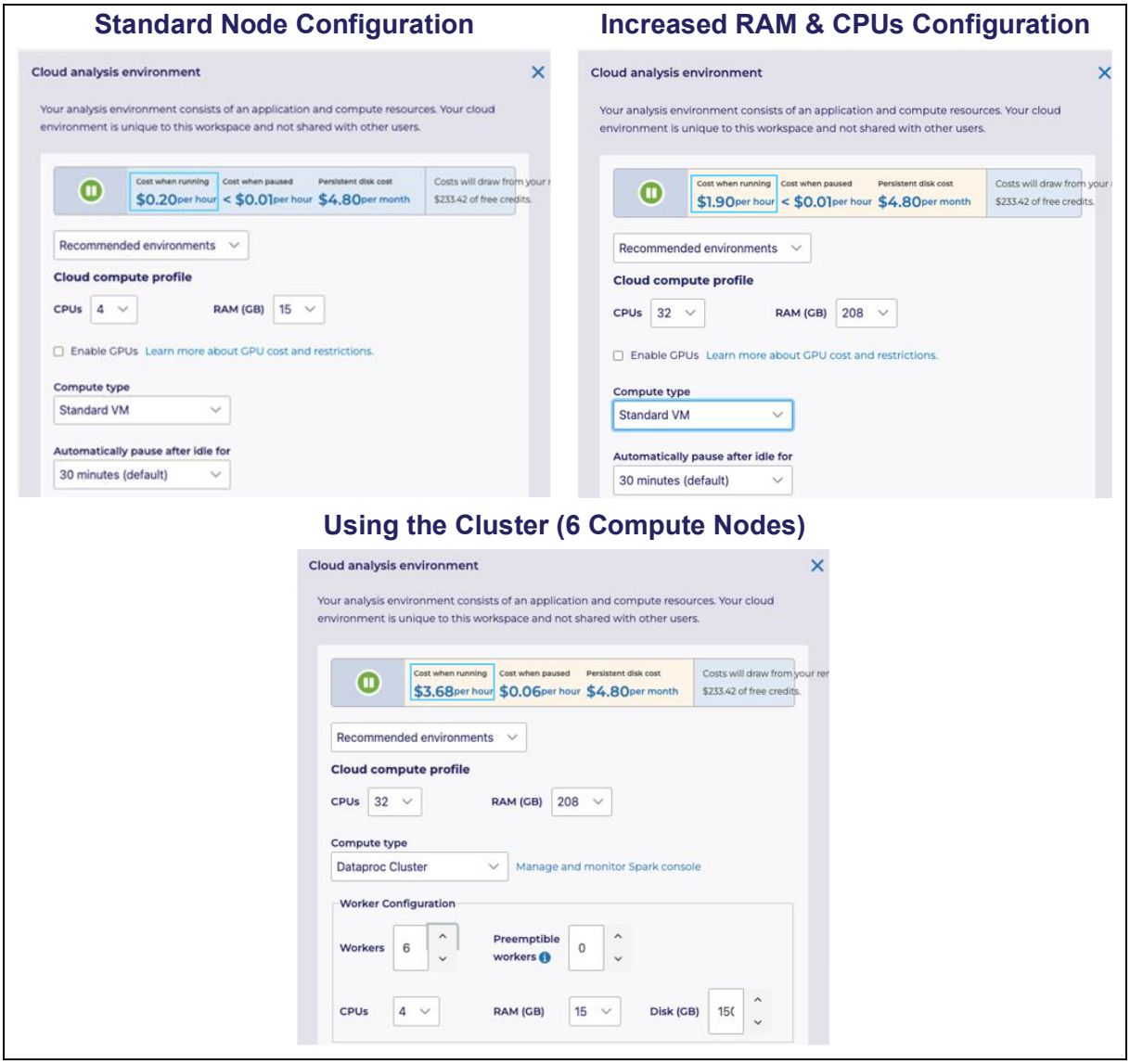
Scenario 2: Inefficient Cloud Compute Configurations

Issue:
A researcher runs their analyses on the default compute configuration without adjusting the CPU, memory, or disk size—even when the job does not require high resources. This leads to over-provisioning and higher costs for tasks that could run on smaller machines.

Solution:
Take the following steps when launching a compute environment:

- **Right-size your resources**: Choose fewer CPUs or reduce RAM if you are working with smaller datasets or light processing tasks (see **Figure 10-2**).
- Use the **"Customize" option** under environment settings to match your compute to the specific demands of the task.
- Monitor your performance; if jobs consistently finish quickly or leave resources underutilized, scale down next time.

Figure 10-2. Configuration Examples



Note: **"Cost when running"** amount will adjust accordingly as you alter cloud analysis environment configuration. This can be useful when estimating how much your analysis will cost using a particular cloud analysis environment configuration.

For additional information on configuring your cloud environments and storage options, visit [Cloud Environments and Storage Options (video and PPT slides)].

Pro Tip

For simple exploratory analyses or visualizations, a 4CPU/12GB memory setup is usually sufficient. For genomic data analyses, it is important to test out the suggested analysis on a small subset of data first to optimize your compute configuration to what will be needed for the main dataset.

Scenario 3: Leaving Compute Instances Running

Issue:
A user finishes a coding session but forgets to shut down the compute instance. Because Terra continues to bill for running environments—even idle ones—this oversight can quietly drain your billing account.

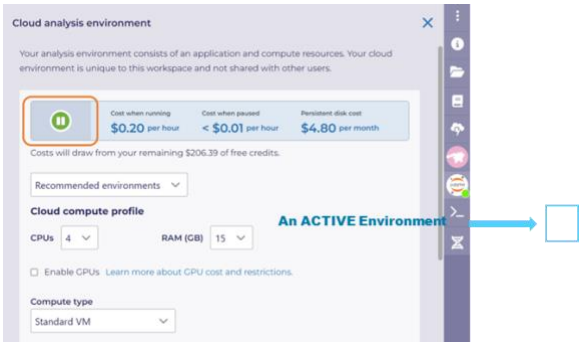
Solution:

- Always **pause or shut** down your cloud environment when you are done working (see **Figure 10-3**). This step will immediately stop additional compute charges.
- Enable auto-pause settings to automatically suspend idle notebooks after a set time of inactivity (e.g., 30 minutes).
- Get into the habit of logging out or disconnecting from the Workbench when you are done with each session.

Figure 10-3. Pausing and Restarting an Environment

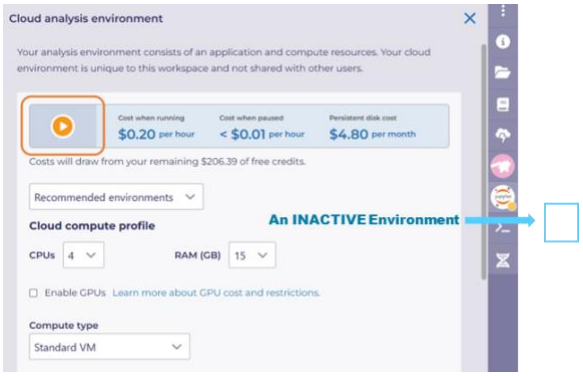
Pausing an Environment

Active environments can be paused by clicking the green "pause" sign in the Jupyter environment panel.



Restarting a Paused Environment

Paused environments can be restarted at any time by clicking the orange "run" button.



For additional detailed information on customizing and optimizing your cloud analysis environments and storage options, visit **Using, Customizing, and Optimizing Jupyter Cloud Environments**.

Pro Tip

Look for the **green dot** in the upper-right of the Jupyter environment—if it is active, you are still being charged!

Proactive cost management is part of being an effective and responsible researcher on the cloud, so teaching other researchers about these tips is important for their success. These small adjustments not only protect your budget but also help ensure accessibility and sustainability across the *All of Us* community.

11. Planning Your Data Analysis on the Workbench

Starting an analysis on the Workbench can be stressful and potentially overwhelming, but we offer multiple resources and examples to help navigate this task, as detailed below. The key to completing a project successfully in the Workbench is knowing what data are available and how to use the data effectively to address your research question and test your hypothesis.

The *User Tips for Avoiding Common Workbench Challenges* resource (included as **Appendix A**) was designed to help Workbench users navigate the platform and avoid—or quickly resolve—some of the most commonly encountered challenges. These practical recommendations are designed to support you as you work, saving you time, reducing frustration, and enabling you to get the most value from the Workbench's features. We encourage you to review this resource.

11.1 Example Analysis from a Workbench Coach

In the PowerPoint slide deck, ***Steps for Success: Planning Your Data Analysis on the* All of Us *Workbench***, available on Amaze, a Workbench coach from the *All of Us* Researcher Academy shares key steps for planning and conducting an analysis in the Workbench, from research ideation all the way through the analysis phase. Researchers will learn how they can use publicly available *All of Us* resources to craft research questions and identify the key information required for an analysis plan.

The following sections provide helpful analysis tips and strategies from Workbench coaches.

11.1.1 Develop a Structured Analysis Plan Early

A structured analysis plan is especially important in helping to **manage Workbench credits effectively.** Each Workbench account begins with $300 in initial credits, and **initial credits expire 365 days after access to the Workbench is granted.**

Creating a well-structured analysis plan before using the Workbench is essential for efficient analysis and cost management. You should clearly define your research question, identify key variables, and verify data availability. Using tools such as the Data Browser and Research Projects Directory can help refine hypotheses and prevent duplicating existing research. Additionally, you should outline your study's methodological approach, including cohort selection criteria, statistical methods, and anticipated challenges. As described in the sidebar, a structured research plan is important in helping to manage Workbench credits effectively.

11.1.2 Understanding How Data Are Structured

Understanding how data are structured and accessed in the Workbench is key to efficient analysis. Data are stored in the cloud and *cannot* be downloaded or directly exported—only summary data can be shared. When building cohorts, concept sets, and datasets, SQL code is automatically generated to load data into your analysis environment. If you need to modify your dataset later, you can either

- remake and export your cohort, concept sets, or dataset to generate updated SQL code (recommended); or

- manually edit the SQL code within your environment (not advised, except for minor changes like adding a variable).

By understanding this workflow, researchers can better navigate data updates while maintaining reproducibility and compliance with *All of Us* data policies.

11.1.3 Optimize Cohort and Dataset Selection

Building a dataset efficiently requires careful selection of cohorts, concept sets, and values. Cohorts define the study population based on specific criteria, whereas concept sets represent key variables such as diagnoses, medications, and procedures. Values define the actual data points researchers want to analyze, such as patient ID, survey responses, or test results. Researchers can use **repackaged concept sets** for quick access to commonly used data points or create **custom concept sets** for a tailored approach. Before finalizing a dataset, it is recommended to use the **Data Preview** tool to exclude unnecessary fields and optimize dataset structure, reducing processing time and storage costs.

11.2 Dissemination Guidelines

The *All of Us* Research Program has established dissemination guidelines to protect participant privacy and ensure responsible data use. Key points include the following:

- **Prohibition on downloading individual-level data**: Researchers are *not* allowed to download or remove any participant-level data from the Workbench.
- **Restrictions on reporting small cell sizes**: When disseminating research findings, any data or aggregate statistics corresponding to fewer than 20 participants should *not* be published or distributed unless appropriate measures are taken to obscure these values using scientifically accepted strategies.

For comprehensive details, please refer to the following resources:

- **Data and Statistics Dissemination Policy**
- **Publication and Presentation Policy–User Support**
- **Overview of *All of Us* Research Program Policies for Researchers**

Adhering to these guidelines is *essential* for maintaining the integrity of the research and participant trust.

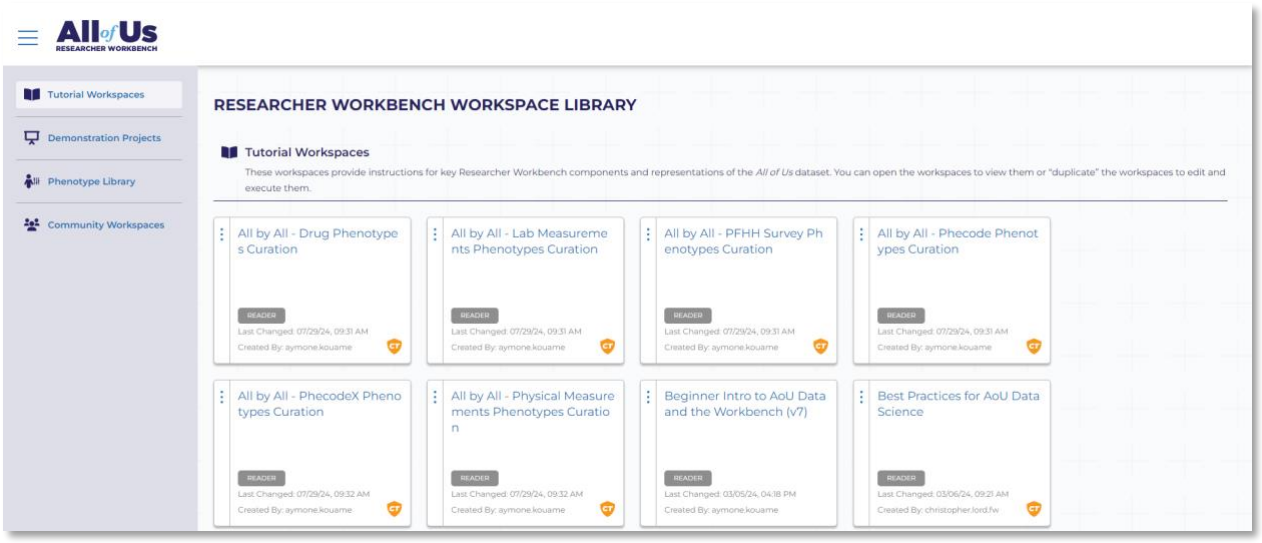
12. Resources and Support

The Workbench provides additional training resources and a User Support Hub to search, communicate, or troubleshoot issues within the platform. The following sections describe these features.

12.1 Featured Workspaces

Featured workspaces provide examples of data querying, wrangling, and analysis to support you when using the Workbench. As shown in **Figure 12-1**, there are four types of featured workspaces: **tutorial workspaces**, **demonstration projects**, **phenotype library**, and **community workspaces**.

Figure 12-1. Featured Workspaces on the Workbench



Note: Featured workspaces are **READ ONLY** by default unless you **duplicate** the featured workspace for your own use.

Tutorial workspaces provide a structured learning environment within the Workbench, guiding users from basic data manipulation to advanced analysis techniques specific to *All of Us* data. These workspaces are an essential starting point for researchers looking to explore and analyze data efficiently within the platform.

Demonstration projects highlight the quality, utility, and range of *All of Us* data by replicating end-to-end analyses from previously published studies within the Workbench. These projects provide valuable insights into the research potential of the platform while serving as practical examples for conducting robust analyses.

Phenotype library workspaces illustrate the implementation of computable electronic phenotypes within the *All of Us* dataset by applying previously published phenotype algorithms. These workspaces serve as practical examples for researchers looking to define and analyze phenotypes within the Workbench.

Community workspaces promote knowledge-sharing, collaboration, and learning by enabling registered Workbench users to share their workspaces with the broader research community. These shared spaces facilitate open exploration and collective advancement in data analysis within the Workbench.

To see a comprehensive list of these featured workspaces and gauge which ones may apply to your research objectives, please see the **Featured Workspaces** tab on your Workbench account or visit **Featured Workspaces**.

Amaze is the *All of Us* Researcher Academy's learning platform that hosts the self-guided courses, curriculum modules, recordings, templates, and a searchable course catalog. These resources and how to sign up to access them are described in full in Section 2 of this manual.

12.2 User Support Hub

The *All of Us* Research Program's User Support Hub offers guidance, tutorials, and educational materials on accessing and analyzing data within the Workbench. The User Support Hub provides technical assistance for troubleshooting, access to community forums and articles, and featured videos. Visit the **User Support Hub** to learn more.

12.3 Staying Current

Stay up to date on the dataset and the program by regularly checking the ***All of Us* Announcements and News page** and by building a few simple habits:

- Subscribe to the ***All of Us* Research Roundup** newsletter for the latest *All of Us* news, funding opportunities, and more.
- Review **CDR/Release Notes** whenever you start or resume a project
- Monitor **Research Highlights** for the most recent publications and examples on how *All of Us* is shaping how we understand health and disease.
- Visit the *All of Us* **Researcher Engagement Hub** for upcoming opportunities for *All of Us* researchers, including training, events, and other engagement activities for researchers at every skill level and career stage.

APPENDIX A:
User Tips for Avoiding Common Workbench Challenges

Train-the-Trainer Program

User Tips for Avoiding Common Workbench Challenges

The *All of Us* Trainer-the-Trainer Program developed this resource to help Researcher Workbench users navigate the platform and avoid—or quickly resolve—some of the most common encountered challenges. These practical recommendations are designed to support you as you work, saving you time, reducing frustration, and enabling you to get the most value from the Workbench's features. Additionally, the *Researchers' Guide to Using the Workbench* provides more in-depth information on these topics.

Progressing from the Cohort and Dataset Builders to Data Analysis (e.g., Jupiter Notebook)

The Cohort Builder and Dataset Builder are free to use, so it is important to take your time when working in these areas. It is better to take your time in the free sections of the Workbench than to waste your time in the costly sections. The more you understand and are able to pull from the Cohort Builder, the less money you will have to spend redoing filtering tasks in the Analysis section.

The **Cohort Builder** identifies and filters participants based on your criteria, but it only returns participant IDs—it does not include any of the underlying data about those individuals. As a result, users often filter for a specific group (e.g., females of a certain age with kidney disease) and later realize in the Dataset Builder that they only have IDs, but not the participant characteristics or condition data they intended to analyze. To avoid this, it is essential to clearly specify which data elements should be included for each participant when building your dataset.

To incorporate those data elements in the **Dataset Builder**, users must create concept sets defining the specific data they want to review (e.g., conditions, demographics, or survey information). The Cohort Builder selects *who* you are interested in, while the Dataset Builder and concept sets determine *what information is pulled for those participants*. These components work together.

If the user knows how to query data from SQL databases, the second option for pulling data elements would be to simply do it in the Analysis section. This is typically done by advanced or experienced coders or data scientists.

Before moving on to data analysis, take time to review and preview everything in the Dataset Builder to ensure all necessary information has been included in your concept set. Doing so ensures that the analysis and workflow tools pull the specific data you intend to analyze. Because these issues are more cumbersome to correct once you reach the Analysis section—where computing time is billed—reviewing and confirming these settings in advance improves your workflow, reduces the risk of unnecessary costs, and ensures you have the data needed for effective analysis.

Helpful Resource:

- Introduction to the Cohort Builder and Dataset Builder *(video and PowerPoint presentation)*

Moving Genomic Data

If your system request is to generate outputs from a large dataset (e.g., 50,000 participants), this request may continue running in the background and quickly deplete your account balance. Before starting any job, double-check that you have selected the correct data file (this is very important for genomic analyses) and have applied the appropriate filters. Without these specifications, the system may process a much larger dataset than intended.

For example, the Automated VCF generator—which pulls and combines VCF files automatically based on the participant IDs you query—can consume significant resources if parameters are not carefully set. To avoid unnecessary charges, confirm all settings in advance and allow adequate time for processing. Some data pulls can take several hours, so patience and careful preparation are essential.

Locating Data Files That You Have Created

When you create a data file, it is important to understand where your files are stored and how they move between systems. A common mistake that users make is to assume that data files are stored in the same location as your analysis code. However, because these tasks are handled in separate systems, the files are stored in different places.

Another common mistake understanding is that objects or datasets created in the Analysis space by one researcher will not be immediately available to other researchers with whom the workspace has been shared. To manage these files properly, you need to understand the file pathways and overall file system so you can locate and reuse your data without duplicating it. Otherwise, you may re-download data unnecessarily. Knowing where to find and manage your files will help you work more efficiently and avoid unnecessary costs.

Within the Workbench, two separate accounts communicate with one another: the Workspace and your Google Cloud account. When data are used, the platform moves data from the original dataset and stores it in your Google Cloud account. A copy of that data is then made available for use in your Workspace.

Within every *All of Us* account, you can view how files are organized in the Researcher Workbench directory. Each Workbench directory typically contains code files, such as Jupyter notebooks or R scripts. This directory does not store data files. Data are stored separately in your Google Cloud bucket.

To access your Google Cloud bucket, go to the **About** tab within your Workspace. At the bottom of the Workspace description, locate the **File management** link. Select this link to go directly to your Google Cloud files. From this file management directory, you can see created objects or datasets and share them with other researchers on the Workbench who have access to that space.

Alternatively, you can access the file management system directly in the Analysis workspace. Go to the **File** dropdown tab at the top left of the Analysis workspace and select **Open...**. This link will take you to the file management system within your Jupyter Notebook session. This is an easy way to find, move, and upload datasets or objects created in your session that can be shared with others.

Helpful Resource:

- Cloud Environments and Storage Options *(video and PowerPoint presentation)*

Coding Issues

When coding in the Workbench, the platform may, at times, present users with error messages. Your Trainer or other program support staff are available to help you work through coding issues and technical challenges as they arise. However, to enable the support team to troubleshoot efficiently and provide accurate guidance, it is helpful to include a few key details when submitting a request. Sharing the following information upfront helps streamline the support process and allows the team to resolve issues more quickly and effectively.

- **Explain the pseudocode, specify the coding language you are using, and clearly describe what you were trying to accomplish when writing the code.** This includes outlining your goal or intended outcome (e.g., pulling data for review) and explaining what you expected the code to achieve. Providing this context helps others better understand your approach and supports more effective troubleshooting and guidance.

- **Capture and share screenshots of any error messages you receive.** Screenshots will allow your Trainer or their support team to more clearly assess what is happening within the system.

> **Important Note!**
>
> When capturing an error message, **be sure to capture only the message itself and exclude any data shown on the screen from the image. You may never share screenshots that contain personal identifiable information or other sensitive data**, so avoid full-screen images that may inadvertently include such information and double-check all images before sharing them.

- **Explain what the dataset input looked like by describing, in general terms, the structure and categories of the data.** Because sensitive data cannot be reviewed directly, it is helpful to outline elements such as column titles or data types so that others can understand how the data are organized. Providing this high-level context will allow your Trainer or their support staff to assist with resolving the issue without accessing the actual data.

Optimizing Your Workbench Funds

Workbench users may often hesitate to run analyses because they feel everything must be perfect before starting. This hesitation can slow progress unnecessarily. If you complete the recommended pre-checks, you can feel confident moving forward. Any charges incurred should be relatively minimal when the proper preparation has been done.

Another common issue is leaving software applications running longer than intended. When you launch an application, you are given the option to either keep it running indefinitely or shut it off after a set period of time. The default setting is to shut off automatically. However, if an application is accidentally left running, it may remain active and continue to incur charges that draw down your funds.

To avoid this, make sure you deactivate applications when you are finished working for the day. To do this, navigate to the **Personal Information** section and then open the BETA CLOUD environment. There, you can see which applications are currently on, active, or paused and take action as needed.

Best practices to avoid unnecessary charges include the following:

- Save all work to your Google Cloud bucket during the analysis phase.
- Delete applications when you are no longer using them so that nothing remains running.
- Disconnect and delete the persistent disk when your work is complete.

When prompted, the system will ask whether you want to delete both the environment and the persistent disk. After confirming that your work has been saved, always select Yes to ensure no resources remain active and billable.

Helpful Resources:

- Using *All of Us* Initial Credits—User Support *(web page)*
- Getting Started and What to Know About Costs *(web page)*
- Billing in the Workbench *(video)*
- Paying for Your Research *(video)*
- Persistent Disk: Managing and Deleting *(web page)*
- Cloud Environments and Storage Options *(video and PowerPoint presentation)*