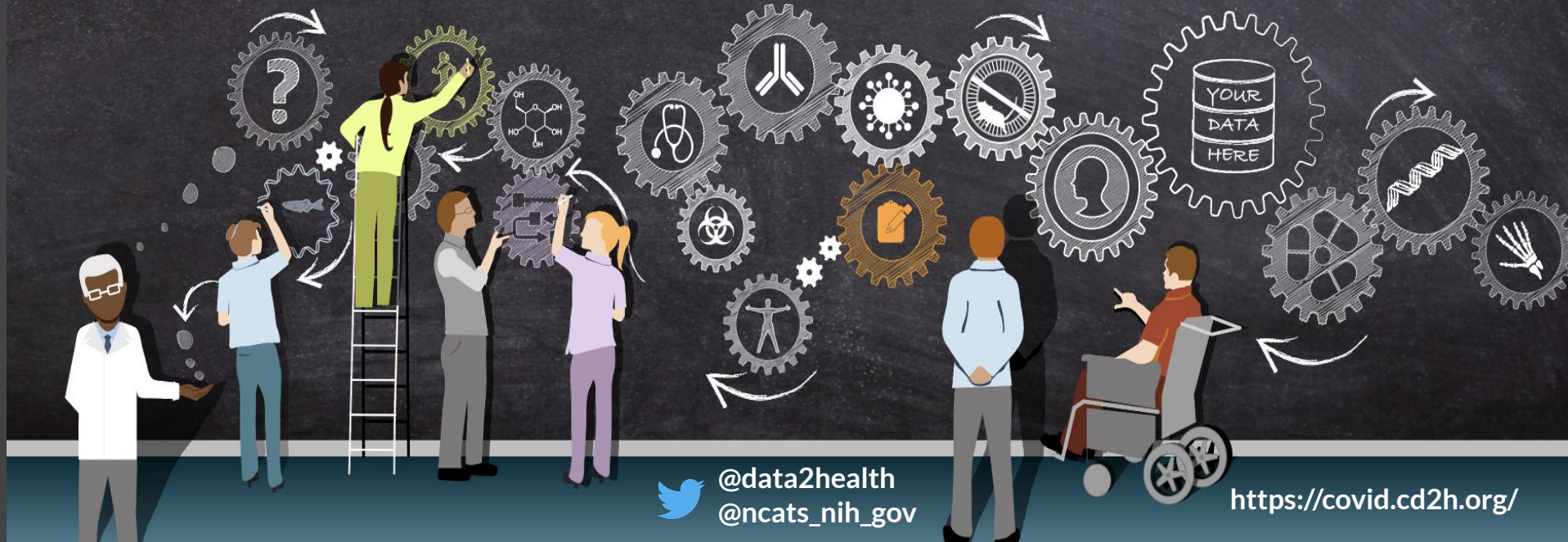
N3C Overview

Education and Training Domain Team Short Course, Spring 2024

Shawn T. O'Neil, CU Anschutz
Stephanie S. Hong, Johns Hopkins



@data2health
@ncats_nih_gov

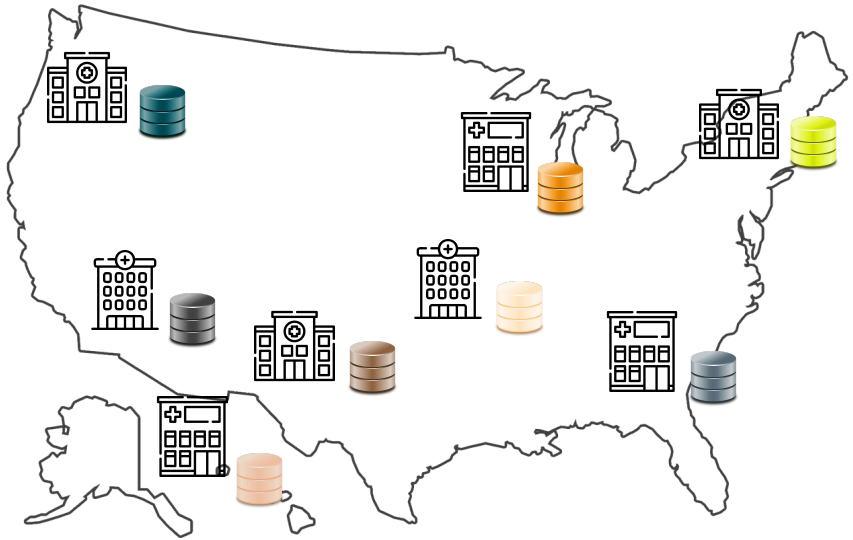
https://covid.cd2h.org/

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH

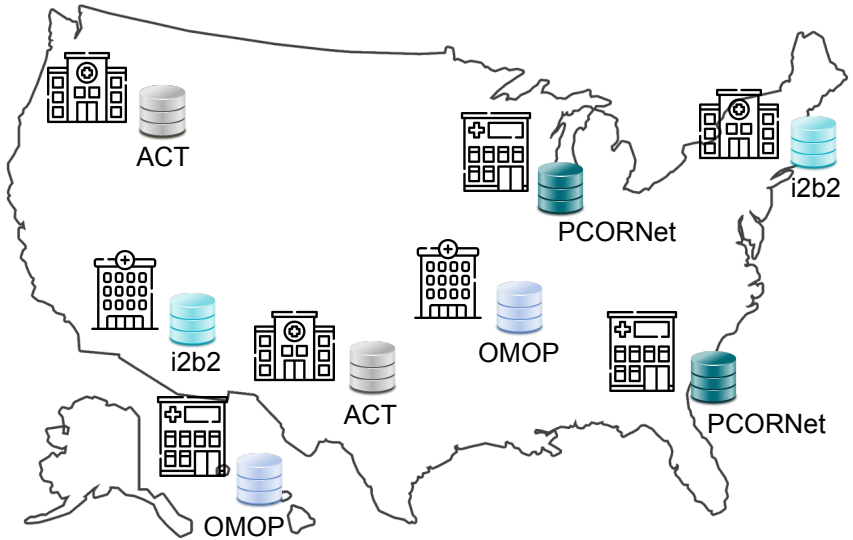


National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH

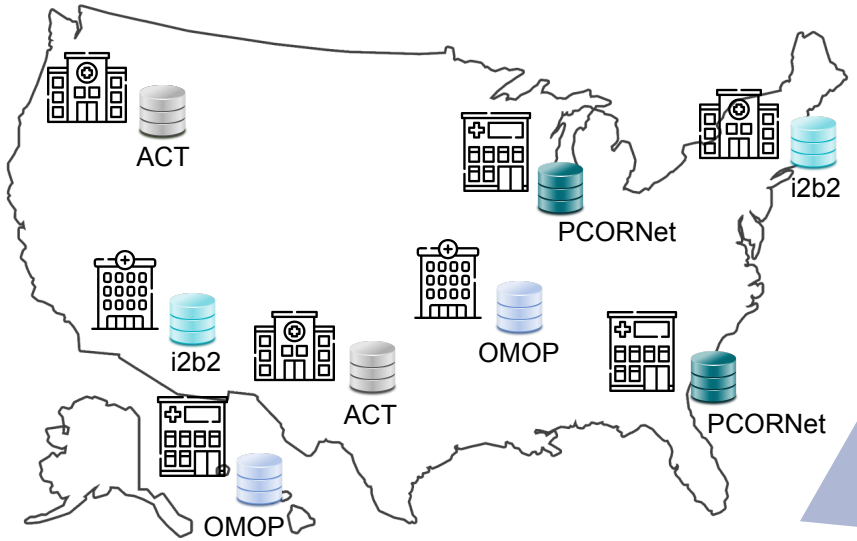


National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

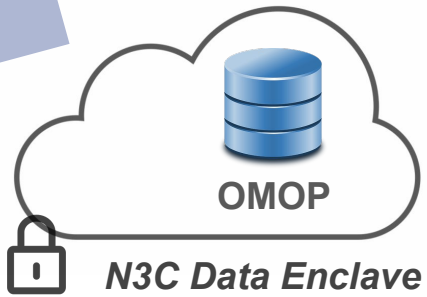
National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH



Data Partners

harmonize



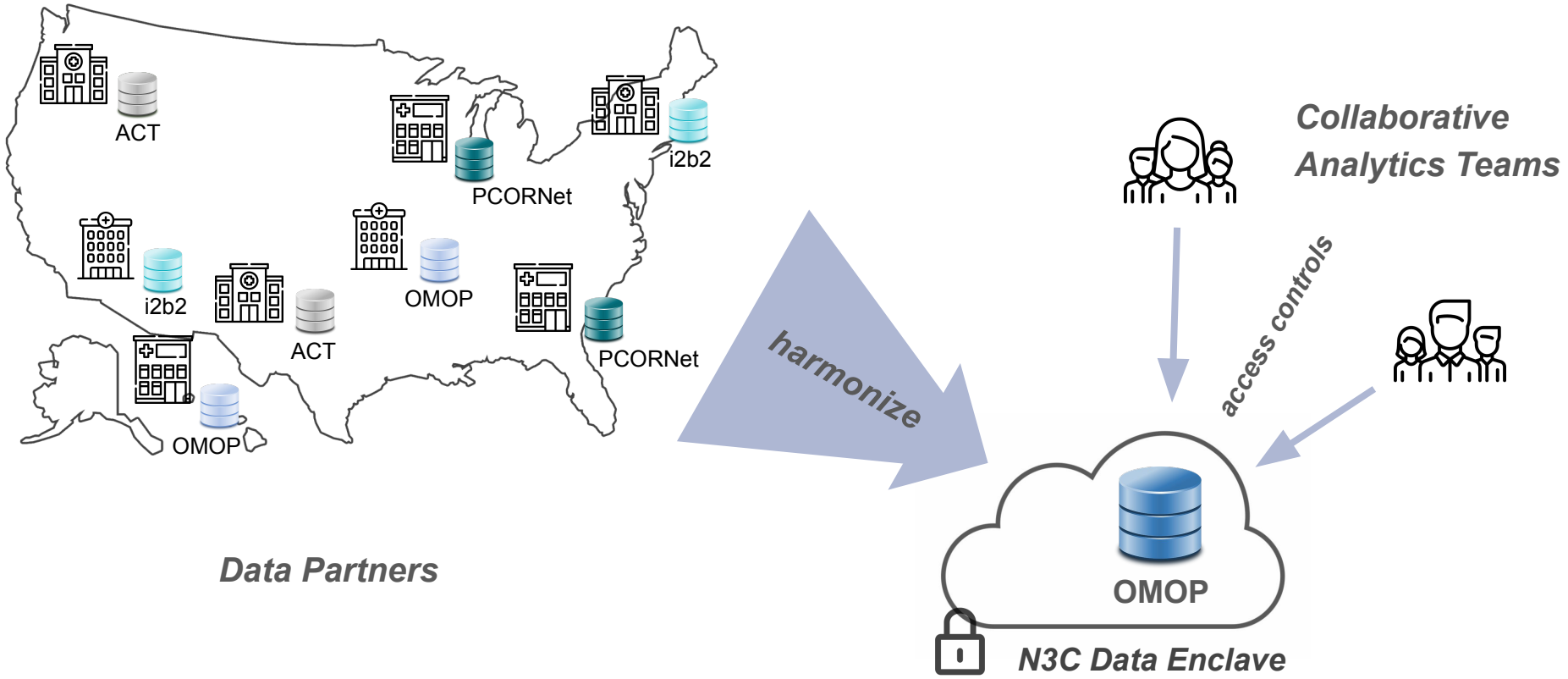
N3C Data Enclave

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH

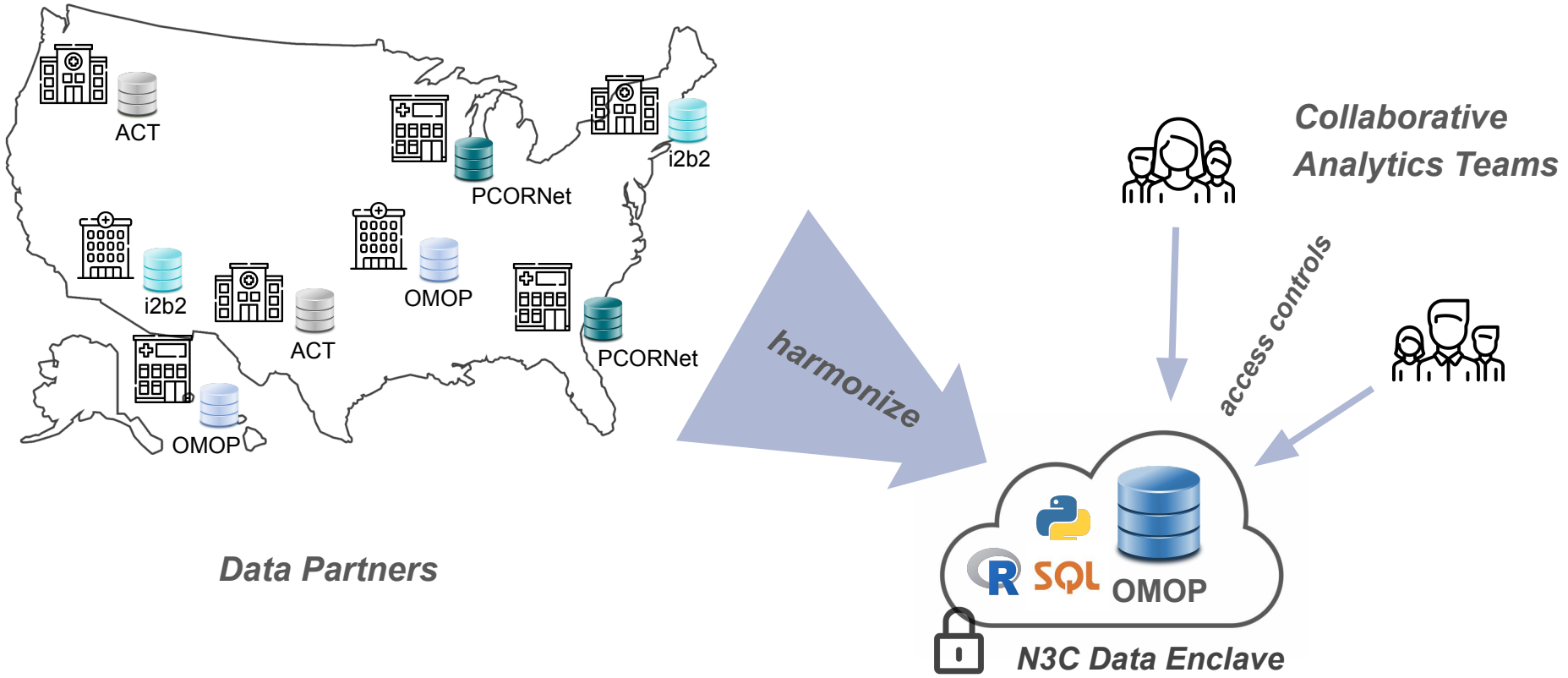


National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH



National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH

17,411,971
TOTAL N3C PATIENTS

6,745,241
CONFIRMED COVID-19 (+)

189,838
POSSIBLE COVID-19 (+)

77
SITES

21.7b
TOTAL ROWS

Each site (data partner) periodically pulls data requested by N3C
1/1/2018 – Present

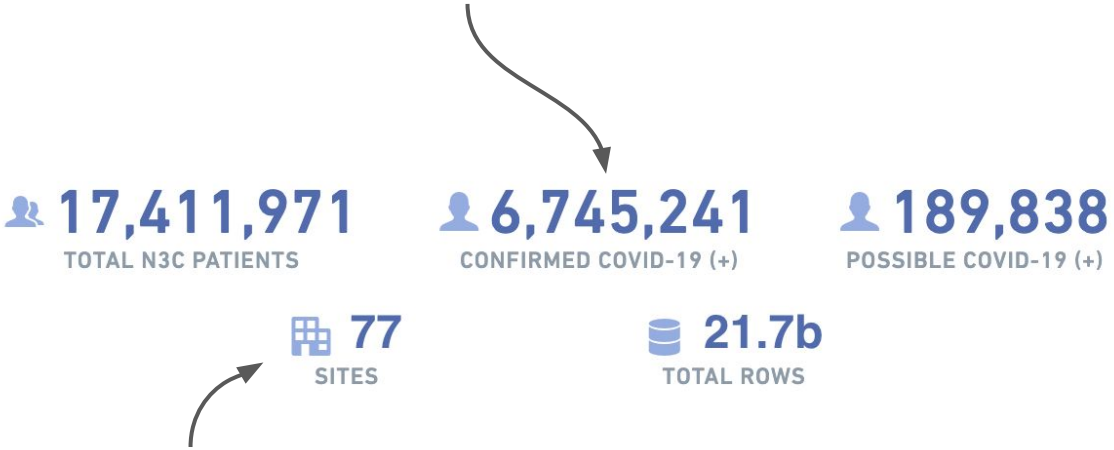
National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH

Positive COVID-19 Lab Test (PCR, Ag, or Antibody)
U07.1 Diagnosis (COVID-19)
U09.9 Diagnosis (Long COVID)
Visit to Long-COVID specialty clinic



Each site (data partner) periodically pulls data requested by N3C
1/1/2018 – Present

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH

Positive COVID-19 Lab Test (PCR, Ag, or Antibody)
U07.1 Diagnosis (COVID-19)
U09.9 Diagnosis (Long COVID)
Visit to Long-COVID specialty clinic

Other strong diagnosis, or two or more symptoms (cough, fever, etc.) Jan - May 2020

Each site (data partner) periodically pulls data requested by N3C
1/1/2018 – Present

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH

Positive COVID-19 Lab Test (PCR, Ag, or Antibody)
U07.1 Diagnosis (COVID-19)
U09.9 Diagnosis (Long COVID)
Visit to Long-COVID specialty clinic

Each Confirmed and Possible COVID-19 patient is matched with 2 others by age, sex, race, ethnicity

Other strong diagnosis, or two or more symptoms (cough, fever, etc.) Jan - May 2020

17,411,971
TOTAL N3C PATIENTS

6,745,241
CONFIRMED COVID-19 (+)

189,838
POSSIBLE COVID-19 (+)

77
SITES

21.7b
TOTAL ROWS

Each site (data partner) periodically pulls data requested by N3C
1/1/2018 – Present

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH

Positive COVID-19 Lab Test (PCR, Ag, or Antibody)
U07.1 Diagnosis (COVID-19)
U09.9 Diagnosis (Long COVID)
Visit to Long-COVID specialty clinic

Each Confirmed and Possible COVID-19 patient is matched with 2 others by age, sex, race, ethnicity

Other strong diagnosis, or two or more symptoms (cough, fever, etc.) Jan - May 2020

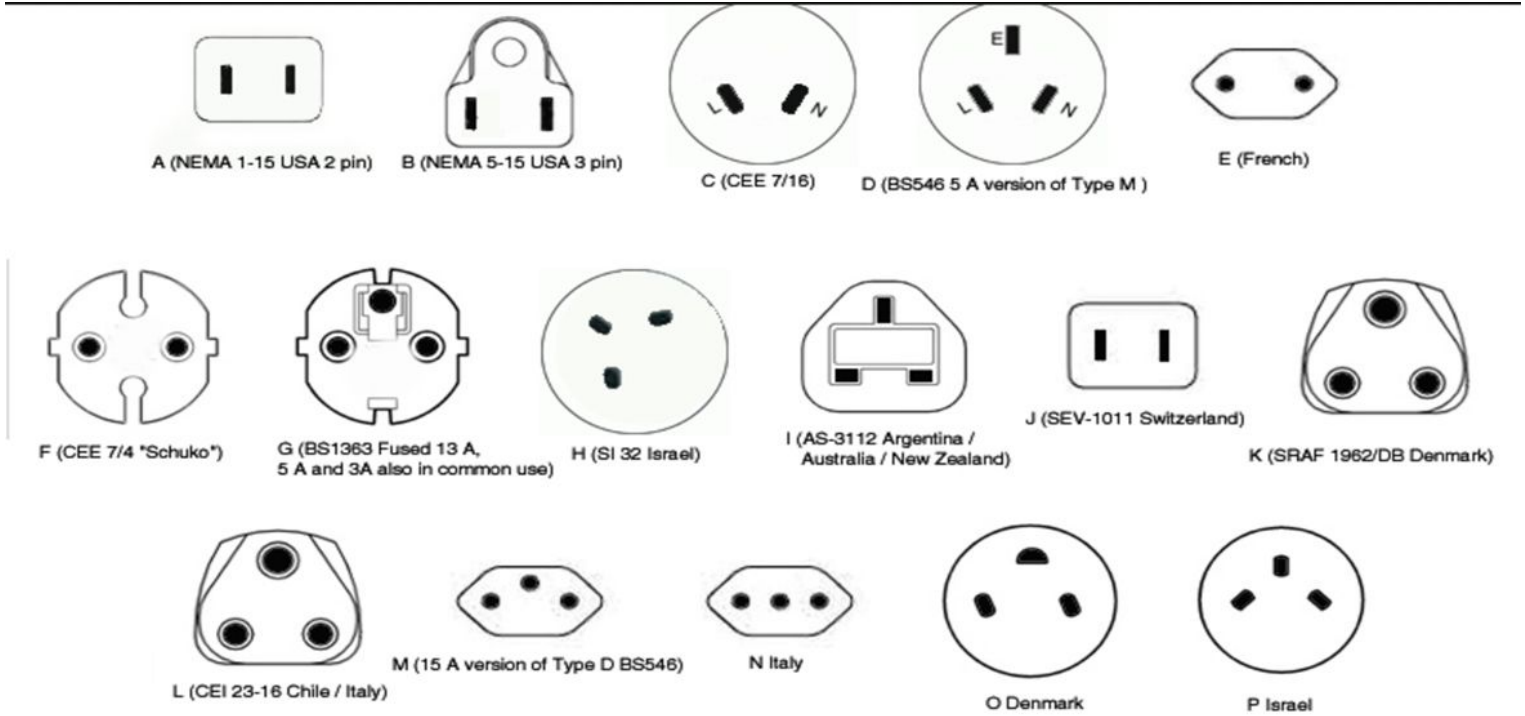
Each site (data partner) periodically pulls data requested by N3C
1/1/2018 – Present

Every week, the most recent data pulls are compiled into a *release*

Harmonization, Data Completeness, Data Quality

Stephanie Hong, FAMIA, Johns Hopkins

Real World Data Standards many standards exist



Stephanie Hong, Johns Hopkins

National COVID Cohort Collaborative

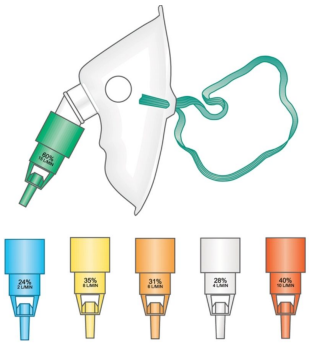
A program of NIH's National Center for Advancing Translational Sciences

Adapt to varying EHR data capture methods and adjust to evolving terminology codes.

NATIONAL CENTER FOR DATA TO HEALTH

O2 Devices -

Significant number of hospitalized COVID patients required oxygen supplementation and/or ventilation support from an oxygen device and/or a ventilator



Vaccination Data

The CDC's National Center of Immunization and Respiratory Diseases (NCIRD) developed and maintains the CVX (vaccine administered) code set.

Long COVID

B94.8 *recommended* before Oct. 1, 2021 "Sequelae of other specified infectious and parasitic diseases") as a placeholder code to signify Long COVID.

CDC announced U09.9 code - officially available effective Oct. 1, 2021

Stephanie Hong, Johns Hopkins

Harmonization

Data formats used by sites introduce heterogeneity

OHDSI — Observational Health Data Sciences and Informatics

Death data supported?	Y
Death date required?	Y
Death cause supported?	Y
Discharge disposition supported?	Y

pcornet — The National Patient-Centered Clinical Research Network

Death data supported?	Y
Death date required?	N
Death cause supported?	Y
Discharge disposition supported?	Y

TriNetX

Death data supported?	Y
Death date required?	Y
Death cause supported?	N
Discharge disposition supported?	N

ACT — The ACT Network

Death data supported?	Y
Death date required?	N
Death cause supported?	N
Discharge disposition supported?	N

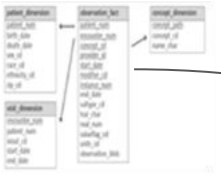
Emily Pfaff, PhD, UNC

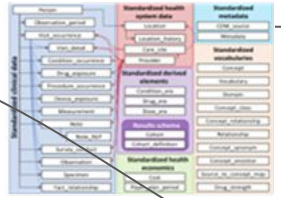
National COVID Cohort Collaborative

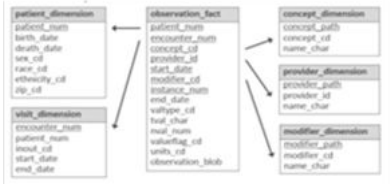
A program of NIH's National Center for Advancing Translational Sciences

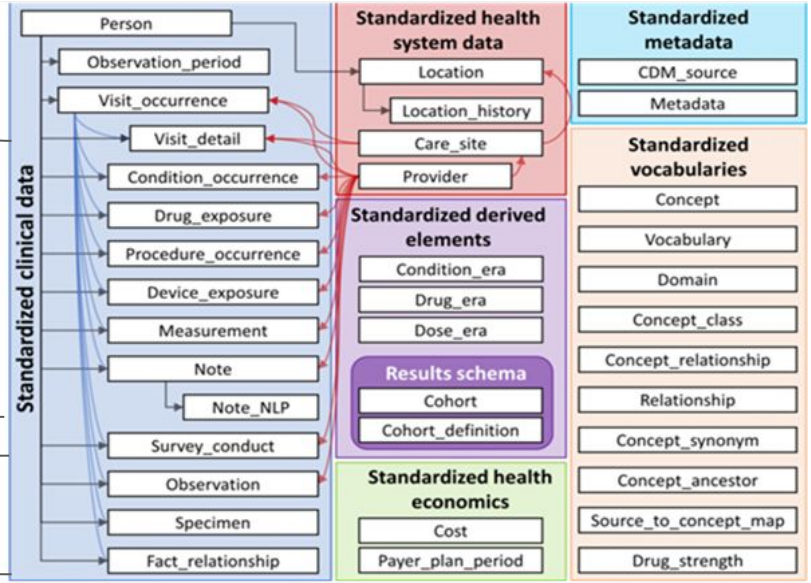
Harmonized into OMOP CDM widely accepted Common Data Model for capturing observational patient level data

NATIONAL CENTER FOR DATA TO HEALTH









Stephanie Hong, Johns Hopkins

Findable, Accessible, Interoperable (Harmonized), Reusable

What does **Interoperable** mean with respect to data? **Harmonized!**

Syntactic Interoperability (harmonization)

- **Structure - data can be found in the expected domain**
- Domain of the data standards and data model communities

Semantic interoperability (harmonization)

- **Unified meaning** (LOINC, ICD10CM, SNOMED CT, CPT, HCPCS => **concept_ids (understandable)**
- Domain of the vocabulary, ontology, classification communities

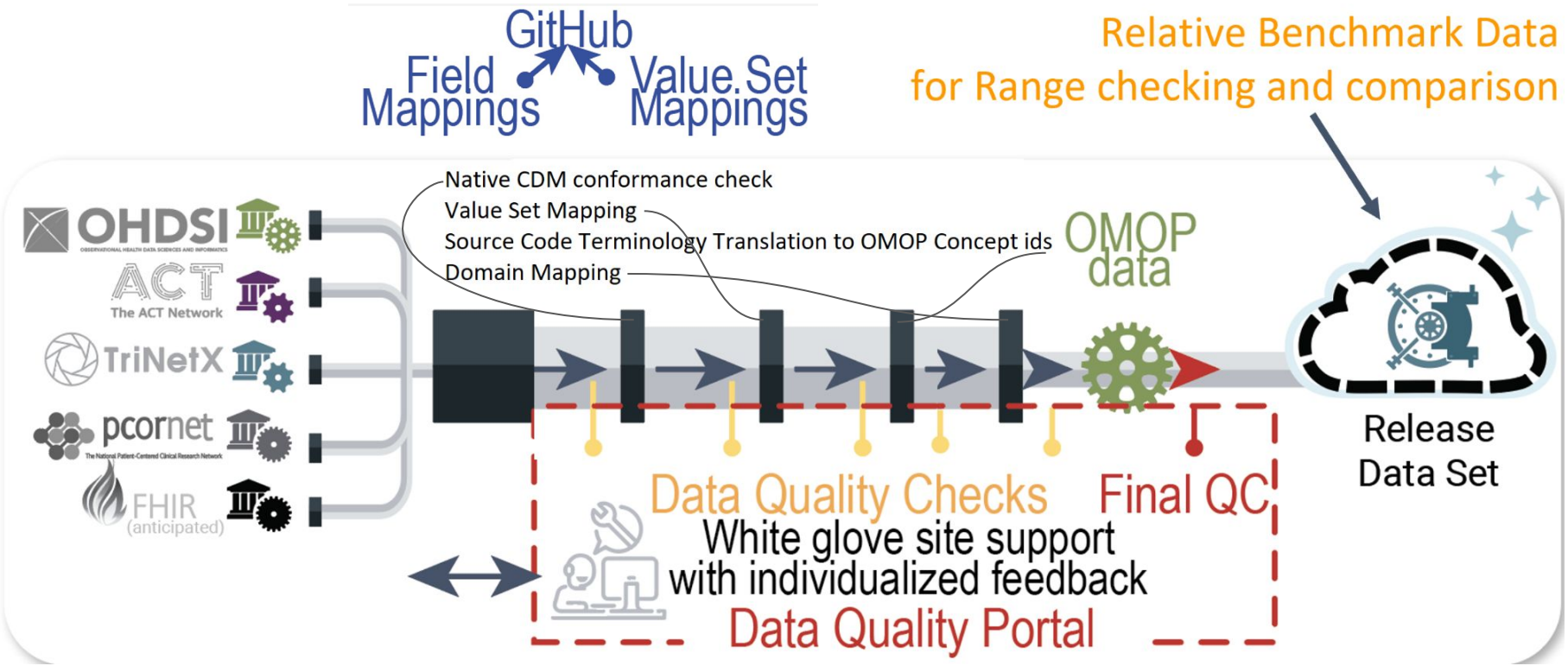
Reusable - Same query can be executed in all OMOP CDM!

Stephanie Hong, Johns Hopkins

National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Limited Data Set (LDS) Data Ingestion Pipeline
Map Every Data Element Field and Its Value

NATIONAL CENTER FOR DATA TO HEALTH



Stephanie Hong, Johns Hopkins

Harmonized

The Data Ingestion and Harmonization Pipeline incorporates a simplified "Data Mapping Framework" to manage diverse data harmonization processes across sites.

CDM Native Codes => value set mapping table

Data Source Specific **Value Sets** - Each CDM has their own enumerated list of permissible coded values that contains specific meaning (e.g Female)

- DEM|SEX:F
- F
- UMLS:HL7V3.0:Gender:F
- 8532

Terminology based Codes => OMOP Concept table

- SNOMED CT
- ICD10CM
- ICD10PCS
- HCPCS
- CPT4
- LOINC
- RxNorm
- NDC
- PPI

N3C Custom Codes - example of example of the CONCEPT records for representing custom codes

- Concept_id = 2004208004
  Concept_name = Other oxygen device
  Domain_id = Device
  Vocabulary_id = N3C

- Concept_id = 2004208005
  Concept_name = Room air (in the context of a device)
  Domain_id = Device
  Vocabulary_id = N3C

Stephanie Hong, Johns Hopkins

N3C Custom Code Management

Data Ingestion and Harmonization Pipeline has Data Mapping Framework to handle **local custom vocabulary codes.** The items in the table below do not have a valid matching SNOMED code. It is possible that a clinician may recognize one of these times as being substantially similar to an existing, valid SNOMED code. However, we know that not all sites will have the ability to consult with a subject matter expert, so the "**Custom Other O2 Device Code**" may be used.

O_2 Device	Code (See data model for specific instructions below)
Bag Mask	Custom Other O2 Device Code
Bubble CPAP	Custom Other O2 Device Code
Face Bucket	Custom Other O2 Device Code
Manual Ventilation Bag	Custom Other O2 Device Code
NIV (non-invasive ventilator), Non-Invasive Positive Pressure	Custom Other O2 Device Code
optiflow	Custom Other O2 Device Code
Vapotherm	Custom Other O2 Device Code
Room Air	Custom Room Air Code

Concept_id	2004208004
Concept_name	Custom Other Oxygen Device Code
Domain_id	Device
Vocabulary_id	***N3C***

Stephanie Hong, Johns Hopkins

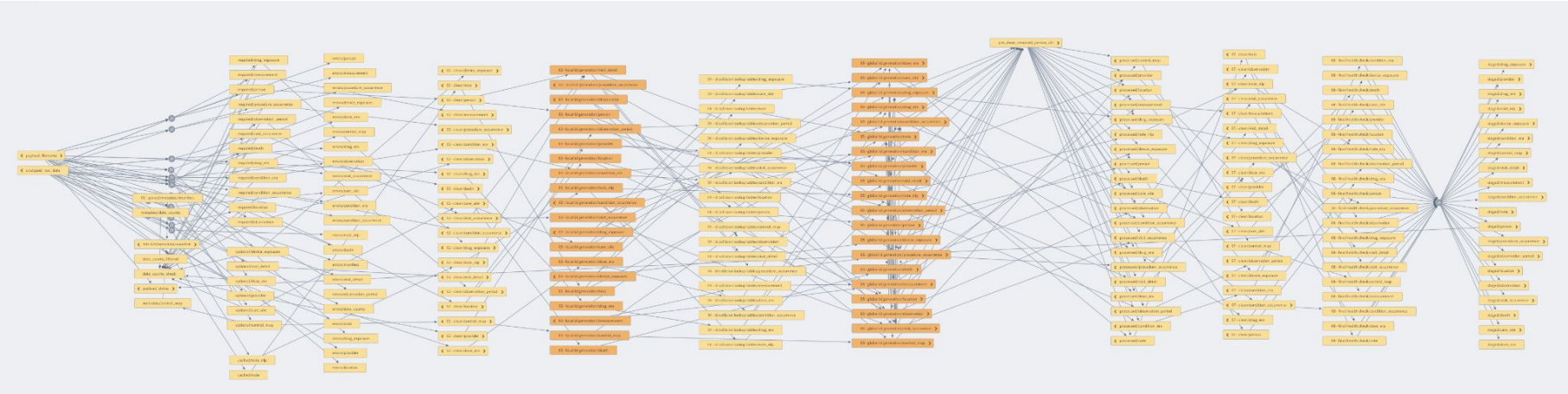
National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

Limited Data Set (LDS) Data Ingestion Pipeline

NATIONAL CENTER FOR DATA TO HEALTH

A unified framework manages stepwise ETL pipeline tasks for ingesting all incoming data. Each site's ingestion pipeline involves over two hundred data transformations, including native CDM checks, terminology translation (semantic and syntactic domain structure alignment), ID generation, and data health checks into OMOP CDM format.



Stephanie Hong, Johns Hopkins

Measurement (numerical & categorical) and Units

Measurement are in numerical value with units or categorical value like "positive", "negative" or "abnormal" - manage by value set mapping table

- Data Element mapping

Stephanie Hong, Johns Hopkins

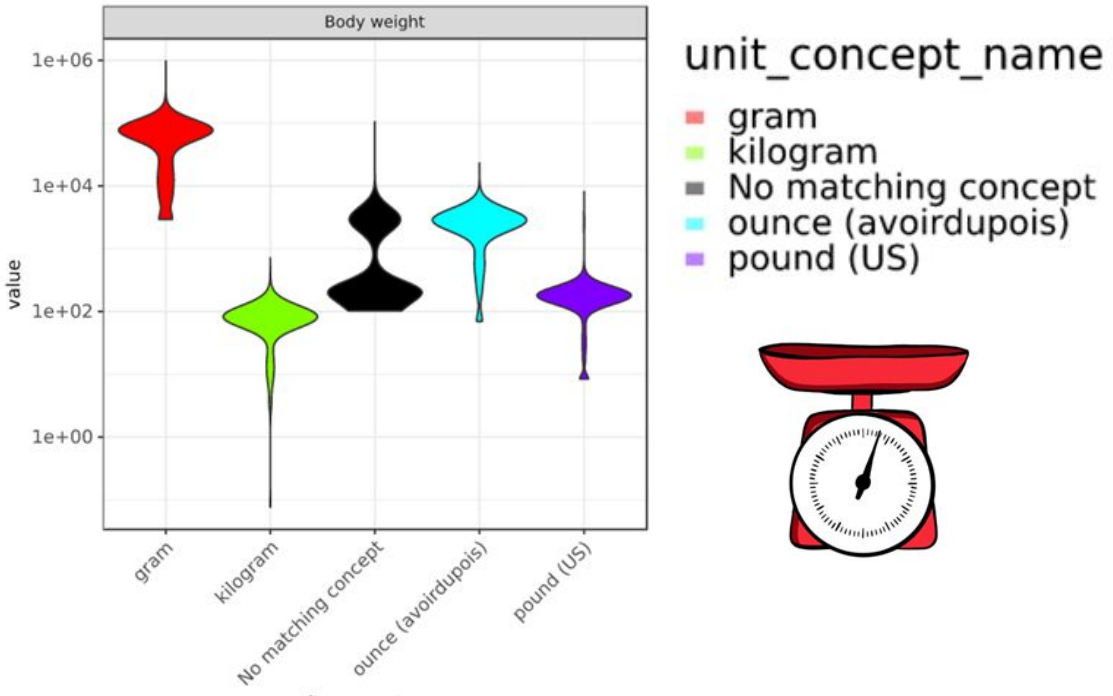
Measurement Unit Harmonization

- **Problem:** Different sites provide their data in different units
- **Solution:** Harmonize each to a standard unit

  Kilograms = **Pounds** / 2.20462

  Kilograms = **Ounces** / 35.274

  Kilograms = **Grams** / 1000



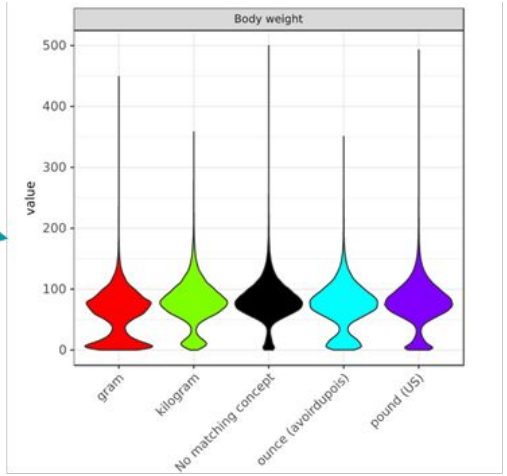
Stephanie Hong, Johns Hopkins

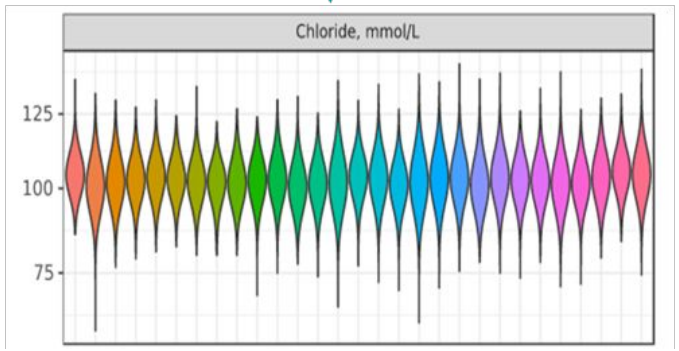
National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Measurement Unit Harmonization

NATIONAL CENTER FOR DATA TO HEALTH

- Harmonized measurements
  - By original unit
  - Across many sites





Homogeneity after harmonization

Humans measured in grams do not look the same as humans measured in kilograms!

Stephanie Hong, Johns Hopkins

National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

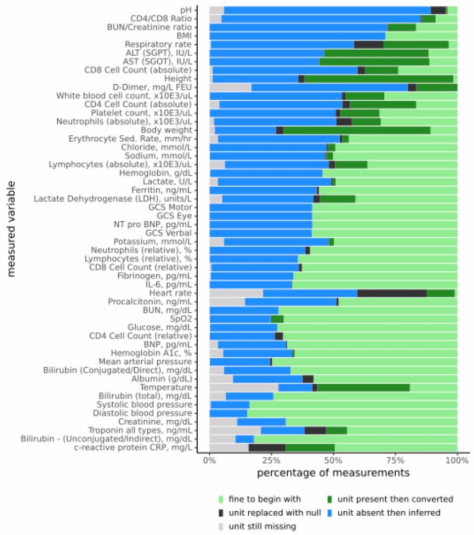
Unit Harmonization

NATIONAL CENTER FOR DATA TO HEALTH

Some Measurement units are missing - Unit Harmonization is performed to rescue measurements with unknown units - harmonized to standard units

Iterate over time and check each ingestion - DOI: 10.1093/jamia/ocac054

Unit harmonization: summary of actions



About half of data are already in the canonical unit. We end up recovering most of the data with unknown units.

Monthly review for new issues

- Changes to formulae
- Changes to thresholds
- Changes to logic

Unit Harmonization- **infer and rescue missing units**

- **Site YYY:** No creatinine data
- **Site XXX:** Creatinine in urine included in blood/serum levels
- **Site ZZZ:** Mass/volume units mixed with molarity

Stephanie Hong, Johns Hopkins

N3C, January 18, 2024
Largest centralized repository of COVID related Patient EHR data in U.S.
But there is more…

Data ‘Levels’ & Extra Data

Shawn O’Neil, CU Anschutz

Levels of Data Access

- **Level 3: Limited Data Set**
  - Accurate Dates (for most data partners)
  - 5-digit zip codes (for most patients)
  - Requires: *Human Subjects Research Protection training, IRB Approval*

Levels of Data Access

- **Level 3: Limited Data Set**
  - Accurate Dates (for most data partners)
  - 5-digit zip codes (for most patients)
  - Requires: *Human Subjects Research Protection training, IRB Approval*

- **Level 2: De-Identified**
  - Dates randomized up to +/- 180 days (per patient)
  - 3-digit zip codes
  - Requires: *Human Subjects Research Protection training*

Levels of Data Access

- **Level 3: Limited Data Set**
  - Accurate Dates (for most data partners)
  - 5-digit zip codes (for most patients)
  - Requires: *Human Subjects Research Protection training, IRB Approval*

- **Level 2: De-Identified**
  - Dates randomized up to +/- 180 days (per patient)
  - 3-digit zip codes
  - Requires: *Human Subjects Research Protection training*

- **Level 1: Synthetic (based on N3C data)**
  - Deprecated

Levels of Data Access

- **Level 3: Limited Data Set**
  - Accurate Dates (for most data partners)
  - 5-digit zip codes (for most patients)
  - Requires: *Human Subjects Research Protection training, IRB Approval*

- **Level 2: De-Identified**
  - Dates randomized up to +/- 180 days (per patient)
  - 3-digit zip codes
  - Requires: *Human Subjects Research Protection training*

- ~~**Level 1: Synthetic (based on N3C data)**~~
  - ~~Deprecated~~

Levels of Data Access

- **Level 3: Limited Data Set**
  - Accurate Dates (for most data partners)
  - 5-digit zip codes (for most patients)
  - Requires: *Human Subjects Research Protection training, IRB Approval*

- **Level 2: De-Identified**
  - Dates randomized up to +/- 180 days (per patient)
  - 3-digit zip codes
  - Requires: *Human Subjects Research Protection training*

- ~~**Level 1: Synthetic (based on N3C data)**~~
  - ~~Deprecated~~

- **"Level 0": Notional synthetic (aka fake)**
  - Requires: *enclave access only*

Levels of Data Access

- **Level 3: Limited Data Set**
  - Accurate Dates (for most data partners)
  - 5-digit zip codes (for most patients)
  - Requires: *Human Subjects Research Protection training, IRB Approval*

Best for: pandemic timelines, geographic work

- **Level 2: De-Identified**
  - Dates randomized up to +/- 180 days (per patient)
  - 3-digit zip codes
  - Requires: *Human Subjects Research Protection training*

- ~~**Level 1: Synthetic (based on N3C data)**~~
  - ~~Deprecated~~

- **"Level 0": Notional synthetic (aka fake)**
  - Requires: *enclave access only*

Levels of Data Access

- **Level 3: Limited Data Set**
  - Accurate Dates (for most data partners)
  - 5-digit zip codes (for most patients)
  - Requires: *Human Subjects Research Protection training, IRB Approval*

Best for: pandemic timelines, geographic work

- **Level 2: De-Identified**
  - Dates randomized up to +/- 180 days (per patient)
  - 3-digit zip codes
  - Requires: *Human Subjects Research Protection training*

Best for: general COVID research
→ We'll access this later!

- ~~**Level 1: Synthetic (based on N3C data)**~~
  - ~~Deprecated~~

- **"Level 0": Notional synthetic (aka fake)**
  - Requires: *enclave access only*

Levels of Data Access

- **Level 3: Limited Data Set**
  - Accurate Dates (for most data partners)
  - 5-digit zip codes (for most patients)
  - Requires: *Human Subjects Research Protection training, IRB Approval*

Best for: pandemic timelines, geographic work

- **Level 2: De-Identified**
  - Dates randomized up to +/- 180 days (per patient)
  - 3-digit zip codes
  - Requires: *Human Subjects Research Protection training*

Best for: general COVID research
→ We'll access this later!

- ~~**Level 1: Synthetic (based on N3C data)**~~
  - ~~Deprecated~~

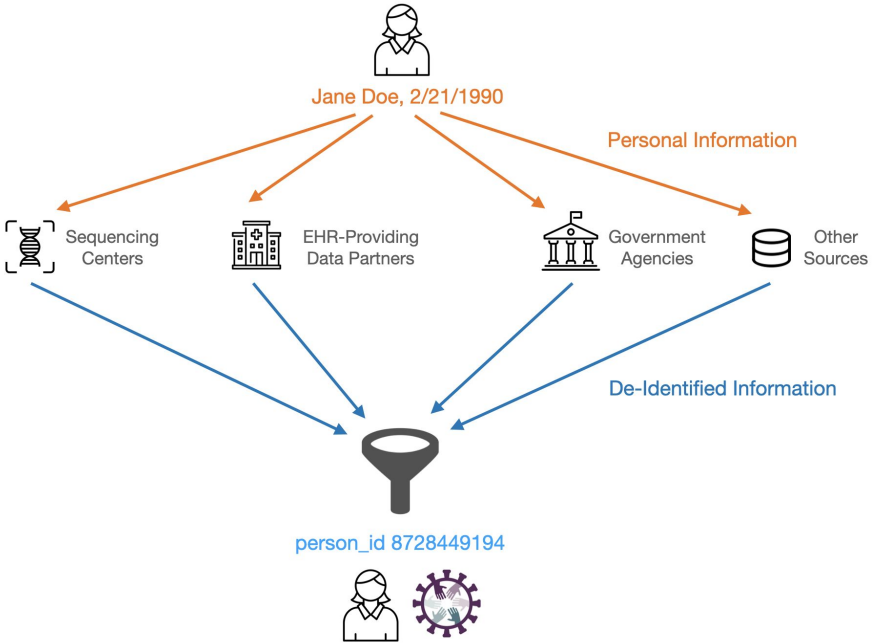
- **"Level 0": Notional synthetic (aka fake)**
  - Requires: *enclave access only*

Best for: practice
→ We'll access this soon!

Privacy-Preserving Record Linkage

PPRL: Securely linking N3C data against additional sources

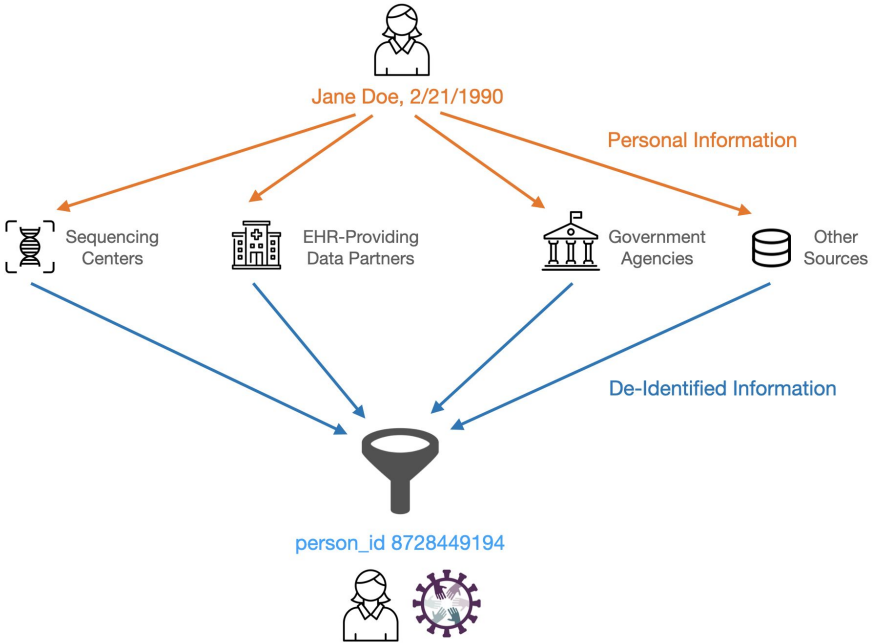
- **Mortality/deaths**
  Private and government sources help fill EHR gaps
- **Viral Variants**
  Sequenced variant (delta, omicron, …)
- **CMS (Medicare/Medicaid) Billing Data**
  Currently only for COVID-diagnosed medicare pts.



Privacy-Preserving Record Linkage

PPRL: Securely linking N3C data against additional sources

- **Mortality/deaths**
  Private and government sources help fill EHR gaps
- **Viral Variants**
  Sequenced variant (delta, omicron, …)
- **CMS (Medicare/Medicaid) Billing Data**
  Currently only for COVID-diagnosed medicare pts.
- Available w/ Level 3 request only (IRB required)
  Only available for subsets of sites & patients
- More info at https://covid.cd2h.org/pprl



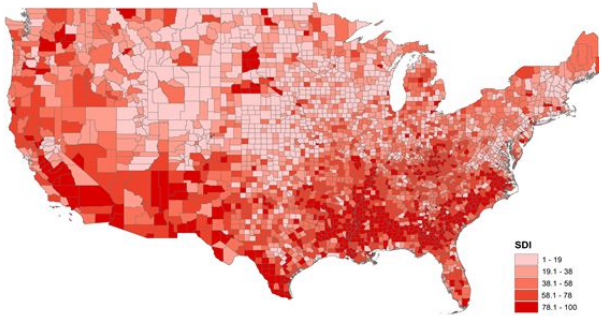
"External" Datasets

Publicly-available datasets available for use

- Researchers cannot upload additional data
- But they can request that data be ingested (after security & legal review)

United States™ Census Bureau

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY



Social Deprivation Index, Robert Graham Center

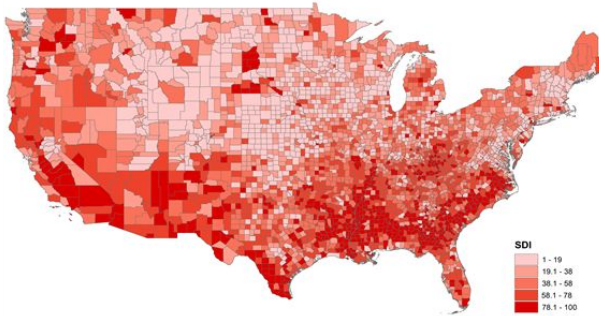
"External" Datasets

Publicly-available datasets available for use

- Researchers cannot upload additional data
- But they can request that data be ingested (after security & legal review)
- Previously reviewed and uploaded data are available to all
- 60 datasets currently
- https://covid.cd2h.org/external-datasets

United States Census Bureau

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY



Social Deprivation Index, Robert Graham Center

Agreements, Policies, Procedures

Shawn O’Neil, CU Anschutz

Data Usage Agreement (DUA)

- **Agreement to behave!**
  - Required for enclave access
  - COVID-research only
  - don't re-identify patients or share data
  - …

NIH National Center for Advancing Translational Sciences

https://covid.cd2h.org/account-instructions

Data Usage Agreement (DUA)

- **Agreement to behave!**
  - Required for enclave access
  - COVID-research only
  - don't re-identify patients or share data
  - …
- **Many institutions sign on behalf of their employees**
  - 370 so far: https://covid.cd2h.org/duas

NIH National Center for Advancing Translational Sciences

https://covid.cd2h.org/account-instructions

Data Usage Agreement (DUA)

- **Agreement to behave!**
  - Required for enclave access
  - COVID-research only
  - don't re-identify patients or share data
  - …
- **Many institutions sign on behalf of their employees**
  - 370 so far: https://covid.cd2h.org/duas
- **Individuals can sign as well**
  - As Citizen Scientists - directly with NIH NCATS (very limited data access)

NIH National Center for Advancing Translational Sciences

https://covid.cd2h.org/account-instructions

Other Rules of the Road

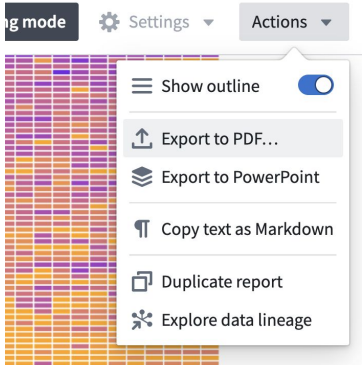
- **Community Guiding Principles:**
  - *Partnership, Inclusivity, Transparency, Reciprocity, Accountability, Security*
  - Also ethics, diversity and inclusion, resolving issues through community arbitration
  - …

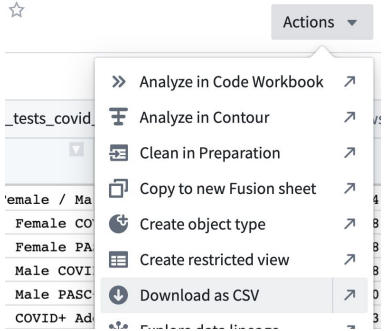
- **User Code of Conduct:**
  - Don't re-identify people, institutions, or populations
  - *No* screenshots or copy/paste data or results, no recording video calls that show data (Non-recorded video calls are OK if everyone has access)
  - …

- https://covid.cd2h.org/policy

Exporting and Publishing

- **Exporting Results**
  - CSV, PNG, PDF outputs possible
  - All results & summary tables must be reviewed and approved before export
  - No “cell sizes” less than 20
  - Mask or randomize site identifiers



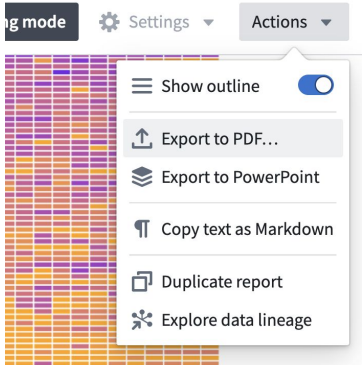


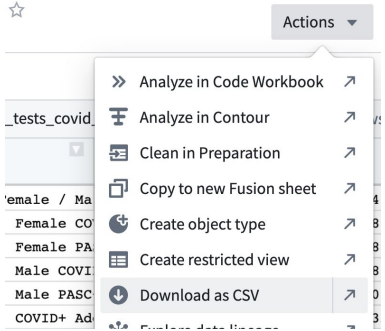
Exporting and Publishing

- **Exporting Results**
  - CSV, PNG, PDF outputs possible
  - All results & summary tables must be reviewed and approved before export
  - No “cell sizes” less than 20
  - Mask or randomize site identifiers

- **Publishing**
  - Draft manuscripts, presentations, posters should follow guidelines
  - Submit intent to publish for (non-scientific) review

- More at [Guide to N3C, Chapter 10]





National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Data Use Request (DUR)

NATIONAL CENTER FOR DATA TO HEALTH

Enclave Access:
Registration + DUA coverage confirmed

N3C Data Enclave (unite.nih.gov)





Notional synthetic (Level 0)


Level 2 (De-IDd)


Level 3 (LDS)

Data

National COVID Cohort Collaborative

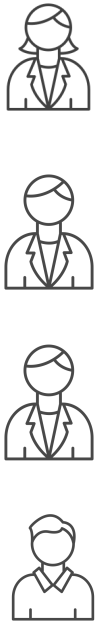
A program of NIH's National Center for Advancing Translational Sciences

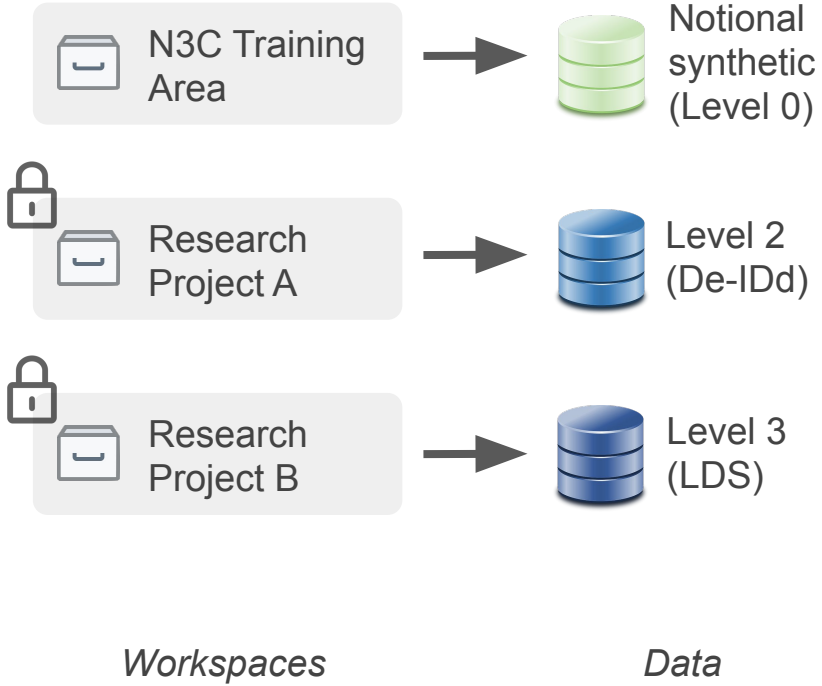
Data Use Request (DUR)

NATIONAL CENTER FOR DATA TO HEALTH

Enclave Access:
Registration + DUA coverage confirmed

N3C Data Enclave (unite.nih.gov)



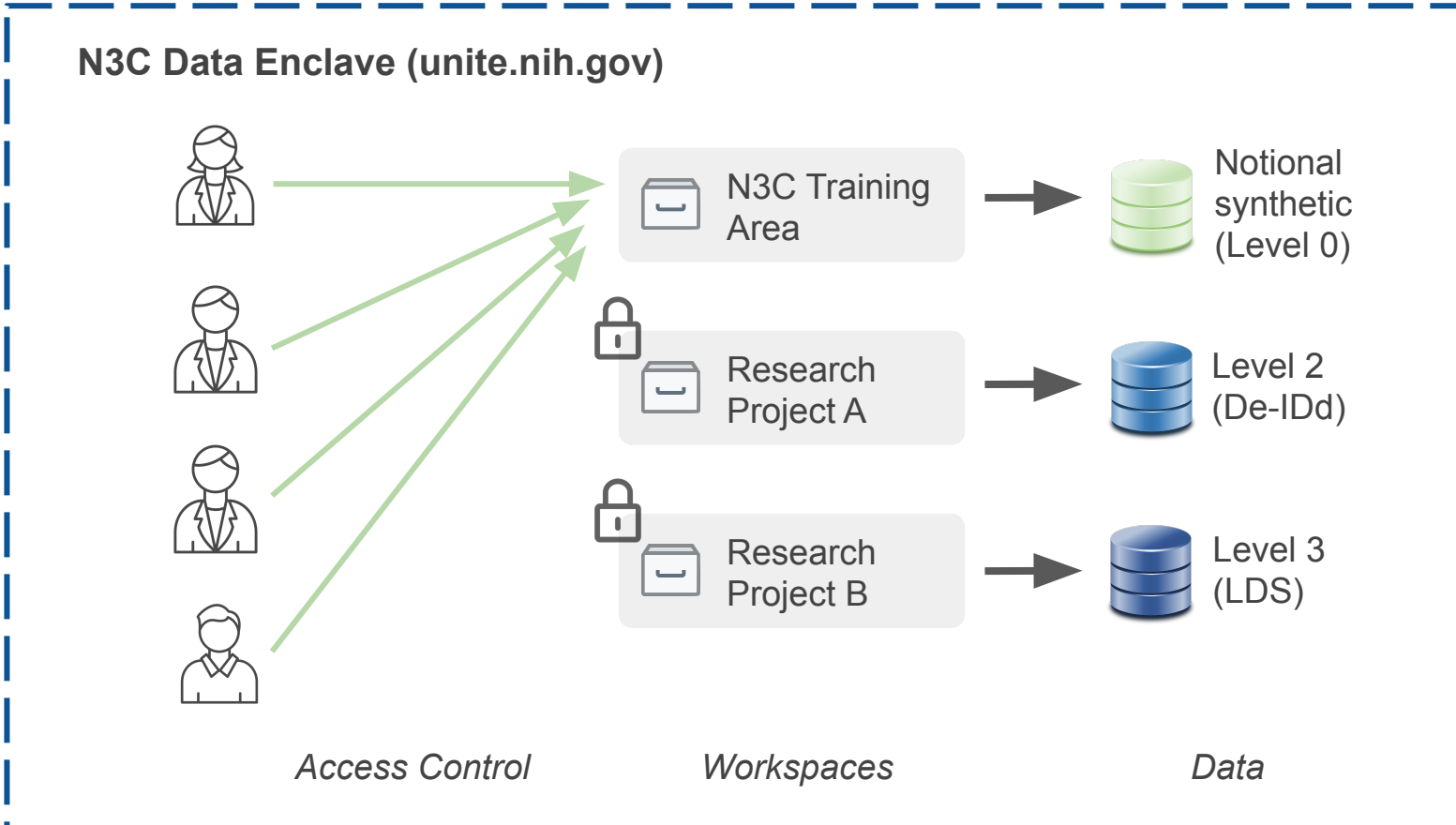


National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Data Use Request (DUR)

NATIONAL CENTER FOR DATA TO HEALTH

Enclave Access:
Registration + DUA coverage confirmed



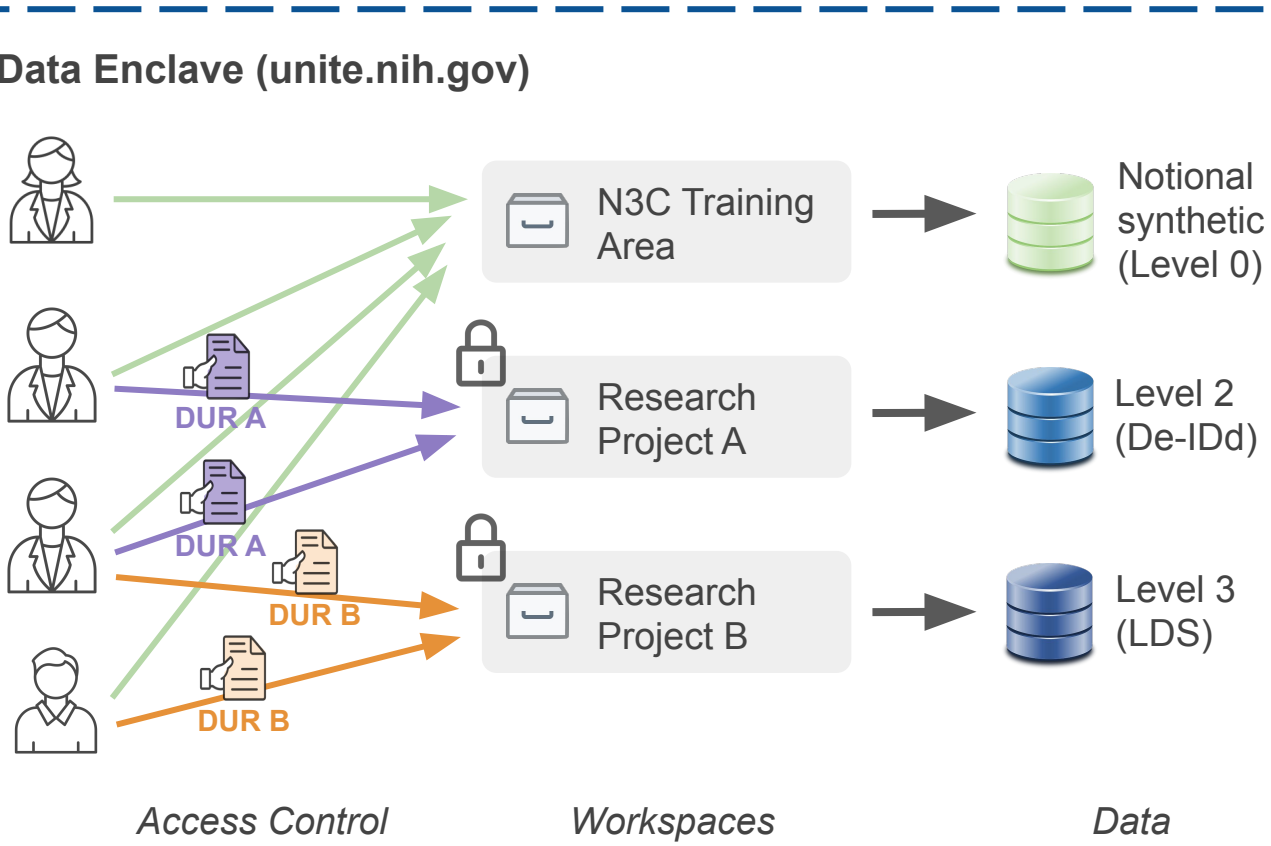
National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Data Use Request (DUR)

NATIONAL CENTER FOR DATA TO HEALTH

Enclave Access:
Registration + DUA coverage confirmed

Workspace Access:
Data **U**se **R**equest form



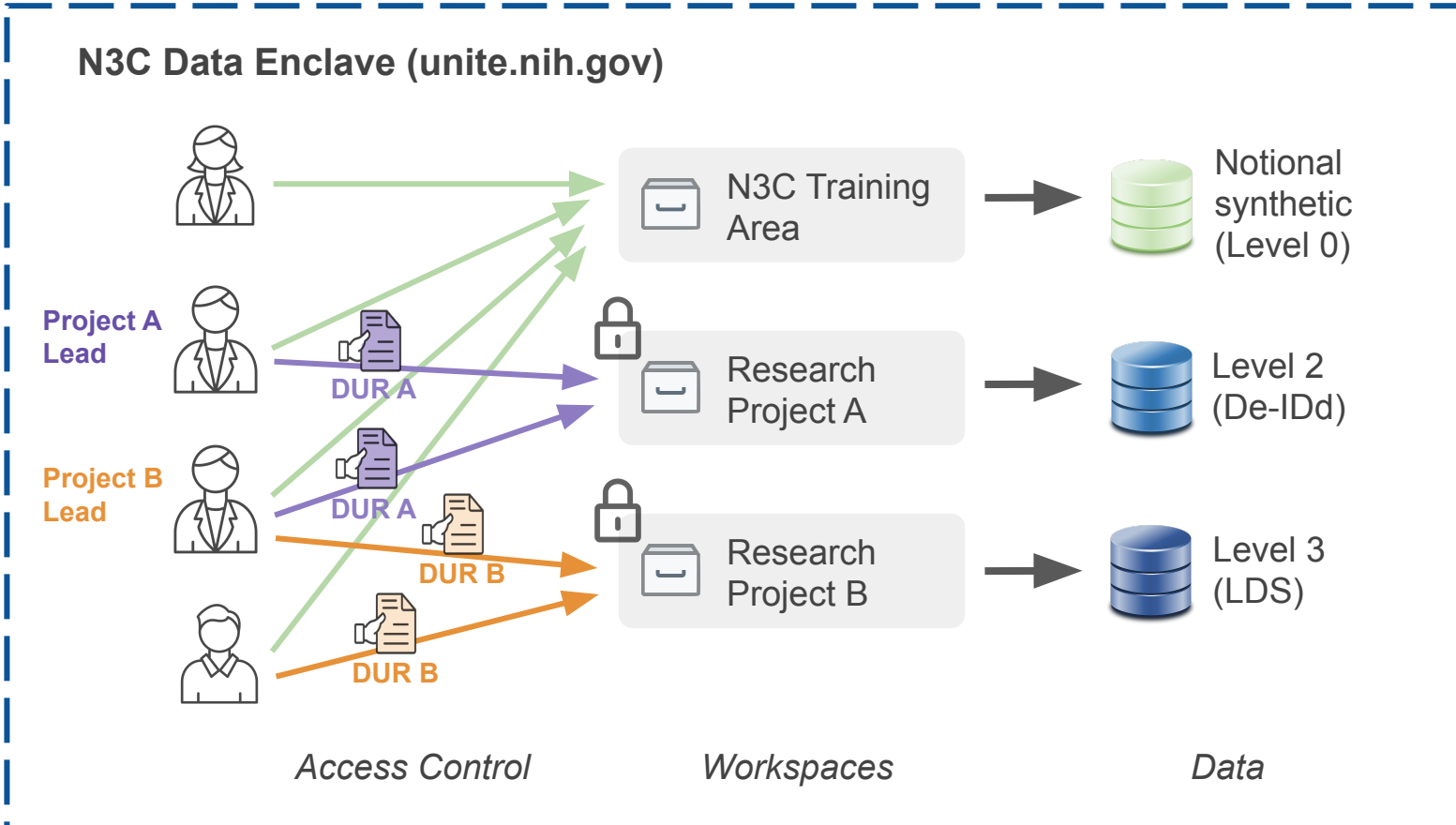
National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Data Use Request (DUR)

NATIONAL CENTER FOR DATA TO HEALTH

Enclave Access:
Registration + DUA coverage confirmed

Workspace Access:
Data **U**se **R**equest form



Enclave Tour!

National COVID Cohort Collaborative

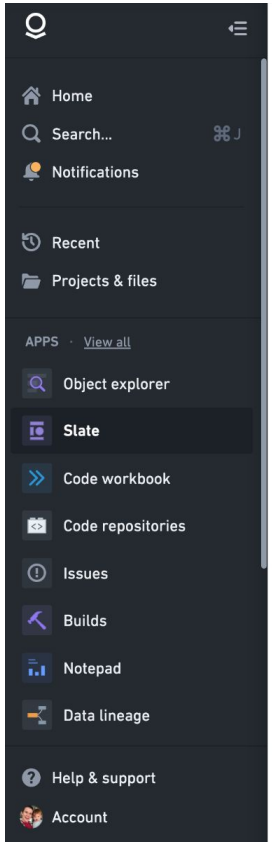
A program of NIH's National Center for Advancing Translational Sciences

Screenshots!

NATIONAL CENTER FOR DATA TO HEALTH

Files →

Analysis Tools →



Welcome to N3C, Shawn

Educational Resources

Training material | N3C Community Notes | Results Download

Training videos & docs

17,411,971 TOTAL N3C PATIENTS
6,745,241 CONFIRMED COVID-19 (+)
189,838 POSSIBLE COVID-19 (+)
77 SITES
21.7b TOTAL ROWS

N3C Cohort Definition
View detailed description of patient-selection criteria for N3C

Phenotype Explorer
Explore demographics and comorbidities by subcohorts

Administrative Resources Project & Workspace Management

Data Use Request (DUR) | My Projects (DURs) | Explore Projects (DURs)

Domain Teams | COVID Publications | N3C Administrative FAQ

Public Health Proposals | File Admin Ticket

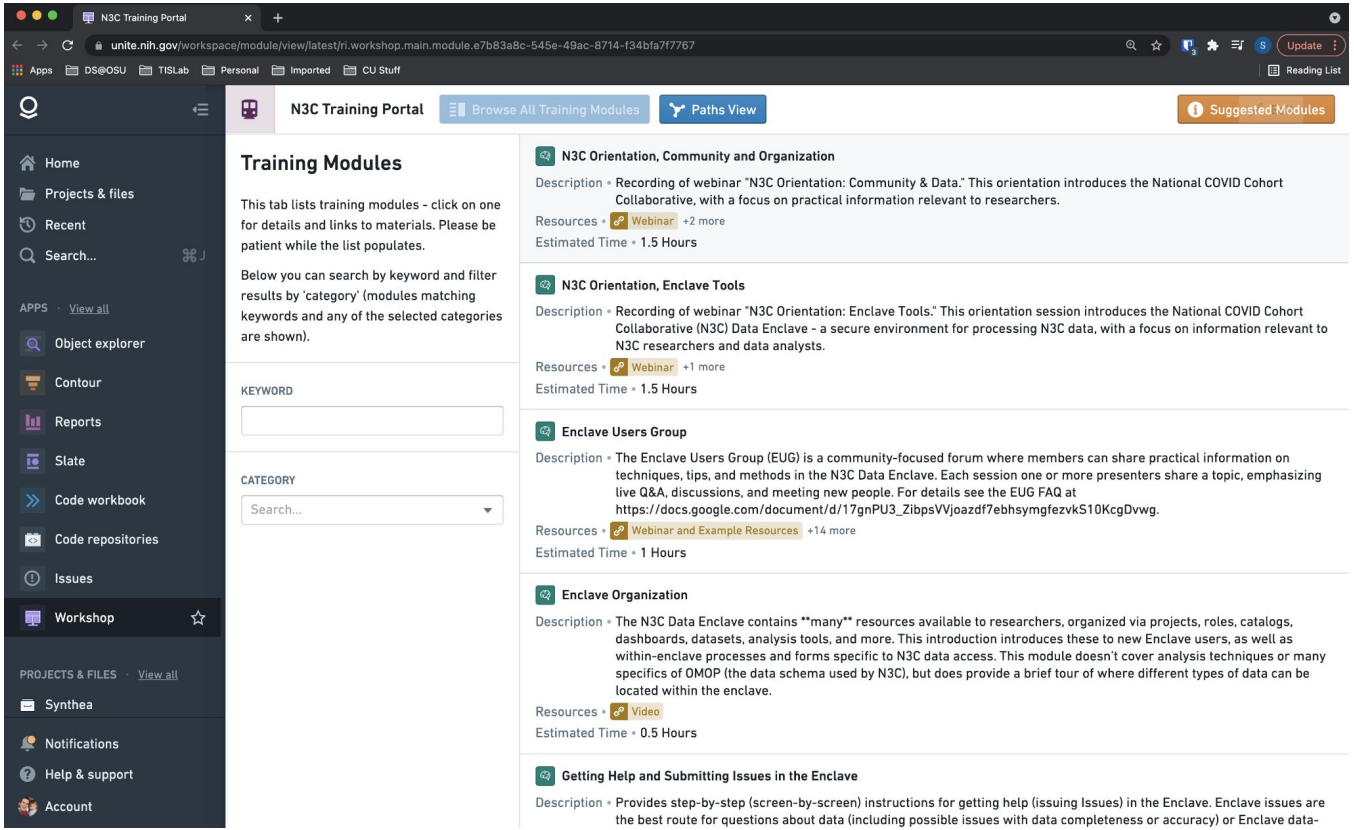
National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

Screenshots!

NATIONAL CENTER FOR DATA TO HEALTH

Training Portal



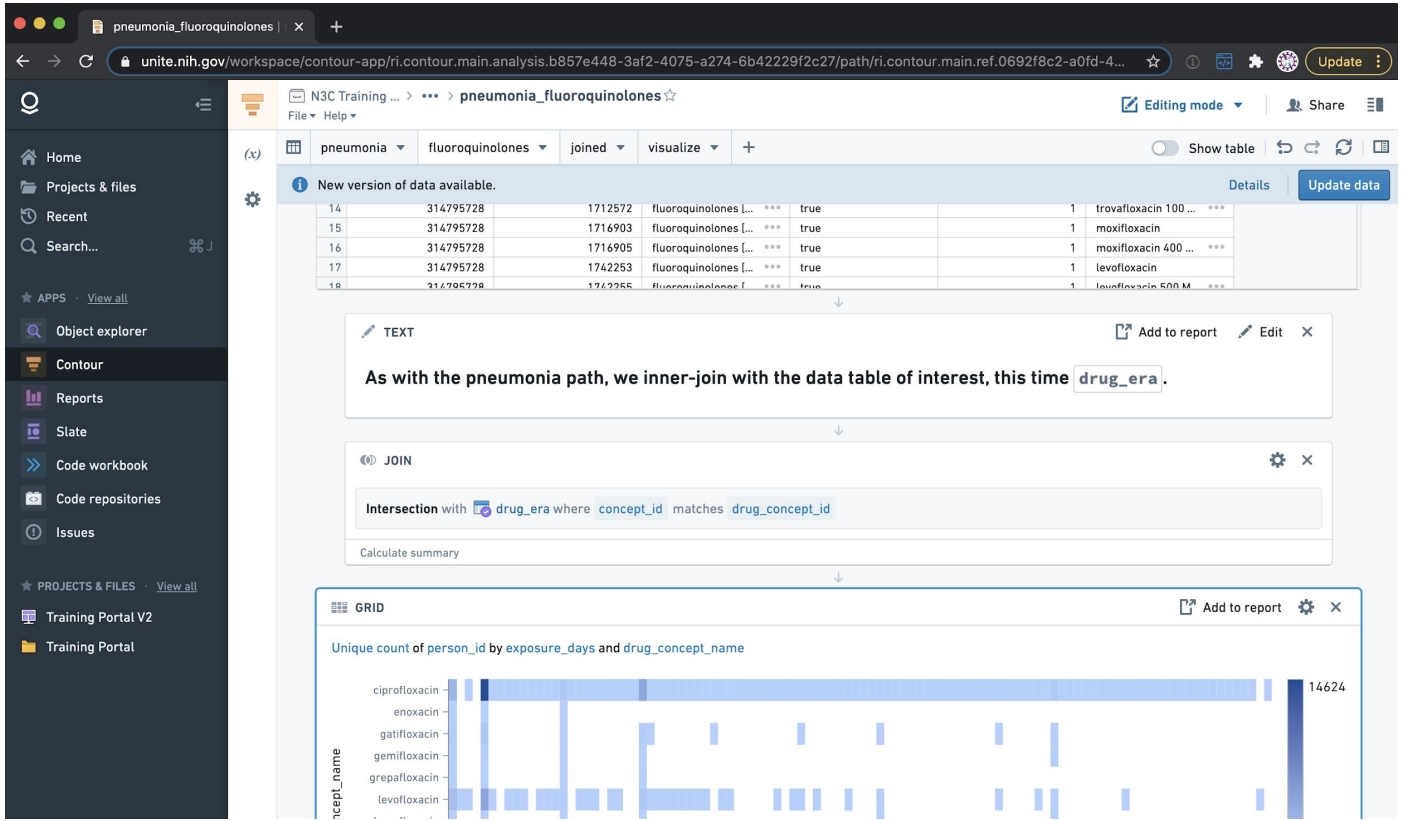
National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

Screenshots!

NATIONAL CENTER FOR DATA TO HEALTH

Contour: GUI for Data Wrangling



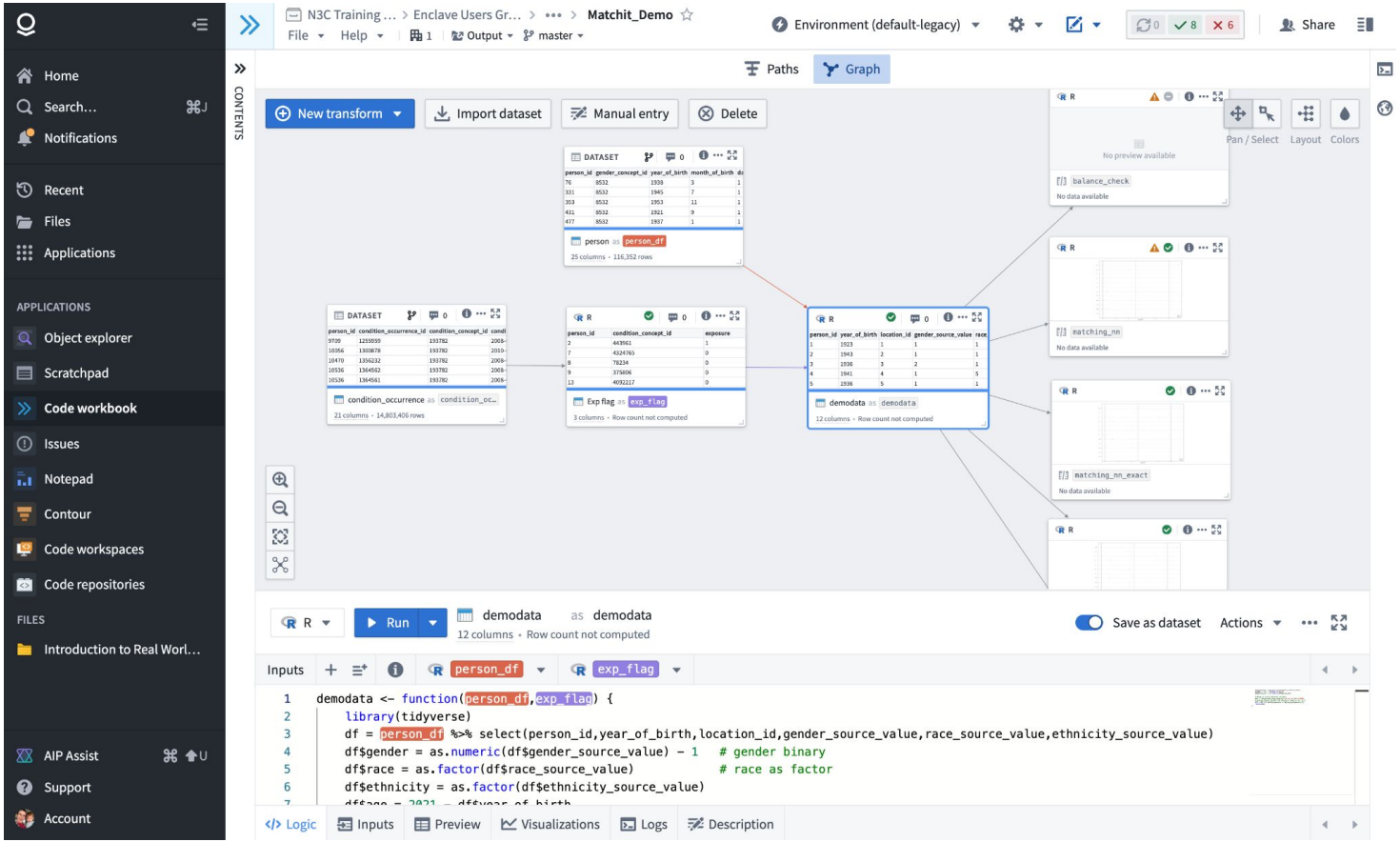
National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

Screenshots!

NATIONAL CENTER FOR DATA TO HEALTH

Code Workbook: SQL + Python + R



- **Wrap-up notes:**
  - Assignment 1 is on GitHub
    - https://github.com/National-COVID-Cohort-Collaborative/short-course-2024-january
    - Explore the enclave - *need enclave access*
    - One part requires you to have the correct permissions
      - Some still need to be fixed - if you have issues email shawn@tislab.org or sharon.patrick@hsc.wvu.edu
  - Support resources:
    - *Thursday* N3C office hours: https://covid.cd2h.org/support
    - Join the Slack workspace! https://cd2h.slack.com
      - Requires providing your slack-connected email during N3C registration
      - Class-specific channel coming soon
    - Sharon Patrick has kindly offered to be our central point of contact: sharon.patrick@hsc.wvu.edu