



Collaborative
Cohort of Cohorts
for COVID-19 Research

C4R Analysis Commons Training

Pallavi Balte, MBBS, PhD

Associate Research Scientist

Department of Medicine | Division of General Medicine

Columbia University Irving Medical Center, NY

August 11, 2021

Agenda

- Overview of C4R and BioData Catalyst/Seven Bridges (3:00 – 3:30pm)
 - Data availability on C4R Analysis Commons
 - Data harmonization in C4R
 - BioData Catalyst/Seven Bridges
 - How to get an account?
 - Organization and access to C4R Analysis Commons
 - Data security
- Live Demo in BioData Catalyst/Seven Bridges (3:30 – 4:00pm)
- Q and A session (4:00 – 5:00pm)

Target population (N=14 cohort studies)

Original cohort focus

Heart



Lung



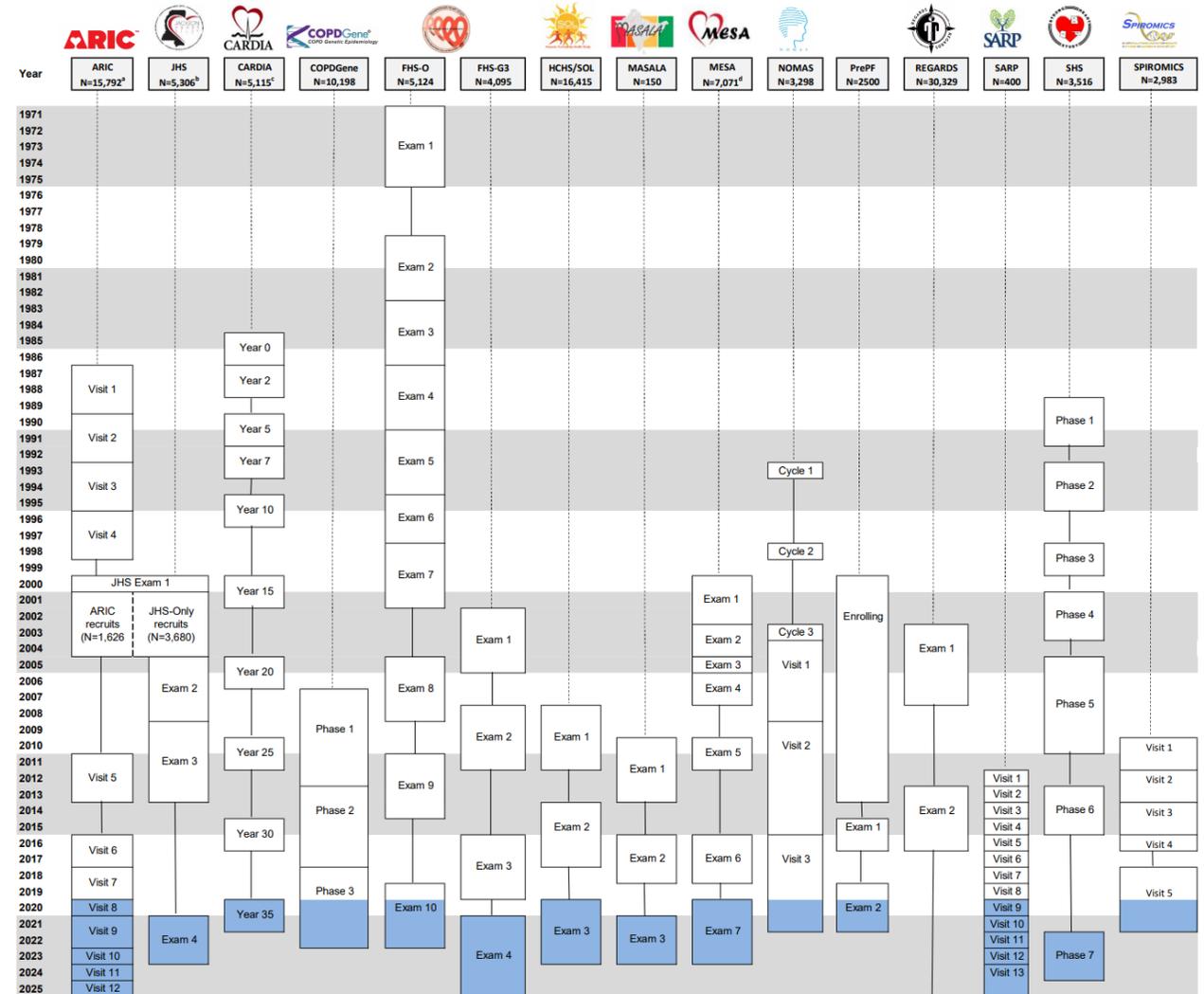
Brain



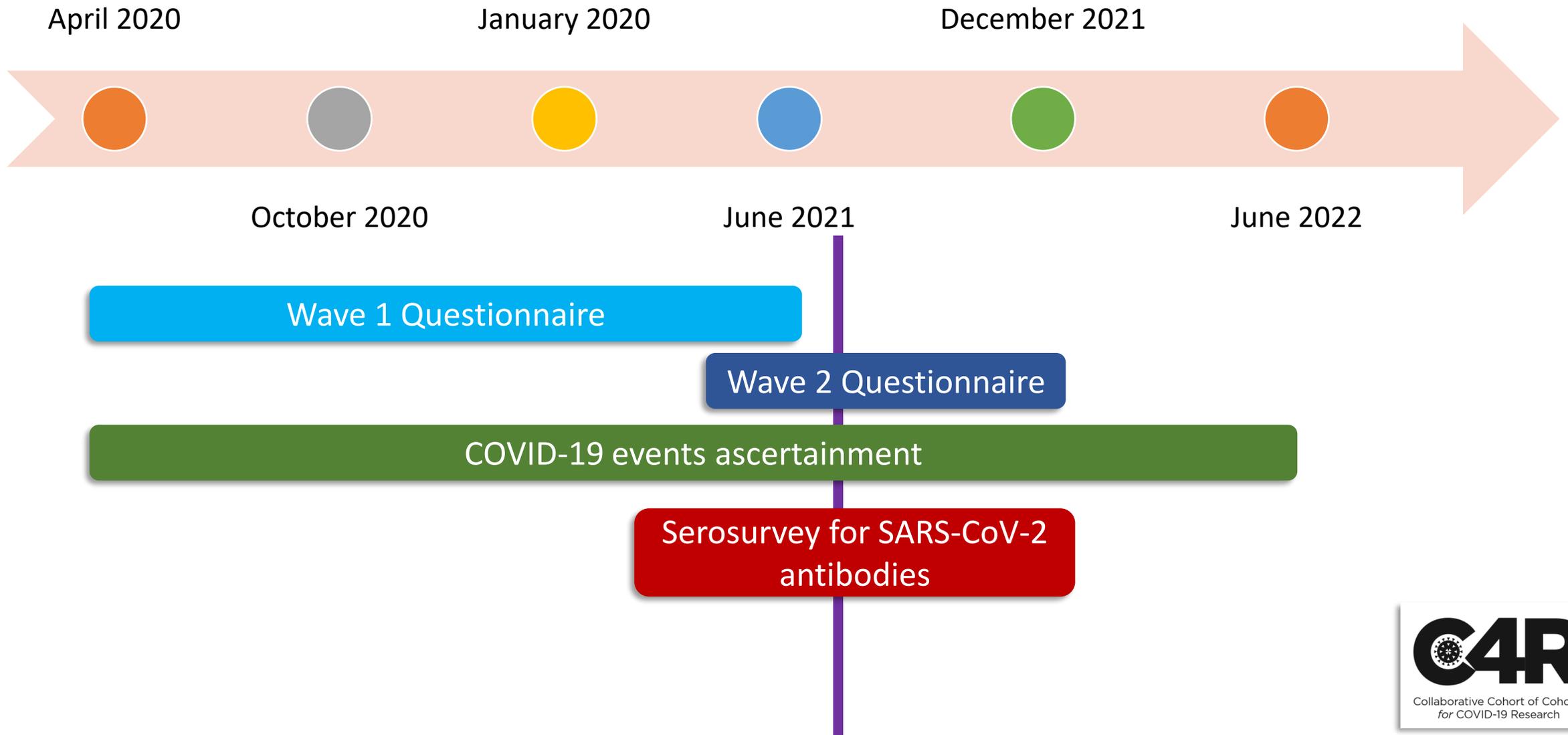
Cohort	N
ARIC	5046
CARDIA	4221
COPD Gene	4000
FIP	2500
Framingham	7258
HCHS-SOL	8400
JHS	2317
MASALA	500
MESA	4683
NOMAS	1267
REGARDS	8000
SARP	380
SPIROMICS	1800
SHS	2701
Total	53073

Over 1,000,000 person-years of follow-up

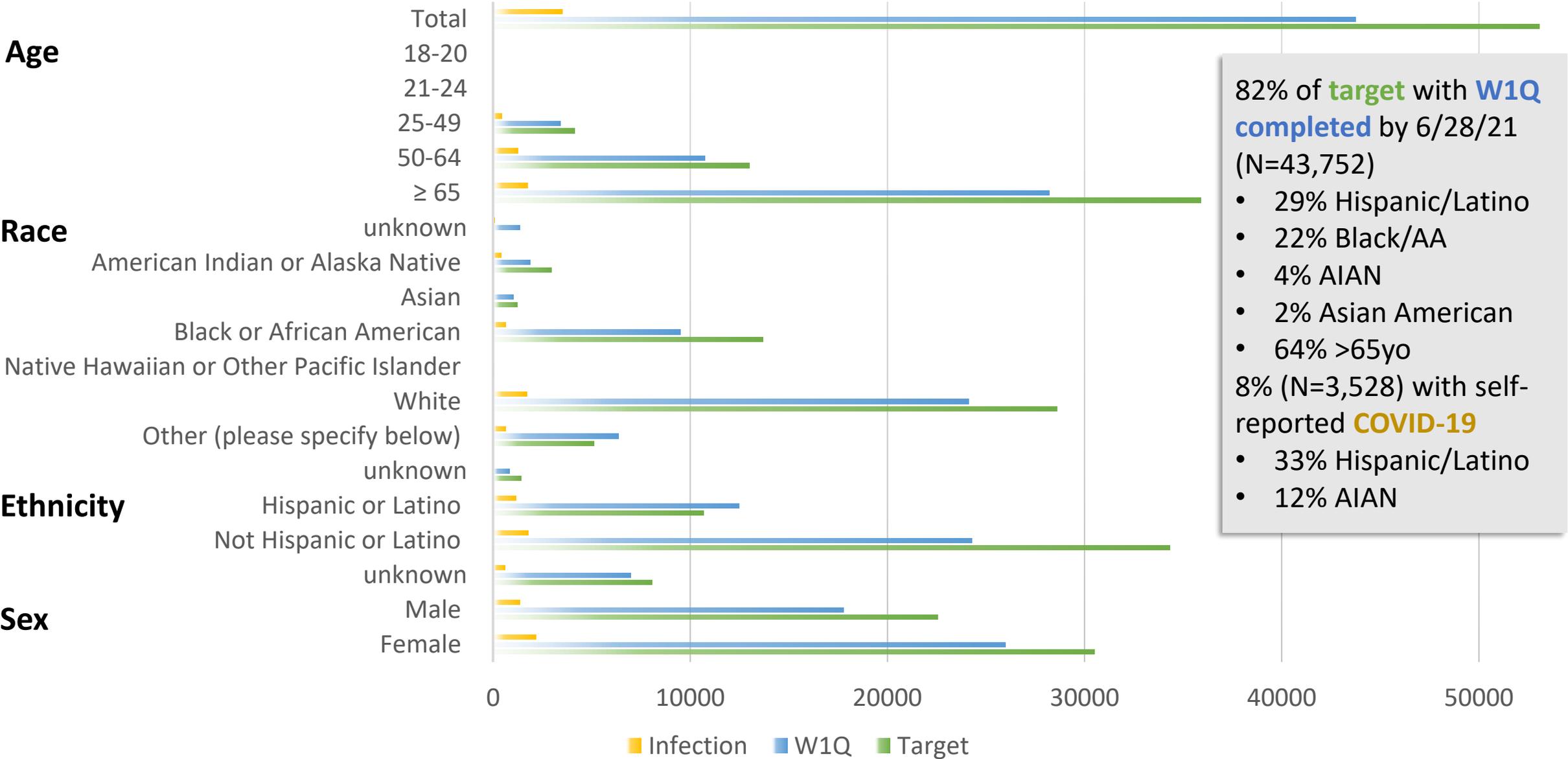
- Repeated measures across multiple organ systems, many biomarkers, and 'Omics
- Pandemic era exams already funded in numerous C4R cohorts



C4R Study Timeline



Wave 1 questionnaire completion (June 28, 2021)



82% of **target** with **W1Q completed** by 6/28/21 (N=43,752)

- 29% Hispanic/Latino
- 22% Black/AA
- 4% AIAN
- 2% Asian American
- 64% >65yo

8% (N=3,528) with self-reported **COVID-19**

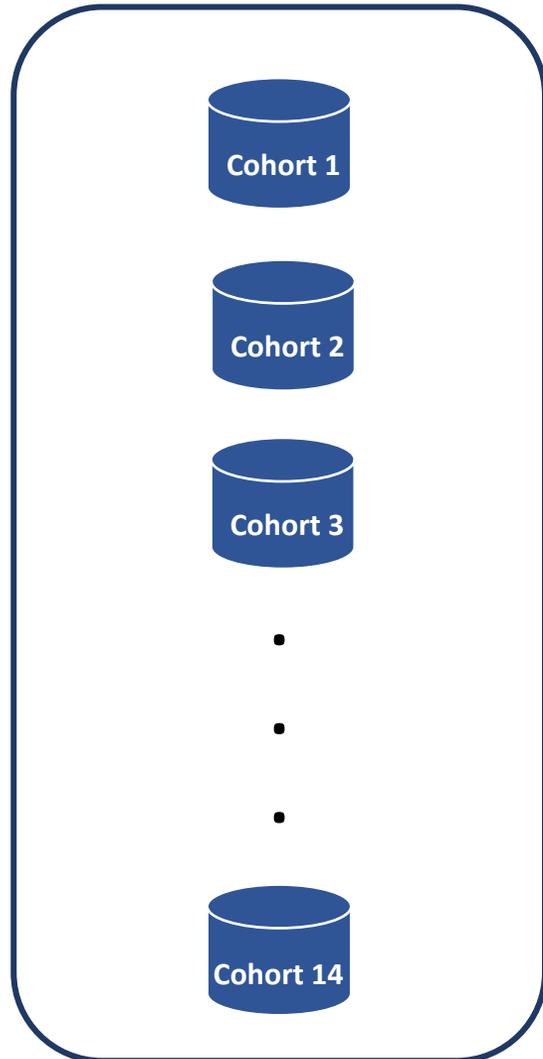
- 33% Hispanic/Latino
- 12% AIAN

C4R Data

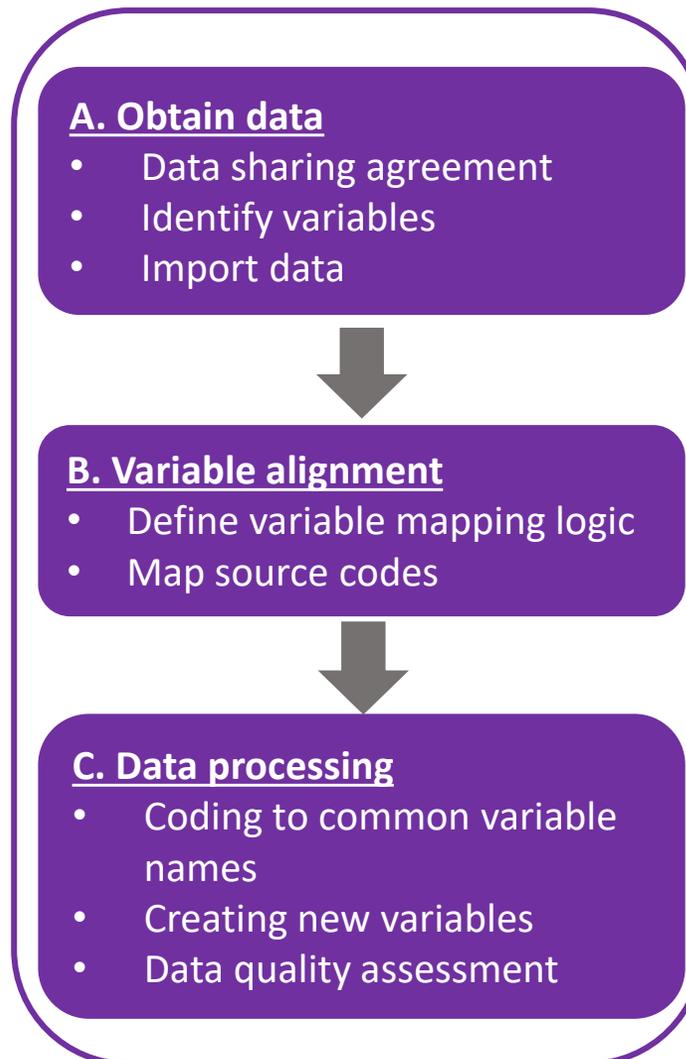
- COVID-19 Endpoints
 - Wave 1 Questionnaire (W1Q)
- Dried Blood Spots (DBS) data
- Adjudicated COVID-19 events
- Socio-demographics
 - Most recent pre-pandemic visit
 - Longitudinal data
- Pre-pandemic measures
 - Most recent pre-pandemic visit
 - Longitudinal data
- Adjudicated cardiac and pulmonary events

Data harmonization within C4R Analysis Commons

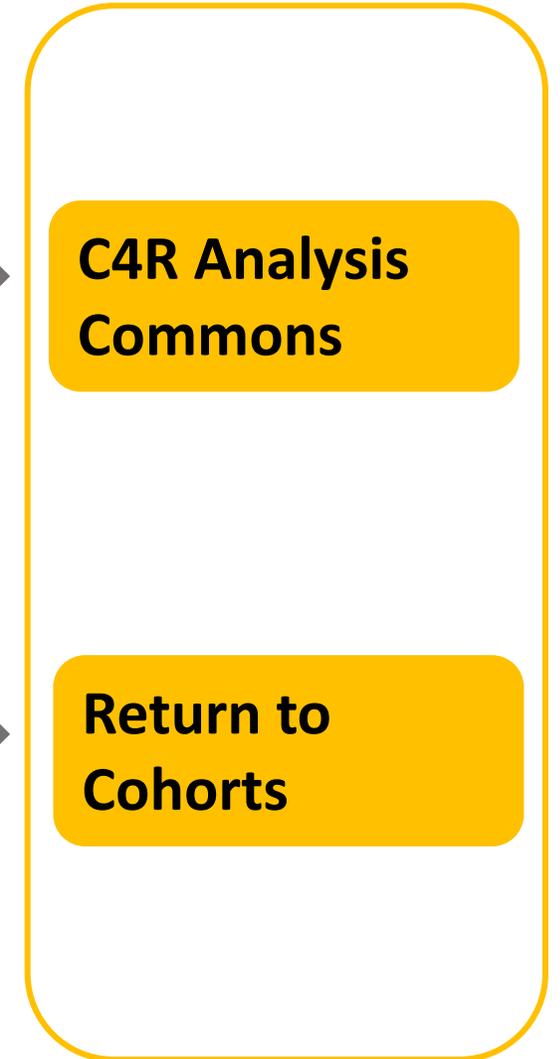
I. Identify Data Sources



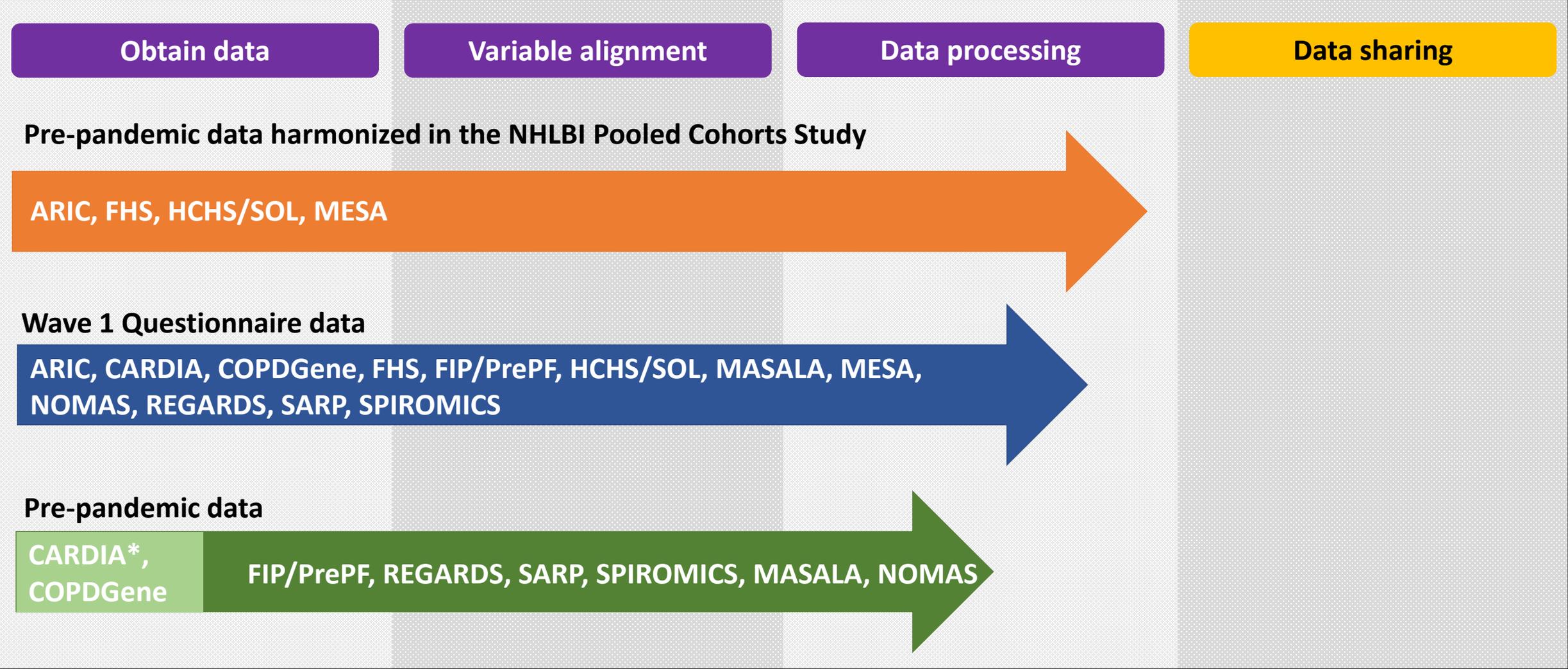
II. Data Harmonization



III. Data Sharing



Harmonization tracking in C4R Analysis Commons



DUAs are pending with JHS and SHS

W1Q COVID-19 Endpoints Variables

	ARIC	CARDIA	COPDGene	FHS	FIP/PrePF	HCHS/SOL	MASALA	MESA	NOMAS	REGARDS	SARP	SPIROMICS
COVID_Infection	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
COVID_Selfreport	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
COVID_HCP	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
COVID_Hospitalized	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
COVID_Testpositive	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
COVID_severity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
COVID_reinfection				✓	✓		✓		✓			
COVID_symptoms	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
COVID_ICU				✓	✓		✓	✓	✓			✓
COVID_recover	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
COVID_vaccine	✓			✓	✓		✓		✓			

DUAs are pending with JHS and SHS

Dried Blood Spot Variables

- Antibodies against the spike (S) or nucleocapsid (N) proteins of the SARS-CoV-2 virus
- Natural COVID-19 infection → *Antibodies against both N and S proteins*
- COVID-19 vaccination → *Antibodies against only S protein*

Threshold for Reactive test:

Antibodies against N protein: ≥ 5000 AU/mL

Antibodies against S protein : ≥ 1000 AU/mL

C4R variable	Explanation
Test_result	Reactive, Non-reactive, Indeterminate, Test cancelled
Num_blood_spots	Number of DBS: 1-5
MFI_N	IgG antibody titers against N protein
MFI_S	IgG antibody titers against S protein
MFI_N_reactive	Reactive to N protein based on threshold of 5000Au/mL: yes, no
MFI_S_reactive	Reactive to S protein based on threshold of 1000Au/mL: yes, no
Test_date	Date of blood spots collection

MFI: Mean fluorescence intensity units

Adjudicated COVID-19 Events Variables

- COVID-related hospitalizations and deaths
 - ICD codes
 - Medical records
- Level of certainty for each diagnosis
 - Definite
 - Probable
 - Not applicable
- Additional data
 - Vital signs (RR, O2, O2 supplementation)
 - Medications

Diagnosis	Severity	Complications
COVID-19...	Infection	Pneumonia
	Hospitalization	Myocardial infarction
	Severe illness	Stroke
	Critical illness	Pulmonary embolism
	Fatal illness	DVT
		Renal failure

Core pre-pandemic measures

- Socio-demographics: age, sex, race/ethnicity, education, birth year
- Anthropometry: height, weight, BMI, hip and waist circumference
- Smoking history: smoking status, pack-years, cigarettes per day
- Spirometry: pre- and post-bronchodilator measurements, QC variables
- Past medical history: hypertension, diabetes, CVD, CKD, cancer, etc.
- Medications: anti-hypertensives, insulin, oral hypoglycemics, statins, steroids, aspirin, etc.
- Blood pressure measurements: systolic and diastolic
- Labs:
 - Renal biomarkers: serum creatinine, uACR, eGFR
 - Lipids: HDL, LDL, total cholesterol, triglycerides
 - Inflammatory biomarkers: CRP, fibrinogen, D-dimer
 - Blood glucose levels, HbA1c
- Respiratory symptoms: cough, phlegm, shortness of breath, etc.
- Adjudicated cardiac events: MI, CVD, CHF, Stroke, etc.
- Adjudicated pulmonary events: asthma, emphysema, chronic bronchitis, COPD

Socio-demographic Characteristics of C4R population

	W1Q completion		Infections to date	
	N	%	N	%
Total	43,752	100%	3,528	100%
Age				
18-20	17	0%	3	0%
21-24	14	0%	0	0%
25-49	3,415	8%	439	12%
50-64	10,727	25%	1,261	36%
≥ 65	28,210	64%	1,743	49%
unknown	1,369	3%	82	2%
Race				
American Indian or Alaska Native	1,881	4%	411	12%
Asian	1,036	2%	34	1%
Black or African American	9,501	22%	647	18%
Native Hawaiian or Other Pacific Islander	34	0%	5	0%
White	24,125	55%	1,711	48%
Other (please specify below)	6,356	15%	653	19%
unknown	819	2%	67	2%
Ethnicity				
Hispanic or Latino	12,474	29%	1,151	33%
Not Hispanic or Latino	24,283	56%	1,770	50%
unknown	6,995	16%	607	17%
Sex Assigned at Birth				
Male	17,770	41%	1,360	39%
Female	25,979	59%	2167	61%

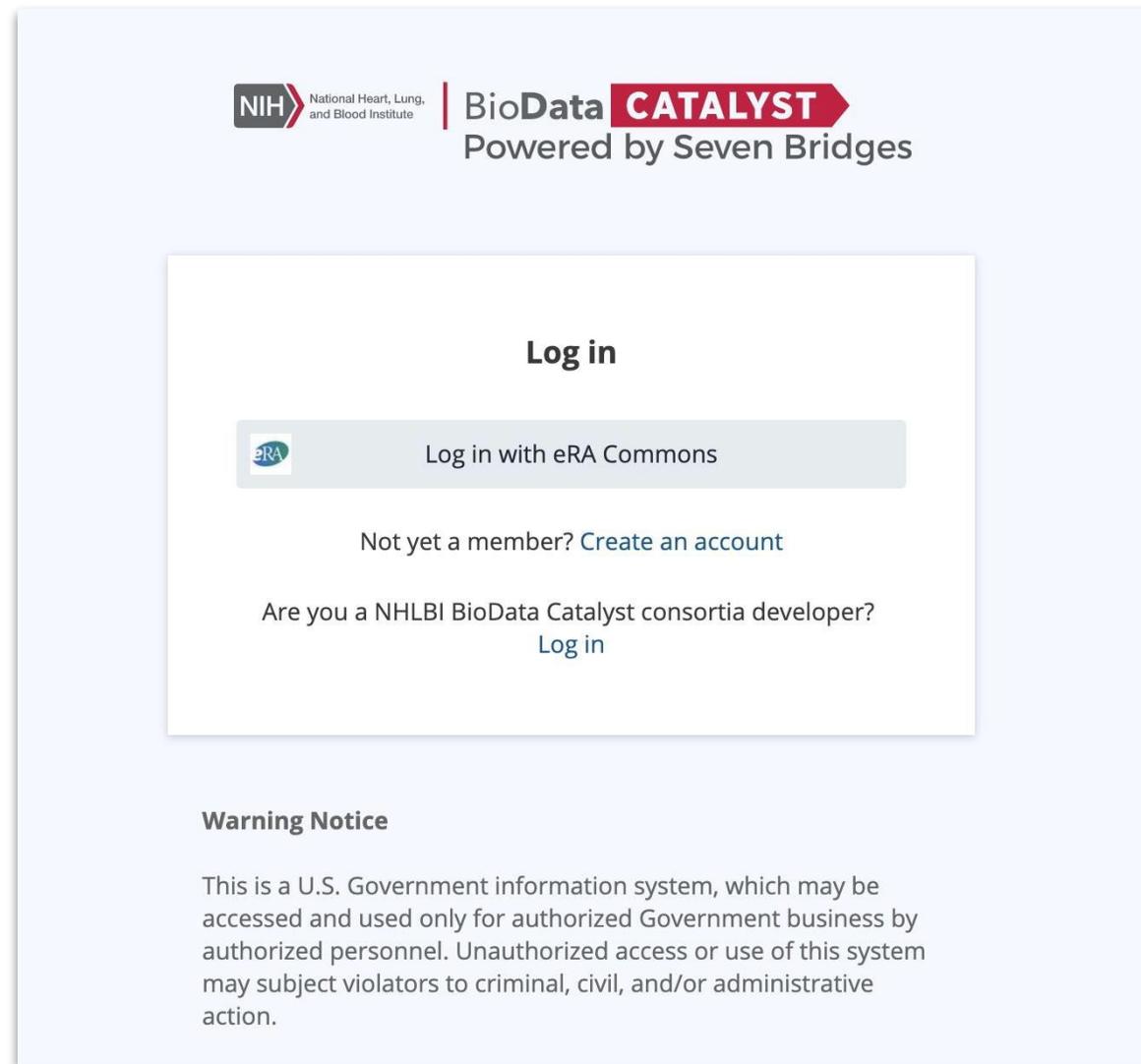
C4R Analysis Commons: BioData Catalyst powered by Seven Bridges

- Cloud-based data sharing and analysis platform
- Secure place to upload data and share with the C4R DCHC as well as researchers/investigators at other institutes
- Analysis software: SAS Studio, R Studio, JupyterLab
- <https://biodatacatalyst.nhlbi.nih.gov/>
- Who can sign up?

The screenshot shows the BioData Catalyst website homepage. At the top left is the NIH logo (National Heart, Lung, and Blood Institute) and the BioData CATALYST logo. A navigation bar includes links for HOME, ABOUT, RESOURCES, FELLOWS, and CONTACT. The main banner features the text "Advancing access to TOPMed data" and "BioData Catalyst provides one point of entry to the most TOPMed datasets, including Freeze 8 data." On the right side of the banner, two statistics are displayed: "406,853 Participants" and "3.42 Petabytes of Data". Below the banner, a central text block reads "Access biomedical data when you need it and how you need it". To the right of this text is a cluster of seven hexagonal icons representing different services: LEARN (graduation cap), DATA (database cylinder), ESTIMATE (dollar sign), SERVICES (wrench and screwdriver), BYOD (hand holding server), JOIN (group of people), and a Help button (question mark icon).

How to get an account on BioData catalyst/Seven Bridges?

- <https://platform.sb.biodatacatalyst.nih.gov/>
- Connect with your eRA Commons ID
- [Getting Started Guide](#)
- C4R DCHC will add you to the appropriate projects with the C4R billing group
- You can test out the platform using your \$500 of pilot funding



NIH National Heart, Lung, and Blood Institute | BioData **CATALYST**
Powered by Seven Bridges

Log in

 Log in with eRA Commons

Not yet a member? [Create an account](#)

Are you a NHLBI BioData Catalyst consortia developer?
[Log in](#)

Warning Notice

This is a U.S. Government information system, which may be accessed and used only for authorized Government business by authorized personnel. Unauthorized access or use of this system may subject violators to criminal, civil, and/or administrative action.

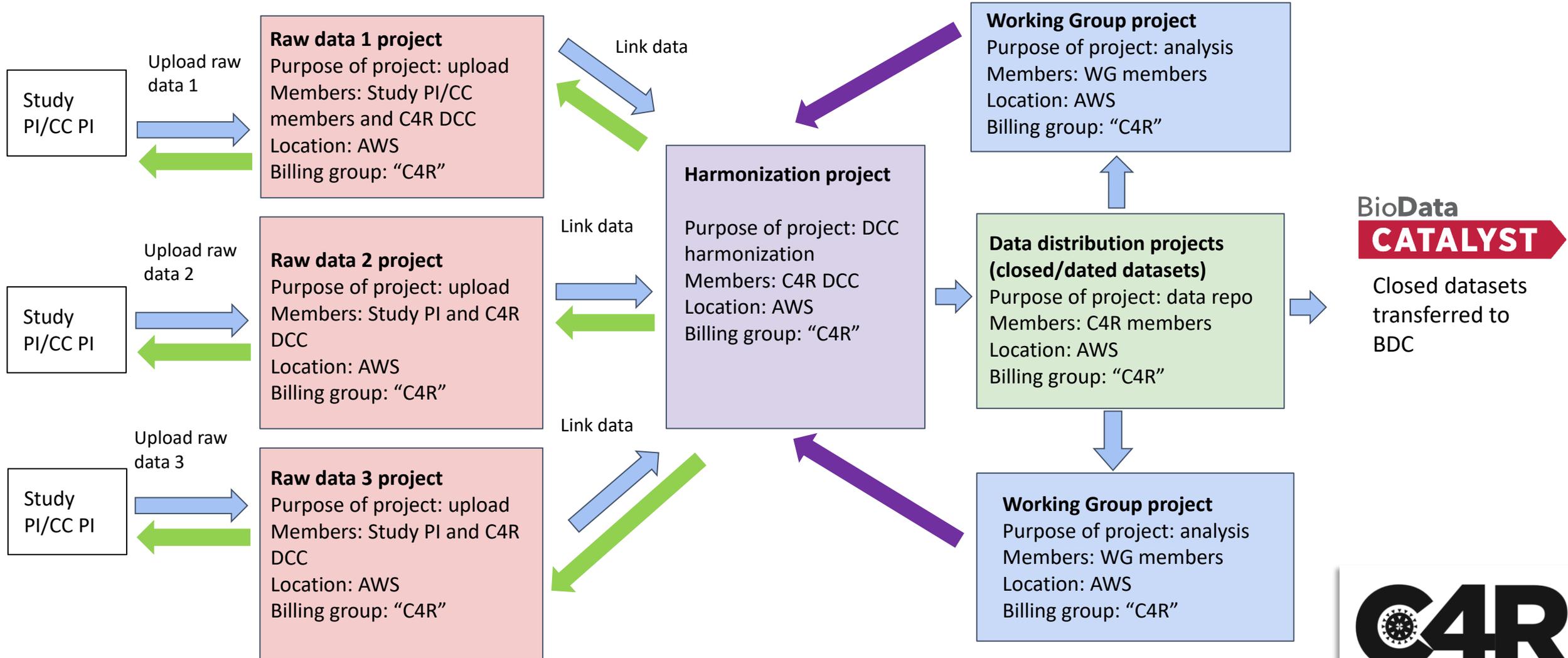
C4R Analysis Commons: Organization and Access

Cohort Access

Consortium Access

Public Access

C4R DCHC Access



Cohorts receive their own harmonized or derived C4R variables, not those from other cohorts

Working groups harmonize necessary additional variables and share back with main project for inclusion in future datasets



BioData Catalyst/Seven Bridges: Data Security

- Users are prohibited from downloading any controlled access, individual-level data
- Project privileges
 - Write
 - Copy
 - Execute
 - Admin
- [BioData Catalyst Security Statement](#)

How to create a project?

The screenshot displays the BioData CATALYST web application interface. At the top, there is a navigation bar with the NIH logo, 'BioData CATALYST Powered by Seven Bridges', and several dropdown menus: 'Projects', 'Data', 'Public Gallery', 'Public projects', and 'Developer'. A user profile 'pbalte' is visible in the top right corner.

The main content area is divided into two columns. The left column is titled 'PROJECTS' and contains a search bar and a list of project entries. The right column is titled 'ANALYSES' and contains a search bar and a section for 'Tasks' and 'Data Cruncher'. A large watermark 'BioData CATALYST' is overlaid on the right side.

The 'PROJECTS' list includes the following entries:

- CONTROLLED** Working group 1
Created by pbalte · July 19, 2021 08:32
- Test_New_Security
Created by pbalte · May 6, 2021 14:15
- SAS Demo - C4R
Created by garrett.rupp · Apr. 30, 2021 15:56
- DBS Data
Created by pbalte · Mar. 31, 2021 10:48
- ARIC-JHS
Created by pbalte · Mar. 4, 2021 14:52
- C4R example project
Created by dave · Feb. 26, 2021 14:23
- C4R Harmonization
C4R
Created by pbalte · Feb. 25, 2021 15:41
- ARIC
C4R **NHLBI-PCS**
Created by pbalte · Feb. 25, 2021 12:16
- SAMPLE Study
C4R
Created by pbalte · Feb. 17, 2021 12:02
- SHS
C4R **NHLBI-PCS**
Created by pbalte · Feb. 17, 2021 10:56

At the bottom of the 'PROJECTS' list, there is a blue button labeled '+ Create a project' and a link 'View all projects'. This button and link are circled in red. To the right of these are navigation arrows.

DESCRIPTION

🏷️ Tags

Title:

Introduction:

Aims:

Edit description

MEMBERS

🔔 Email notifications



pbalte **OWNER**

Write, Copy, Execute,
Admin

Don't work alone.
The best research happens in teams.

👤+ Invite new members

Share your tools, data, and ideas with collaborators

ANALYSES

Search



Tasks

Data Cruncher

Your executions will appear here.
Before you start, learn more about them.

Project Dashboard

Project Files

The screenshot displays the BioData CATALYST interface. At the top, the header includes the NIH logo, 'BioData CATALYST Powered by Seven Bridges', and navigation tabs for 'Projects', 'Data', 'Public Gallery', 'Public projects', and 'Developer'. A user profile 'pbalte' is visible in the top right. Below the header, a secondary navigation bar shows 'Dashboard', 'Files', 'Apps', and 'Tasks'. A red 'CONTROLLED' badge is present next to 'Working group 1'. The main content area is titled 'Files' and contains a search bar, filter buttons for 'Type: All', 'Sample ID: All', 'Task ID: All', and 'Tags: All', and a 'Clear filters' button. A table lists files with columns for 'Name' and 'Size'. One file, 'DCHC_Datasets', is listed. In the top right of the file area, a red circle highlights a 'New folder' button, a '+ Add files' button, and a three-dot menu. The '+ Add files' button is expanded, showing a dropdown menu with the following options: 'Public Files', 'Projects', 'Your Computer', 'FTP / HTTP', 'GA4GH Data Repository Service (DRS)', 'Data Tools', and 'Volumes'.

How
an

Dashboard Files Apps Tasks **CONTROLLED** Working group 1 Interactive Analysis Settings Notes

Interactive data science and scientific computing across multiple programming languages.

Your analyses will appear here. Learn more

Create your first analysis

Email notifications

teams.

collaborators

ANALYSES Search

Tasks Data Cruncher

Your executions will appear here. Before you start, learn more about them.

Starting analysis in SAS Studio

Create new analysis

Basic information Compute requirements

Analysis name
Analysis-1 SAS

Environment

- JupyterLab**
Web-based UI for Project Jupyter
- RStudio**
IDE for R
- SAS Studio BETA**
Analytics and data management platform

Environment setup ?
SAS Data Science

Previous **Next**

Create new analysis

Basic information Compute requirements

Select an instance type with adequate CPU, memory and storage allocation for your analysis. This can be changed between analysis runs, but not while the analysis is running.

Instance type
c5.2xlarge (1024GB EBS, 8vCPUs, 16GB R...
Price: \$0.48 per hour

Suspend time ? On
30 Minutes

Save Draft Previous **Start the analysis**

SAS Studio Environment

The screenshot displays the SAS Studio interface for an analysis session titled "Analysis-1 SAS". The top navigation bar includes a "Stop" button and a "Help" dropdown menu. The main interface is divided into several sections:

- Start Page:** Contains a "New" menu, "Options", and "View" settings. It features a "Start Page" tab and a "GET STARTED" section with options like "New SAS Program", "New Import", and "New Query". There is also a "LEARN" section and a "STAY CONNECTED" section.
- Explorer:** A panel for navigating through the project files.
- Tasks:** A panel for managing tasks, including a search filter.
- Snippets:** A panel for managing code snippets, also with a search filter.
- Libraries:** A panel for managing data libraries.
- Git Repositories:** A panel for managing Git repositories, featuring a diagram of a computer connected to a cloud with databases and the text "ADD OR CLONE A GIT REPOSITORY." and "Add Repository".

On the left side, there is a vertical toolbar with icons for various actions, each highlighted with a colored box: a document icon (orange), a gear icon (green), a document with a plus icon (blue), a document with a trash icon (red), and a document with a link icon (yellow).

The bottom right corner shows a "Submission (0)" status indicator.

Starting analysis in R Studio

Create new analysis ✕

Basic information **▶** Compute requirements

Analysis name
Analysis-1 R

Environment

- JupyterLab**
Web-based UI for Project Jupyter
- RStudio**
IDE for R
- SAS Studio** BETA
Analytics and data management platform

Environment setup ⓘ
SB Bioinformatics - R 4.0 ▼

Previous **Next**

Create new analysis ✕

Basic information **▶** Compute requirements

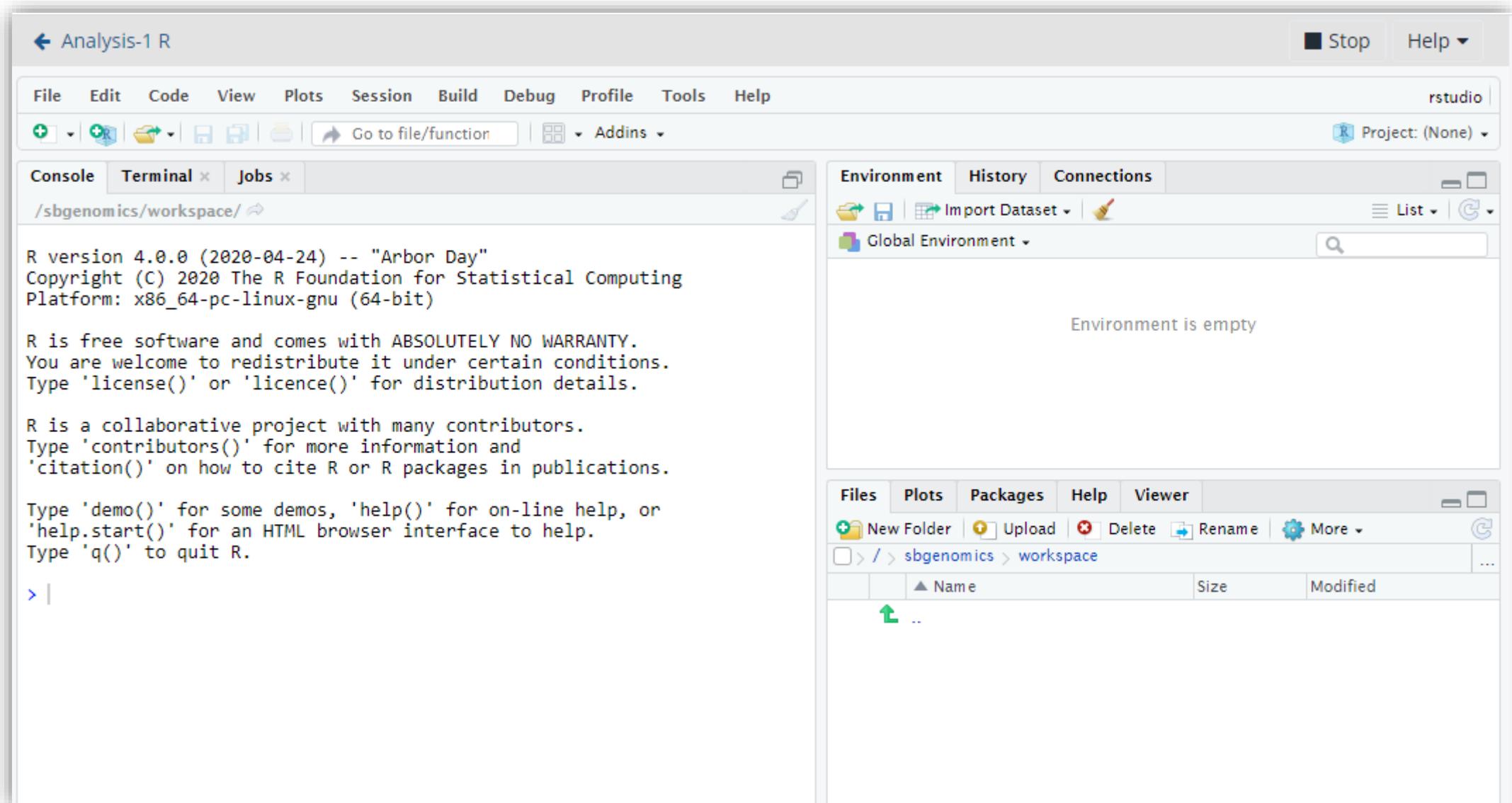
Select an instance type with adequate CPU, memory and storage allocation for your analysis. This can be changed between analysis runs, but not while the analysis is running.

Instance type
c5.2xlarge (1024GB EBS, 8vCPUs, 16GB R...
Price: \$0.48 per hour

Suspend time ⓘ On
30 Minutes

Save Draft Previous **Start the analysis**

R Studio Environment



Locations of data library in analysis environment

- Data import location: /sbgenomics/project-files/**name of the folder**
- Data export location: /sbgenomics/output-files
- All exported files are saved in the “Files” of the project
- If desired, exported files can be manually moved to other folders after stopping the analysis

```
libname lib1"/sbgenomics/project-files/DCHC_Datasets";  
libname out"/sbgenomics/output-files";
```