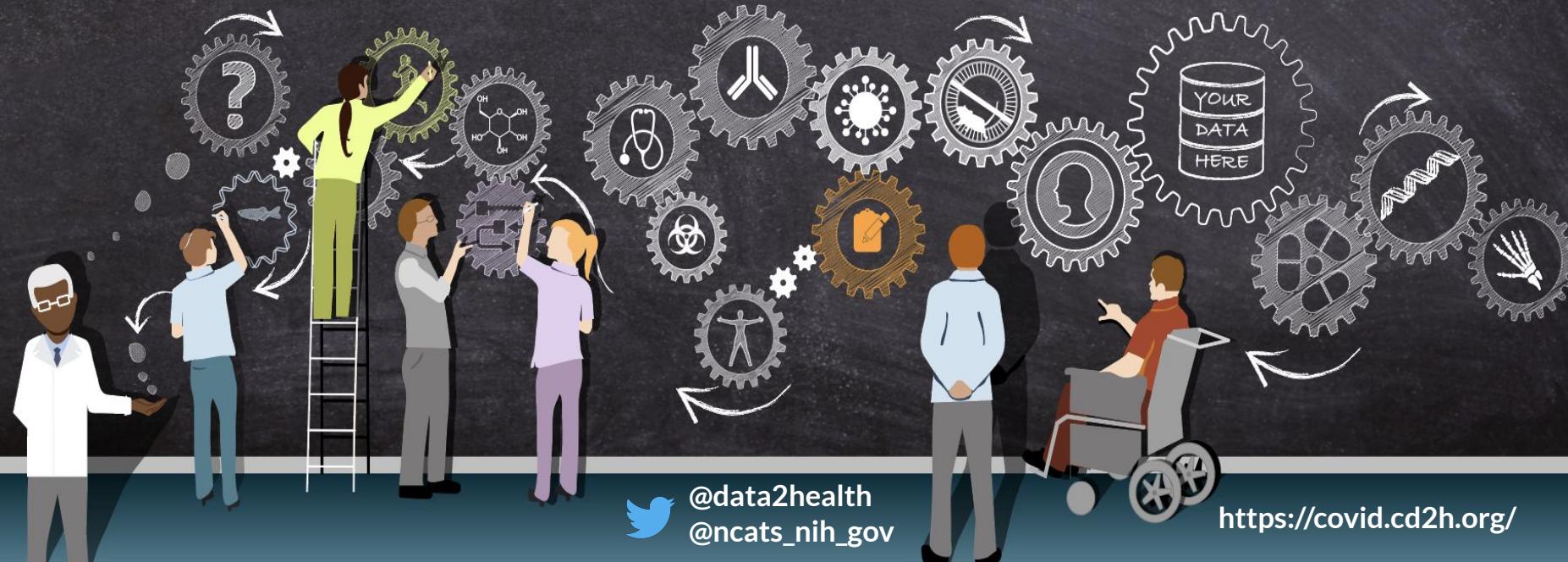


N3C Overview

Education and Training Domain Team Short Course, Spring 2024

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



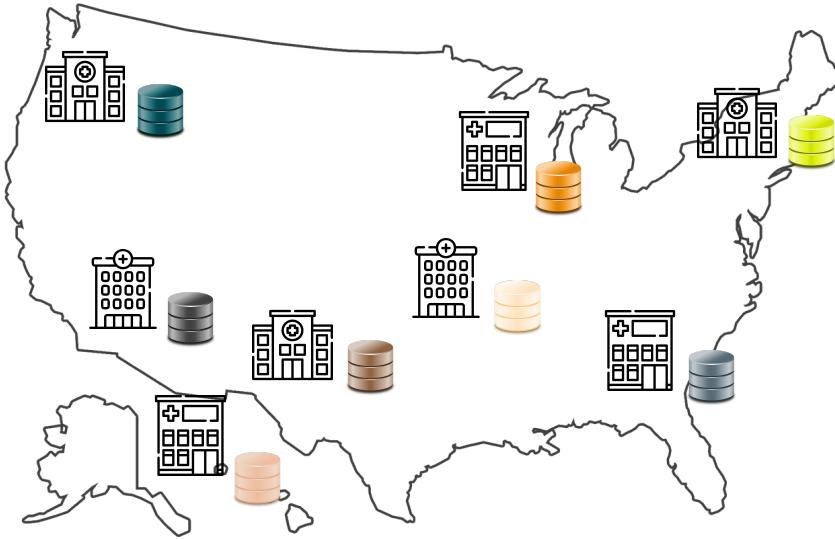
@data2health
@ncats_nih_gov

<https://covid.cd2h.org/>



National
COVID
Cohort
Collaborative

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



National COVID Cohort Collaborative

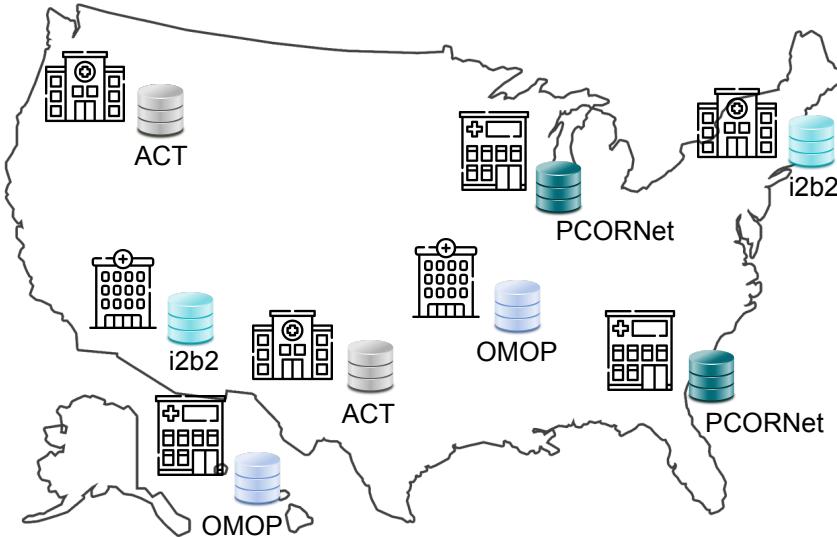


NATIONAL CENTER
FOR DATA TO HEALTH



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



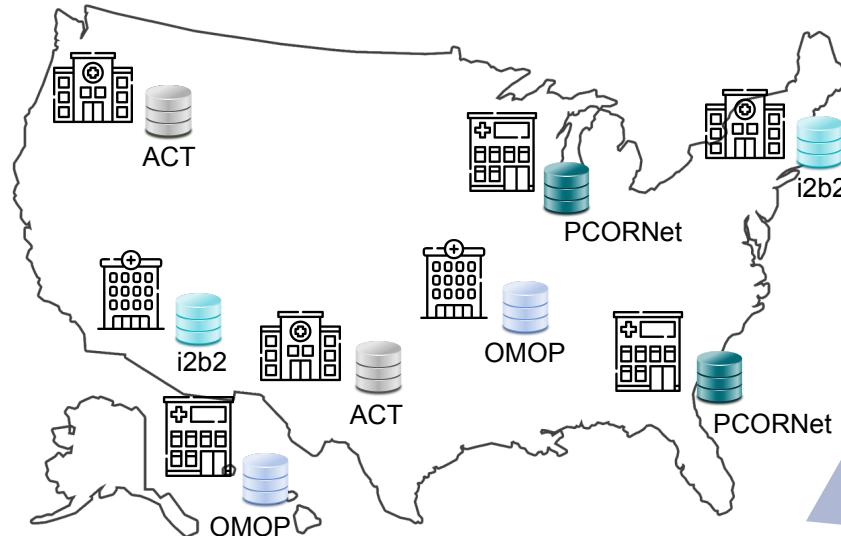
National
COVID
Cohort
Collaborative

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

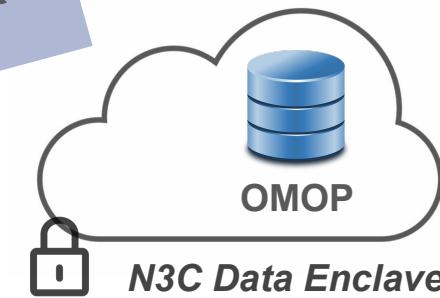
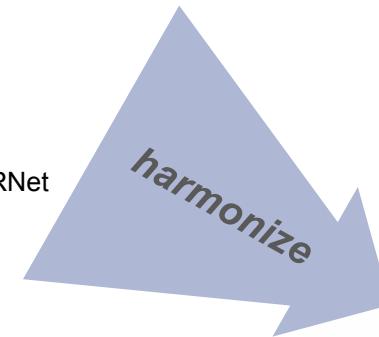


NATIONAL CENTER
FOR DATA TO HEALTH

National COVID Cohort Collaborative



Data Partners





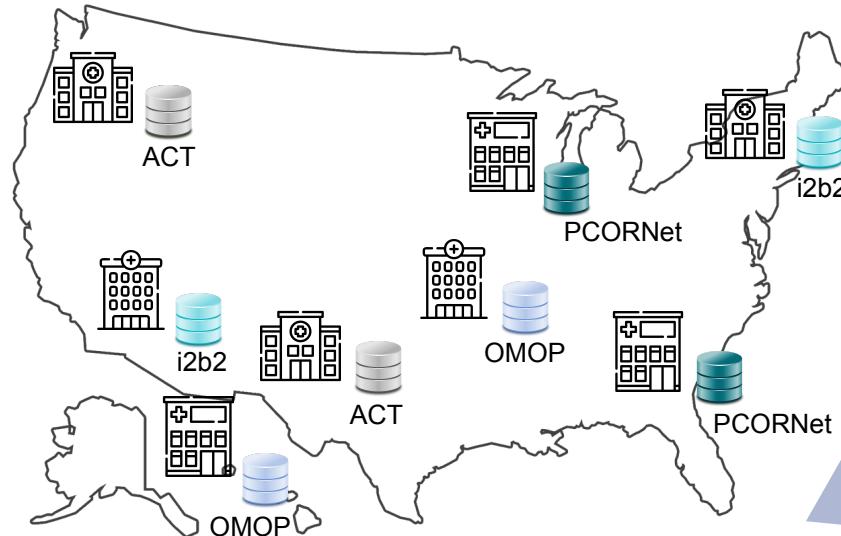
National
COVID
Cohort
Collaborative

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

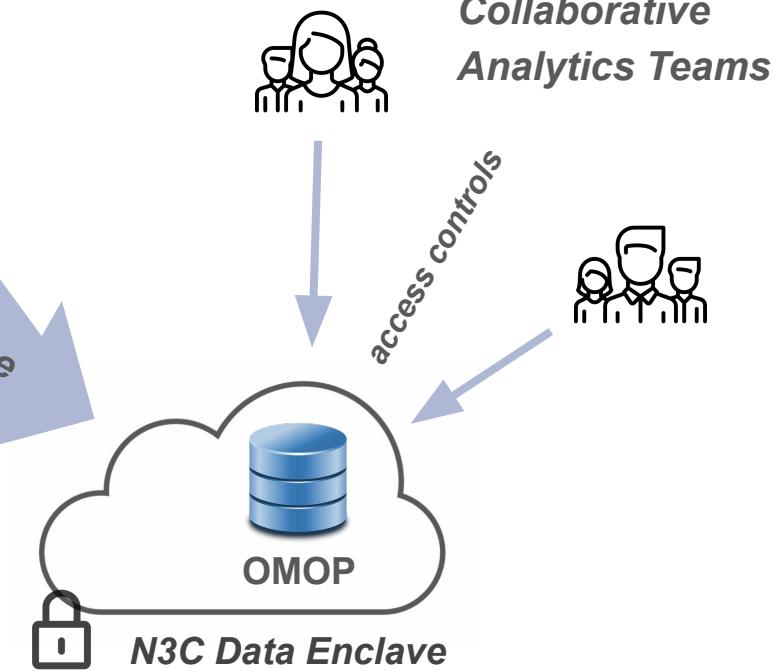
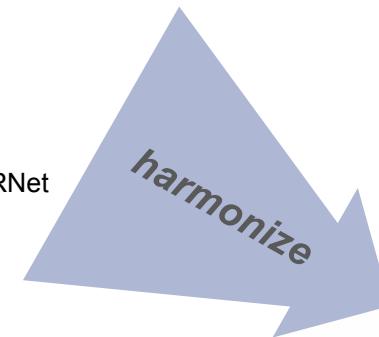


NATIONAL CENTER
FOR DATA TO HEALTH

National COVID Cohort Collaborative



Data Partners





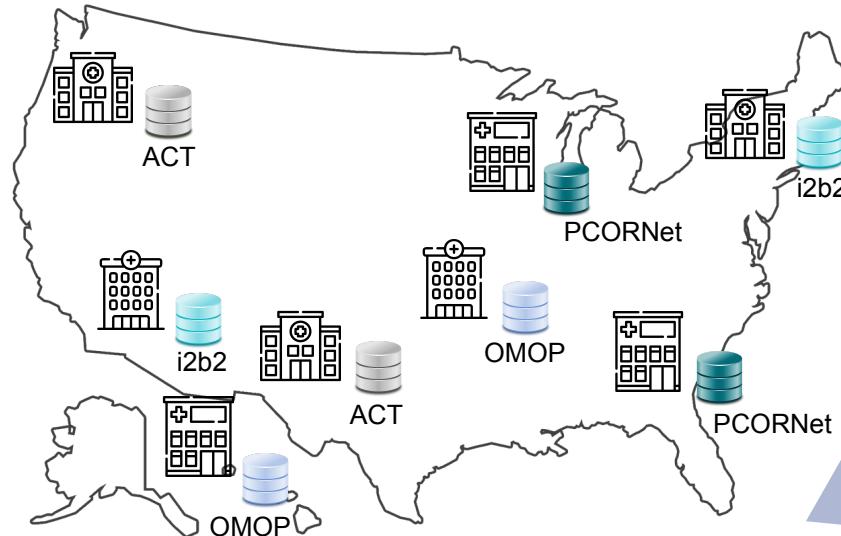
National
COVID
Cohort
Collaborative

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

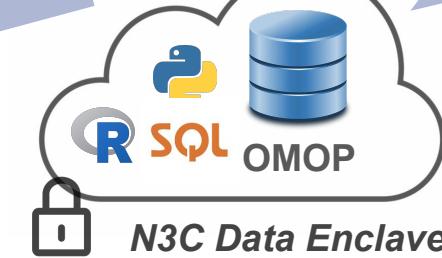
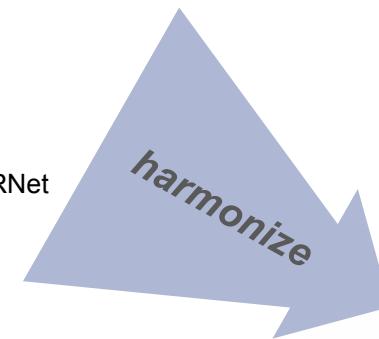


NATIONAL CENTER
FOR DATA TO HEALTH

National COVID Cohort Collaborative



Data Partners



N3C Data Enclave



*Collaborative
Analytics Teams*





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

👤 17,411,971
TOTAL N3C PATIENTS

👤 6,745,241
CONFIRMED COVID-19 (+)

👤 189,838
POSSIBLE COVID-19 (+)

🏢 77
SITES

蓄 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

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

👤 **17,411,971**

TOTAL N3C PATIENTS

👤 **6,745,241**

CONFIRMED COVID-19 (+)

👤 **189,838**

POSSIBLE COVID-19 (+)

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

🏢 **77**

SITES

蓄 **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

👤 **17,411,971**

TOTAL N3C PATIENTS

👤 **6,745,241**

CONFIRMED COVID-19 (+)

👤 **189,838**

POSSIBLE COVID-19 (+)

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

🏢 **77**

SITES

ddb **21.7b**

TOTAL ROWS

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



National
COVID
Cohort
Collaborative

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

Real World Data Standards many standards exist



NATIONAL CENTER
FOR DATA TO HEALTH



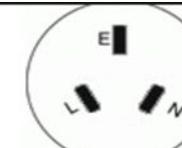
A (NEMA 1-15 USA 2 pin)



B (NEMA 5-15 USA 3 pin)



C (CEE 7/16)



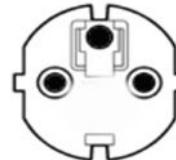
D (BS546 5 A version of Type M)



E (French)



F (CEE 7/4 "Schuko")



G (BS1363 Fused 13 A,
5 A and 3A also in common use)



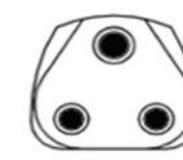
H (SI 32 Israel)



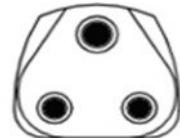
I (AS-3112 Argentina /
Australia / New Zealand)



J (SEV-1011 Switzerland)



K (SRAF 1962/DB Denmark)



L (CEI 23-16 Chile / Italy)



M (15 A version of Type D BS546)



N Italy



O Denmark



P Israel

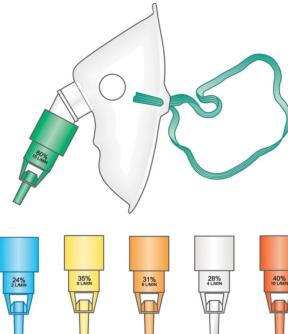
Stephanie Hong, Johns Hopkins



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

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



National
COVID
Cohort
Collaborative

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



NATIONAL CENTER
FOR DATA TO HEALTH

Harmonization

Data formats used by sites introduce heterogeneity



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



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



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

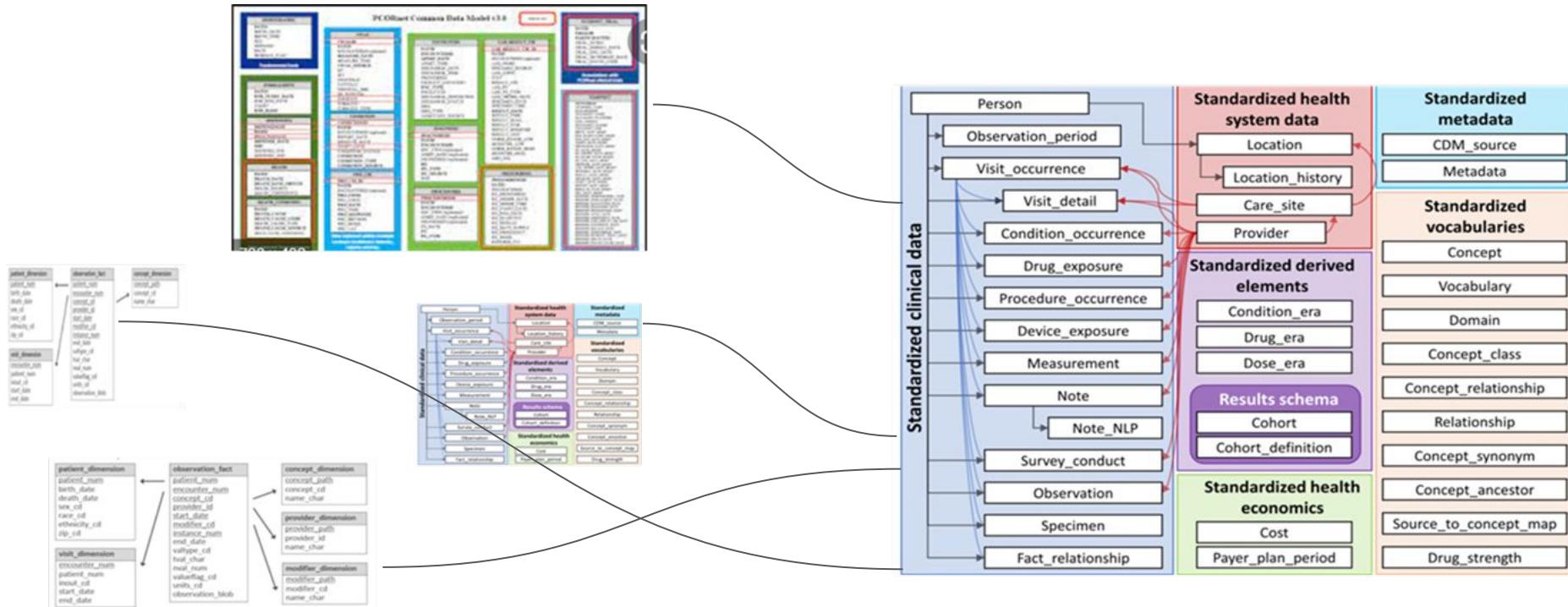


Death data supported?	Y
Death date required?	N
Death cause supported?	N
Discharge disposition supported?	N

Emily Pfaff, PhD, UNC



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



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!



National
COVID
Cohort
Collaborative

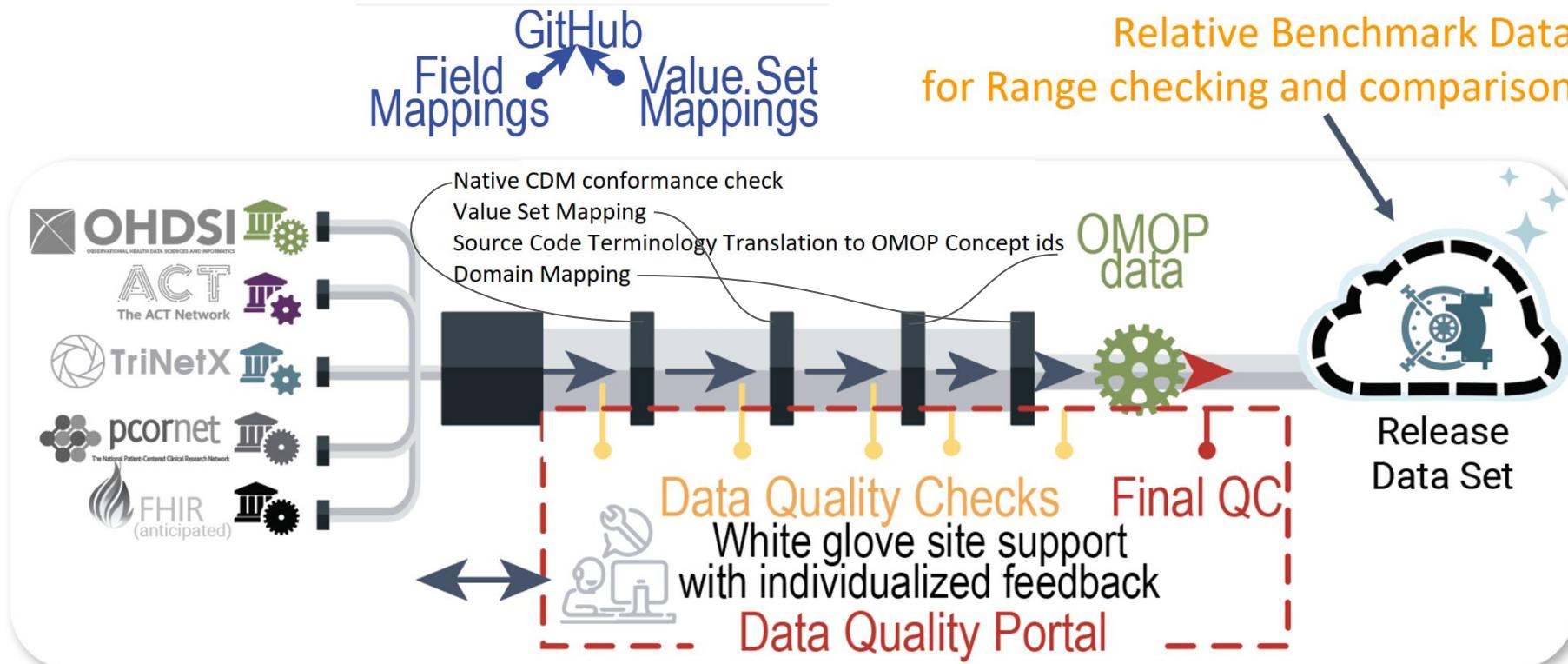
A program of NIH's National Center for
Advancing Translational Sciences

Limited Data Set (LDS) Data Ingestion Pipeline

Map Every Data Element Field and Its Value



NATIONAL CENTER
FOR DATA TO HEALTH





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



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	<u>N3C</u>



Limited Data Set (LDS) Data Ingestion Pipeline



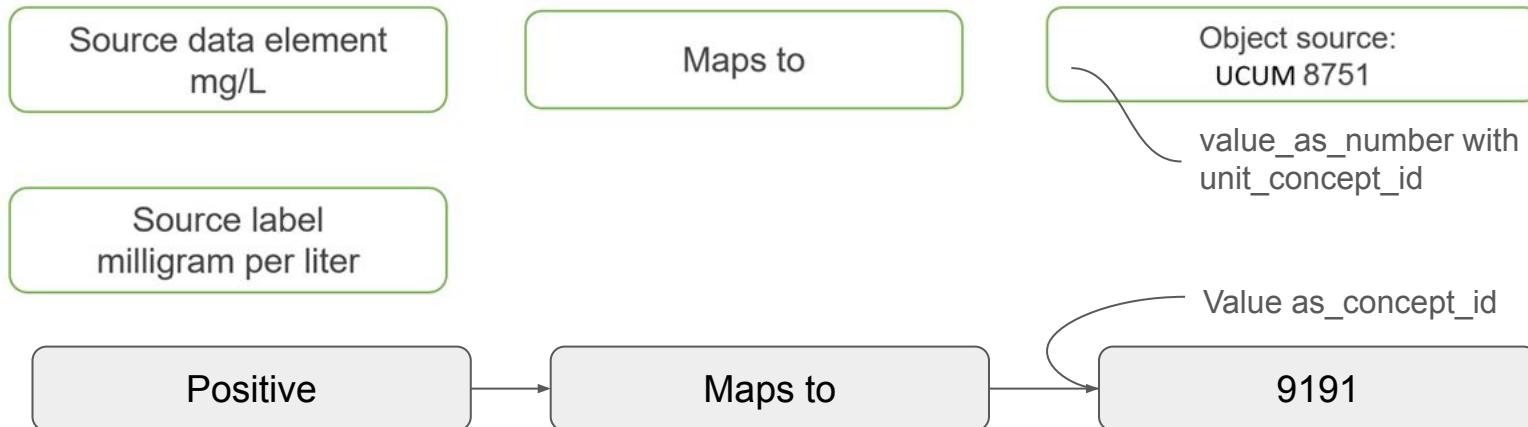
A unified framework manages stepwise ETL pipeline tasks for ingesting all incoming data. Each site's ingestion pipeline involves over two hundred data transformations, including native CDM checks, terminology translation (semantic and syntactic domain structure alignment), ID generation, and data health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



Measurement (numerical & categorical) and Units

Measurements are in numerical value with units or categorical value like “positive”, “negative” or “abnormal” - managed by value set mapping table

- Data Element mapping





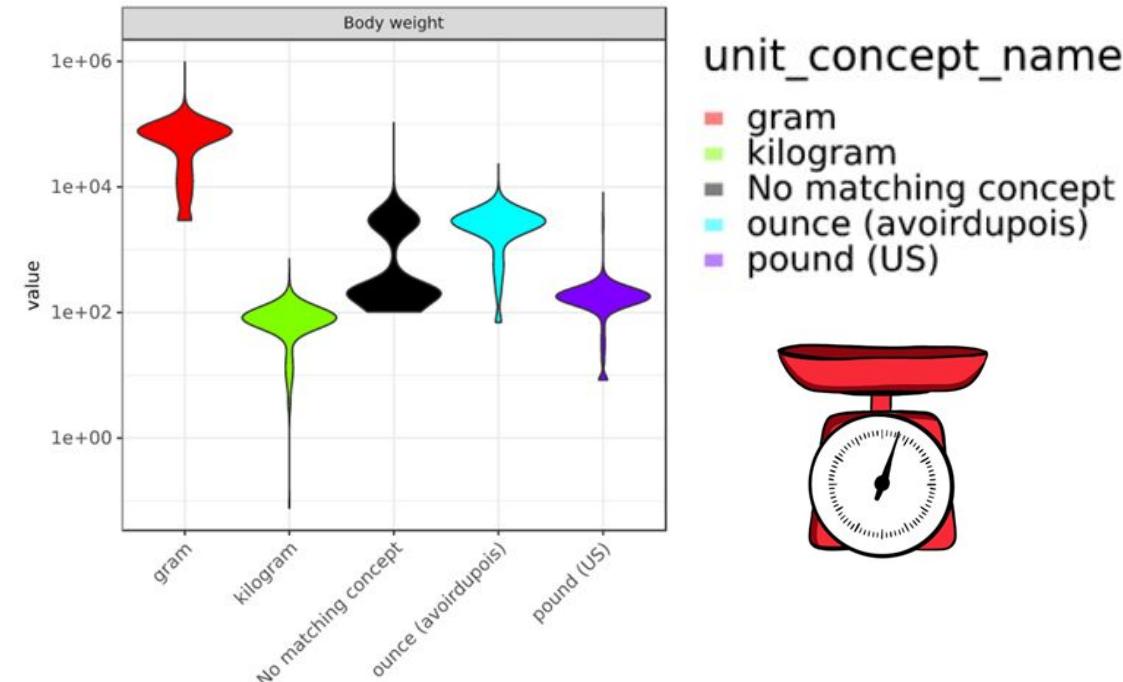
Measurement Unit Harmonization

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

Kilograms = Pounds /
2.20462

Kilograms = Ounces /
35.274

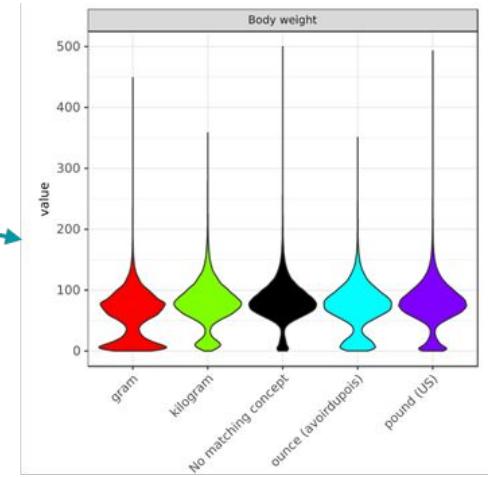
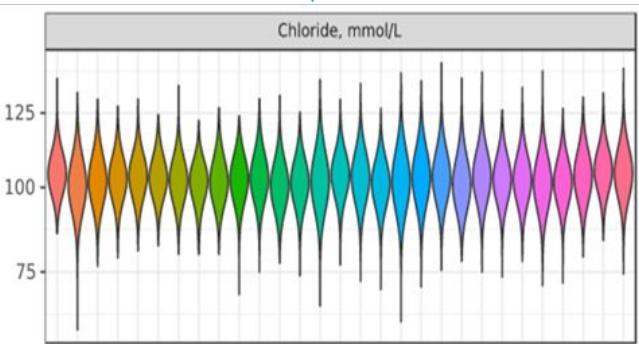
Kilograms = Grams / 1000





Measurement Unit Harmonization

- 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**!

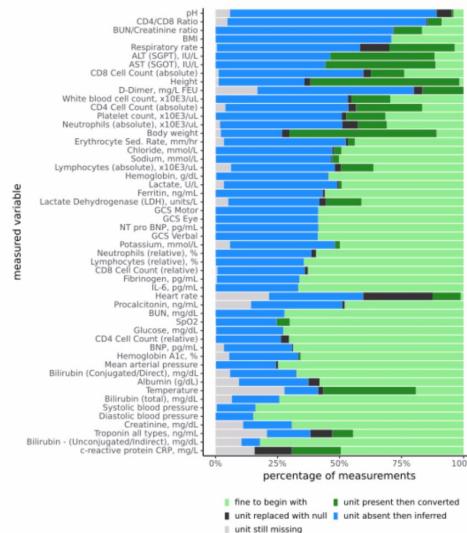




Unit Harmonization

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](https://doi.org/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

 **21,725,331**

TOTAL N3C PATIENTS

 **8,472,961**

CONFIRMED COVID-19 (+)

 **220,902**

POSSIBLE COVID-19 (+)

 **83**

SITES

 **30.8b**

TOTAL ROWS

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**

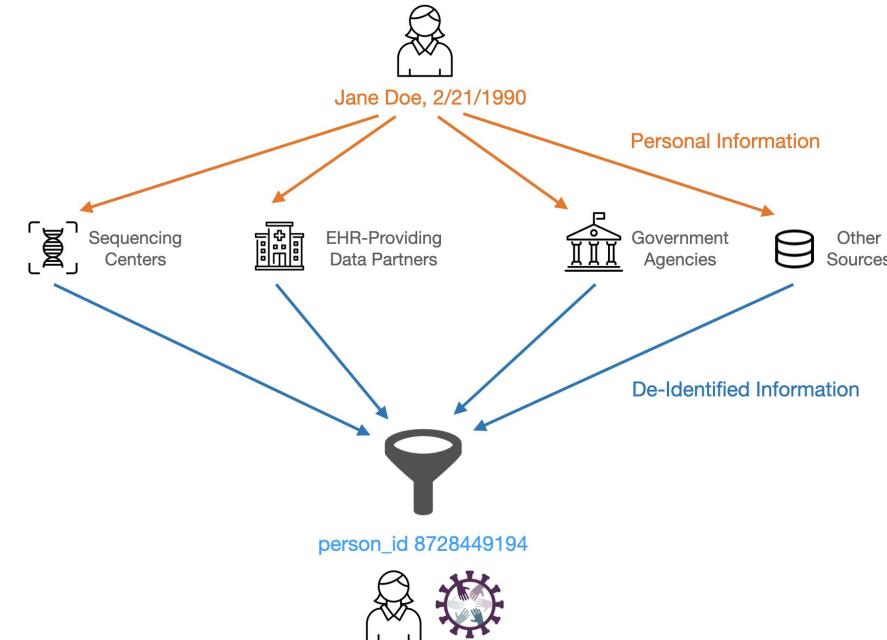
Private and government sources help fill EHR gaps

- **Viral Variants**

Sequenced variant (delta, omicron, ...)

- **CMS (Medicare/Medicaid) Billing Data**

Currently only for COVID-diagnosed medicare pts.

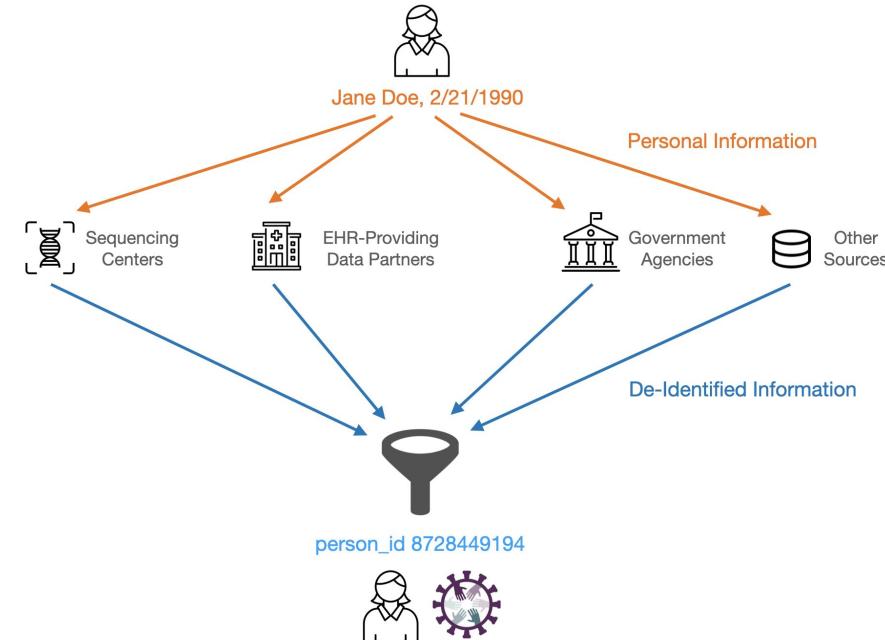




Privacy-Preserving Record Linkage

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>





National
COVID
Consort
Collaborative

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

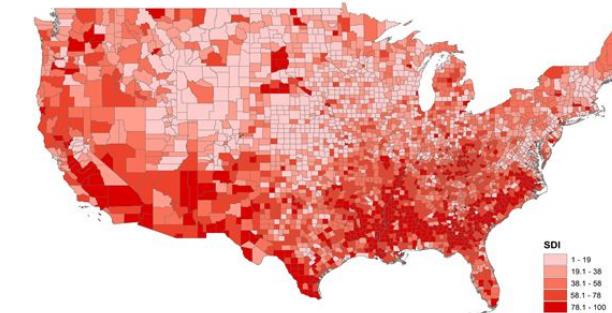


NATIONAL CENTER
FOR DATA TO HEALTH

“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)



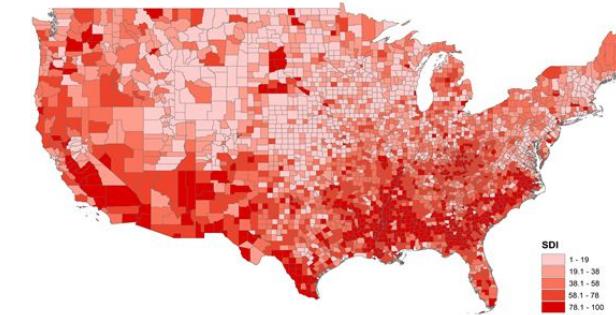
Social Deprivation Index, Robert Graham Center



“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>



Social Deprivation Index, Robert Graham Center

Agreements, Policies, Procedures

Shawn O'Neil, CU Anschutz



National
COVID
Cohort
Collaborative

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

Data Usage Agreement (DUA)



NATIONAL CENTER
FOR DATA TO HEALTH

- **Agreement to behave!**

- Required for enclave access
- COVID-research only
- don't re-identify patients or share data
- ...



National Center
for Advancing
Translational Sciences



National
COVID
Cohort
Collaborative

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

Data Usage Agreement (DUA)



NATIONAL CENTER
FOR DATA TO HEALTH

- **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>



National Center
for Advancing
Translational Sciences



National
COVID
Cohort
Collaborative

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

Data Usage Agreement (DUA)



NATIONAL CENTER
FOR DATA TO HEALTH

- **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)



National Center
for Advancing
Translational Sciences



Other Rules of the Road

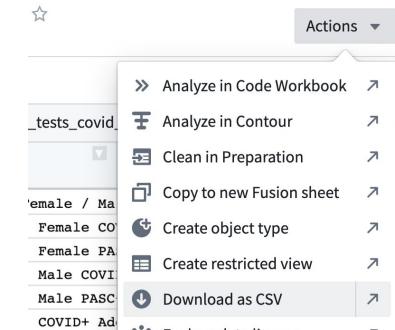
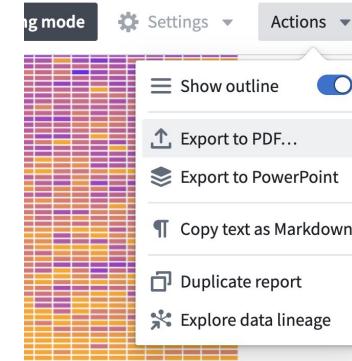
- **Community Guiding Principles:**
 - *Partnership, Inclusivity, Transparency, Reciprocity, Accountability, Security*
 - Also ethics, diversity and inclusion, resolving issues through community arbitration
 - ...
- **User Code of Conduct:**
 - Don't re-identify people, institutions, or populations
 - 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

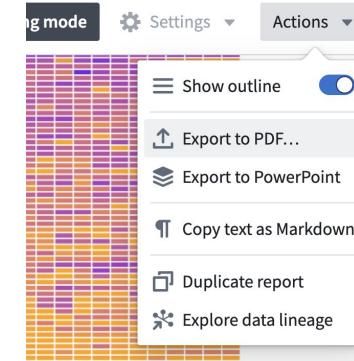




Exporting and Publishing

• Exporting Results

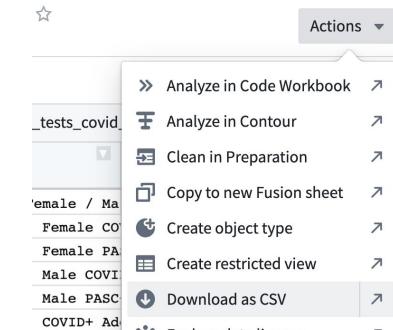
- CSV, PNG, PDF outputs possible
- All results & summary tables must be reviewed and approved before export
- No “cell sizes” less than 20
- Mask or randomize site identifiers



• 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



NATIONAL CENTER
FOR DATA TO HEALTH

Data Use Request (DUR)

Enclave Access:
Registration + DUA
coverage confirmed

N3C Data Enclave (unite.nih.gov)



Notional
synthetic
(Level 0)



Level 2
(De-IDd)



Level 3
(LDS)

Data



Data Use Request (DUR)

Enclave Access:
Registration + DUA
coverage confirmed

N3C Data Enclave (unite.nih.gov)



N3C Training
Area



Research
Project A



Research
Project B



Notional
synthetic
(Level 0)



Level 2
(De-IDd)



Level 3
(LDS)

Workspaces

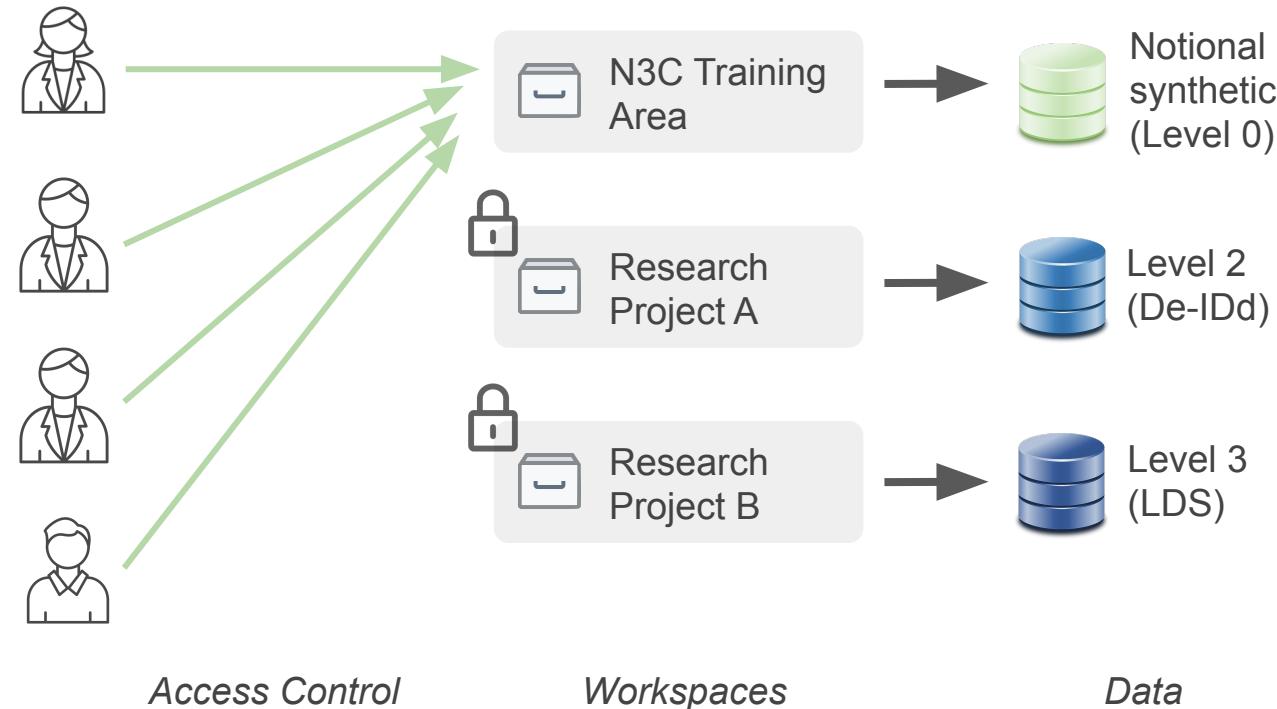
Data



Data Use Request (DUR)

Enclave Access:
Registration + DUA
coverage confirmed

N3C Data Enclave (unite.nih.gov)

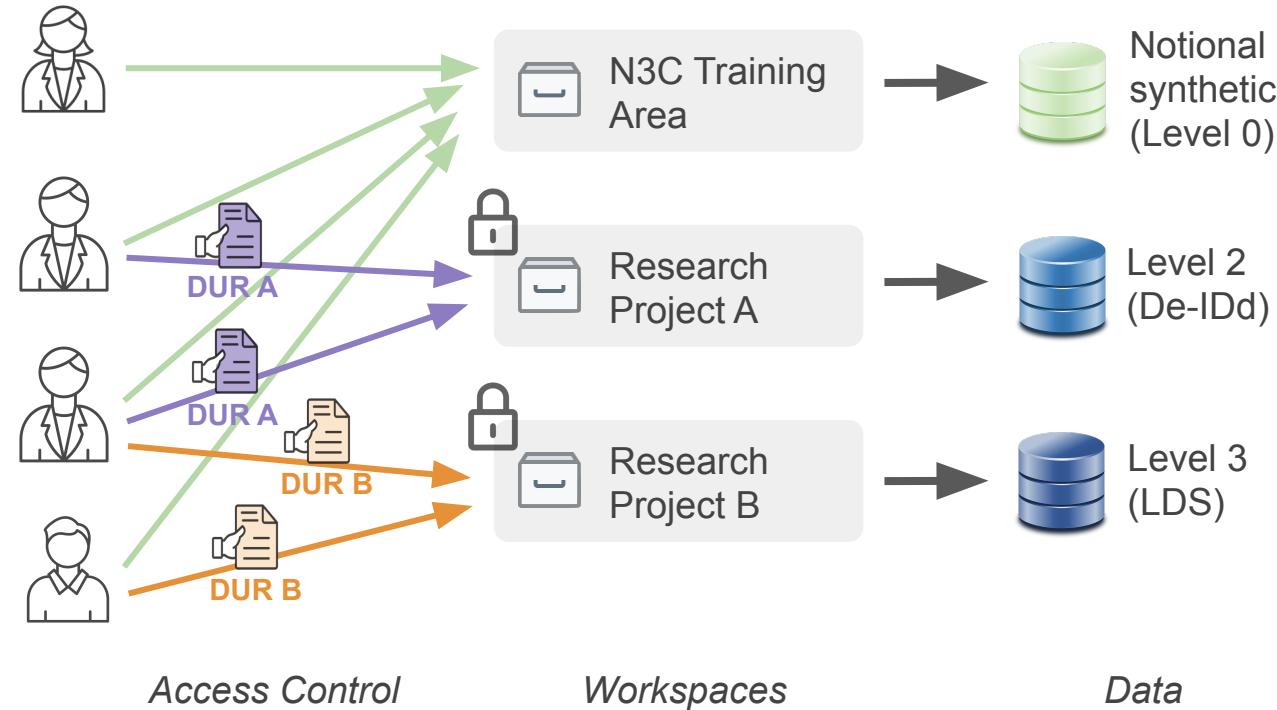




Data Use Request (DUR)

Enclave Access:
Registration + DUA
coverage confirmed

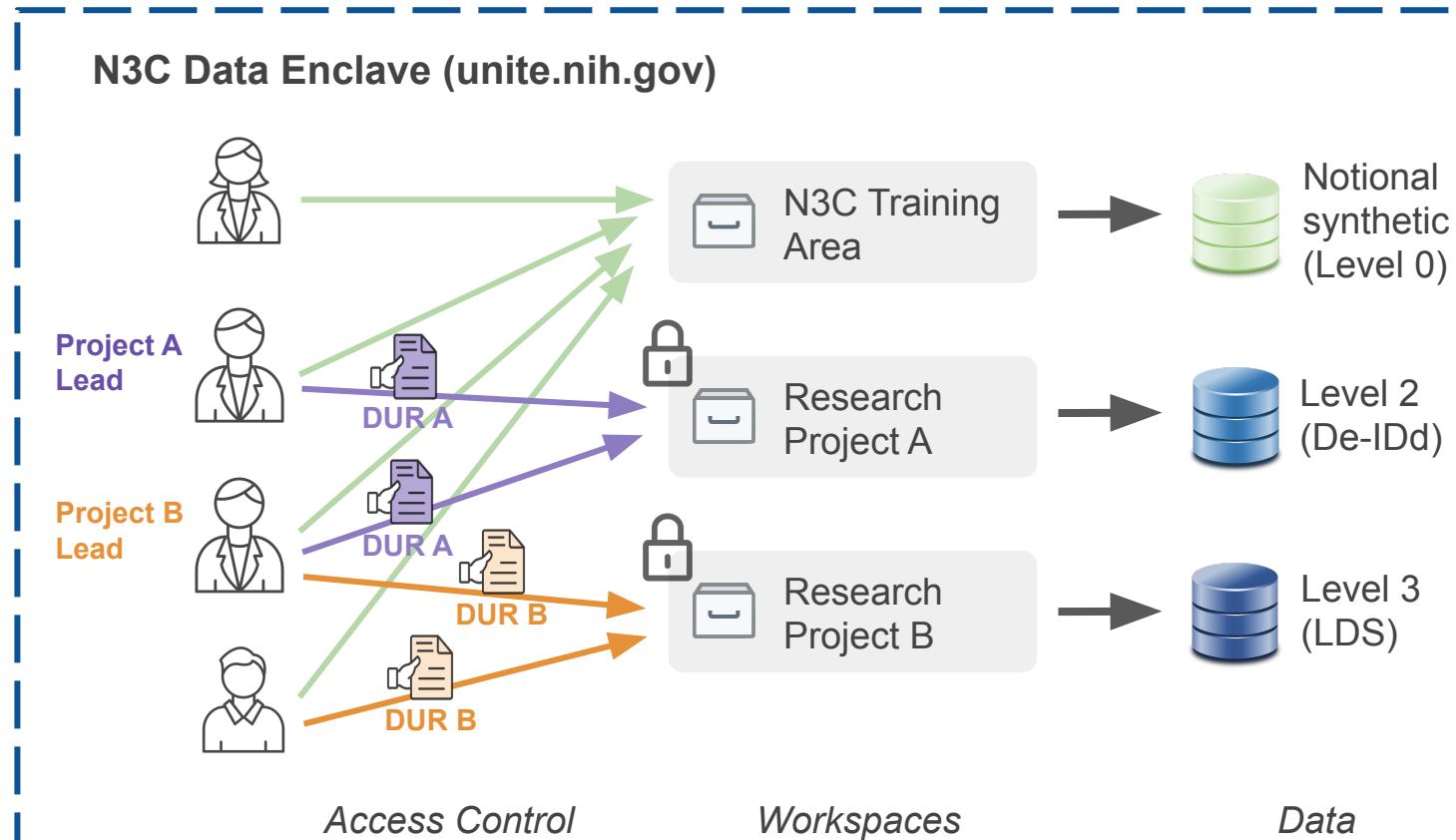
N3C Data Enclave (unite.nih.gov)





Data Use Request (DUR)

Enclave Access:
Registration + DUA
coverage confirmed



Enclave Tour!



National
COVID
Cohort
Collaborative

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



NATIONAL CENTER
FOR DATA TO HEALTH

Screenshots!

Files →

Analysis
Tools →

The screenshot shows the left sidebar of the N3C platform. It includes a user icon at the top, followed by a list of navigation items:

- Home
- Search...
- Notifications
- Recent
- Projects & files
- APPS · View all
- Object explorer
- State
- Code workbook
- Code repositories
- Issues
- Builds
- Notepad
- Data lineage
- Help & support
- Account

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



Screenshots!

Training Portal

The screenshot shows the N3C Training Portal interface. On the left is a dark sidebar with various navigation links: Home, Projects & files, Recent, Search..., View all, Object explorer, Contour, Reports, Slate, Code workbook, Code repositories, Issues, Workshop (selected), View all, Synthea, Notifications, Help & support, and Account.

The main content area has a header with tabs: N3C Training Portal (selected), Browse All Training Modules, Paths View, and Suggested Modules. Below the header is a section titled "Training Modules" with a sub-section "N3C Orientation, Community and Organization". It includes a description, resources (Webinar), and estimated time (1.5 Hours). There are also sections for "N3C Orientation, Enclave Tools", "Enclave Users Group", "Enclave Organization", and "Getting Help and Submitting Issues in the Enclave". Each section follows a similar structure with a description, resources (e.g., Webinar, Video), and estimated time.



Screenshots!

Contour: GUI for Data Wrangling

The screenshot shows the Contour application interface on a web browser. The left sidebar contains navigation links for Home, Projects & files, Recent, Search, APPS, Object explorer, Contour (selected), Reports, Slate, Code workbook, Code repositories, and Issues. Below these are sections for Training Portal V2 and Training Portal.

The main workspace displays a table titled "pneumonia_fluoroquinolones". The table has columns: ID, concept_id, exposure_days, drug_concept_name, and true. The data shows various fluoroquinolone drugs like trovafloxacin, moxifloxacin, and levofloxacin.

Below the table is a "TEXT" section containing the note: "As with the pneumonia path, we inner-join with the data table of interest, this time `drug_era`".

Under the "JOIN" section, it says "Intersection with `drug_era` where `concept_id` matches `drug_concept_id`".

The bottom section is a "GRID" visualization titled "Unique count of person_id by exposure_days and drug_concept_name". It shows a horizontal bar chart where each bar represents a unique drug name. The longest bar is for ciprofloxacin, with a value of 14624.



Code Workbook: SQL + Python + R

The screenshot displays a data analysis workspace interface, likely from a cloud-based platform like Jupyter Notebook or a similar environment. The top navigation bar shows the path: N3C Training ... > Enclave Users Gr... > Matchit_Demo. The environment is set to 'Environment (default-legacy)'. The main area features a 'Graph' view showing the flow of data between various datasets and R environments.

Graph View: The graph illustrates the data pipeline. It starts with a 'person_df' dataset (25 columns, 116,352 rows) which feeds into an 'exp_flag' environment (3 columns). This is followed by a 'demodata' environment (12 columns). The 'demodata' environment then feeds into two separate 'R' environments: 'matching_nm' and 'matching_nm_exact', both of which show 'No data available'.

Code Workbook View: On the left, the sidebar lists applications such as Object explorer, Scratchpad, Code workbook (which is currently selected), Issues, Notepad, Contour, Code workspaces, and Code repositories. The main workspace shows the following code in the 'Logic' tab:

```
1 demodata <- function(person_df,exp_flag) {  
2   library(tidyverse)  
3   df = person_df %>% select(person_id,year_of_birth,location_id,gender_source_value,race_source_value,ethnicity_source_value)  
4   df$gender = as.numeric(df$gender_source_value) - 1 # gender binary  
5   df$race = as.factor(df$race_source_value) # race as factor  
6   df$ethnicity = as.factor(df$ethnicity_source_value)  
7   df$age = 2021 - df$year_of_birth
```

The 'Inputs' tab at the bottom shows the inputs for the 'demodata' function: 'person_df' and 'exp_flag'.

- 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