National
COVID
Cohort
Collaborative

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

CTSA Site Data Scorecards

RECOVER
Researching COVID to Enhance Recovery

An Initiative Funded by the National Institutes of Health

Emily Pfaff, PhD, MS

Assistant Professor, UNC Chapel Hill School of Medicine

CTSA Steering Committee 3.13.2023

N3C Scorecards Overview

Sofia Dard

Tim Schwab

Chris Roeder

Abhishek Bhatia

Stephanie Hong

Bryan Laraway

Eric Kim

Maya Choudhury

James Cavallon

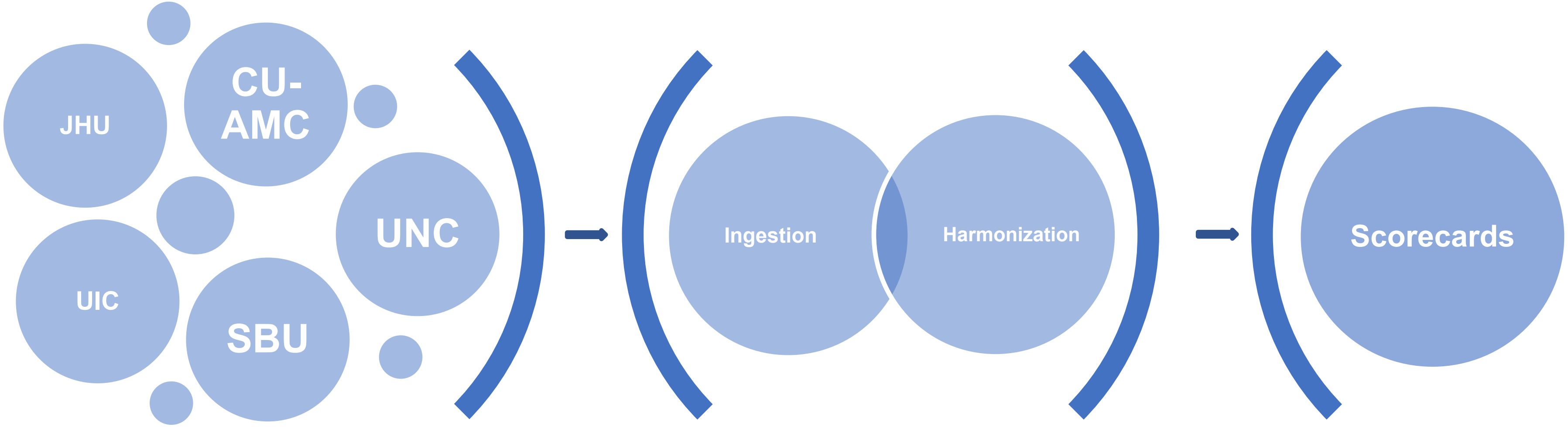
Kate Bradwell

DI&H team

& our beta testers

N3C Scorecard Team

N3C Workflow



18,151,603 TOTAL N3C PATIENTS

7,006,103 CONFIRMED COVID-19 (+)

196,564 POSSIBLE COVID-19 (+)

77 SITES

22.9b TOTAL ROWS

National
COVID
Cohort
Collaborative

NATIONAL CENTER
FOR DATA TO HEALTH

N3C Scorecards Goals

We hope to…

- Inform data partners about their data quality (highlight good & areas for improvement)
- Stop data quality regression & maintain data quality across subsequent payloads
- Provide perspective; benchmark set by comparing to other sites

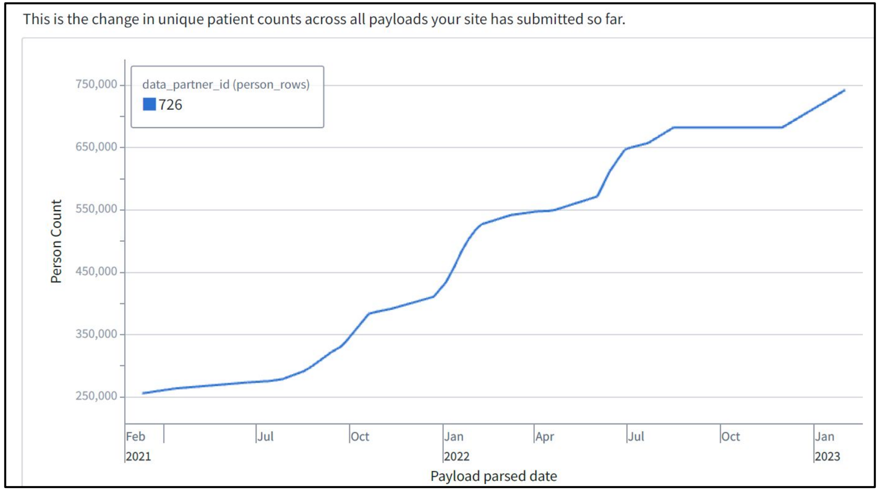
NIH National Center for Advancing Translational Sciences

National
COVID
Cohort
Collaborative

Overview

National
COVID
Cohort
Collaborative

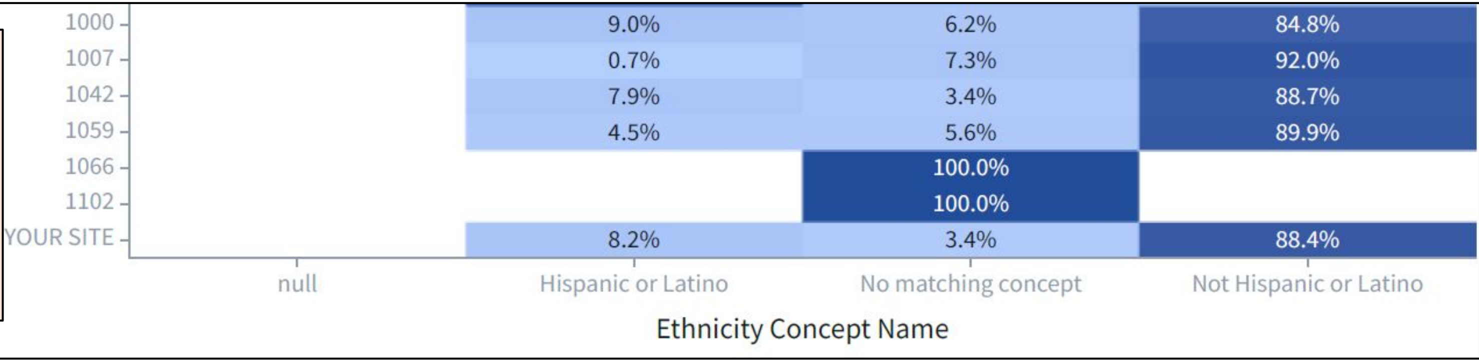
Demographics



This is the change in unique patient counts across all payloads your site has submitted so far.

Sites can ensure the trajectory of their COVID population increase looks reasonable (and catch where it is not).

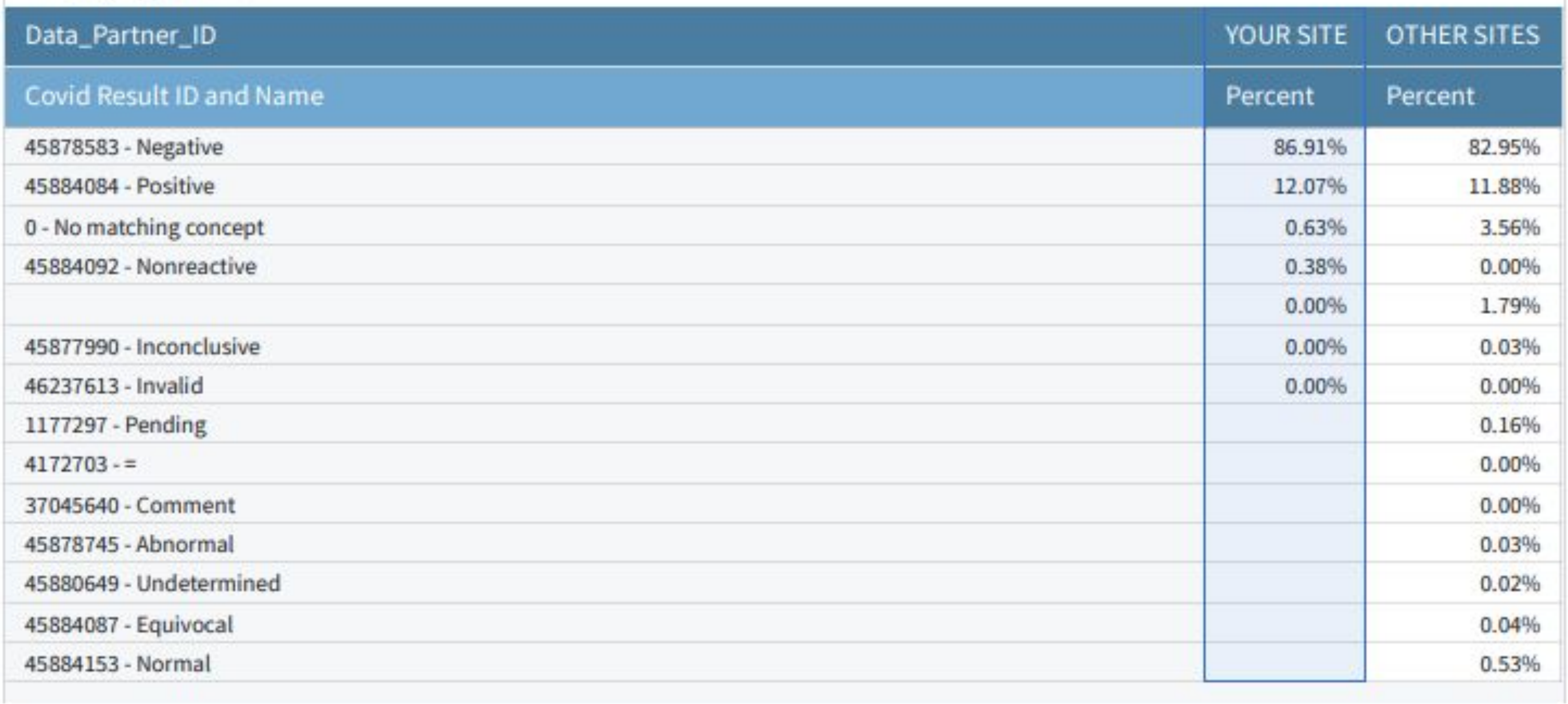
Sites can compare their demographic categories and proportions with peers, and note areas for improvement.



	null	Hispanic or Latino	No matching concept	Not Hispanic or Latino
1000		9.0%	6.2%	84.8%
1007		0.7%	7.3%	92.0%
1042		7.9%	3.4%	88.7%
1059		4.5%	5.6%	89.9%
1066			100.0%	
1102			100.0%	
YOUR SITE		8.2%	3.4%	88.4%

Ethnicity Concept Name

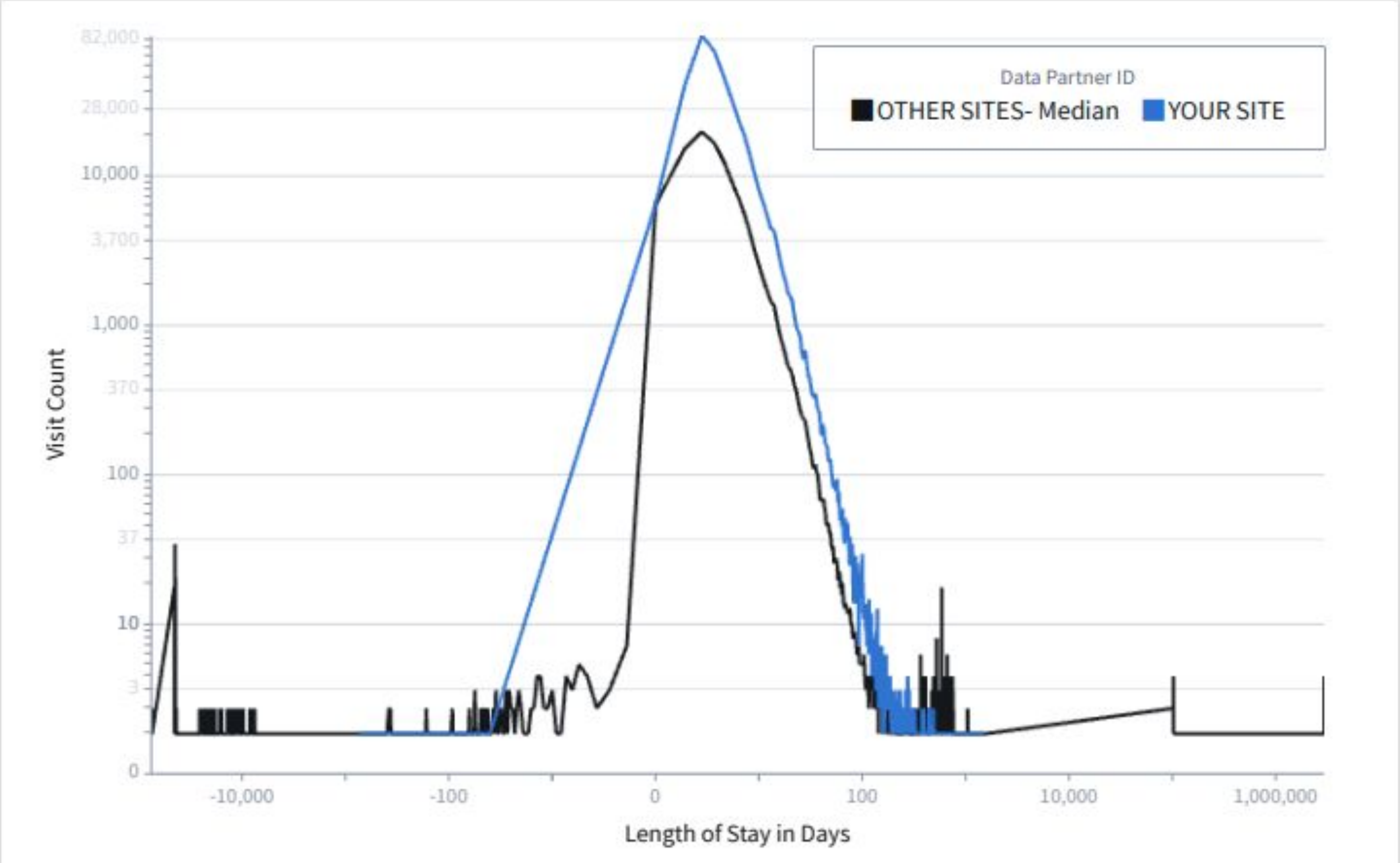
COVID-19 Metrics



Data_Partner_ID	YOUR SITE	OTHER SITES
Covid Result ID and Name	Percent	Percent
45878583 - Negative	86.91%	82.95%
45884084 - Positive	12.07%	11.88%
0 - No matching concept	0.63%	3.56%
45884092 - Nonreactive	0.38%	0.00%
	0.00%	1.79%
45877990 - Inconclusive	0.00%	0.03%
46237613 - Invalid	0.00%	0.00%
1177297 - Pending		0.16%
4172703 - =		0.00%
37045640 - Comment		0.00%
45878745 - Abnormal		0.03%
45880649 - Undetermined		0.02%
45884087 - Equivocal		0.04%
45884153 - Normal		0.53%

National
COVID
Cohort
Collaborative

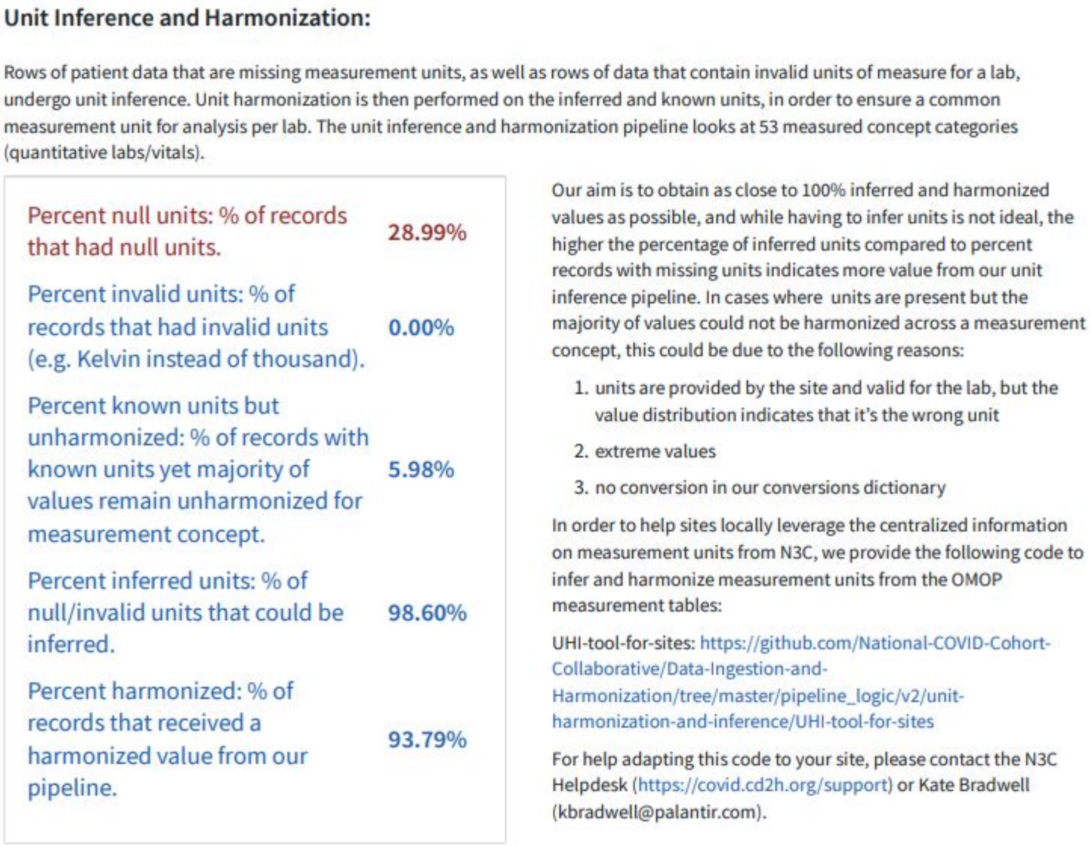
Visits



NIH National Center
for Advancing
Translational Scie

NATIONAL CENTER
FOR DATA TO HEALTH

Measurements/ Labs



Unit Inference and Harmonization:

Rows of patient data that are missing measurement units, as well as rows of data that contain invalid units of measure for a lab, undergo unit inference. Unit harmonization is then performed on the inferred and known units, in order to ensure a common measurement unit for analysis per lab. The unit inference and harmonization pipeline looks at 53 measured concept categories (quantitative labs/vitals).

Percent null units: % of records that had null units.	28.99%
Percent invalid units: % of records that had invalid units (e.g. Kelvin instead of thousand).	0.00%
Percent known units but unharmonized: % of records with known units yet majority of values remain unharmonized for measurement concept.	5.98%
Percent inferred units: % of null/invalid units that could be inferred.	98.60%
Percent harmonized: % of records that received a harmonized value from our pipeline.	93.79%

Our aim is to obtain as close to 100% inferred and harmonized values as possible, and while having to infer units is not ideal, the higher the percentage of inferred units compared to percent records with missing units indicates more value from our unit inference pipeline. In cases where units are present but the majority of values could not be harmonized across a measurement concept, this could be due to the following reasons:

1. units are provided by the site and valid for the lab, but the value distribution indicates that it's the wrong unit
2. extreme values
3. no conversion in our conversions dictionary

In order to help sites locally leverage the centralized information on measurement units from N3C, we provide the following code to infer and harmonize measurement units from the OMOP measurement tables:

UHI-tool-for-sites: https://github.com/National-COVID-Cohort-Collaborative/Data-Ingestion-and-Harmonization/tree/master/pipeline_logic/v2/unit-harmonization-and-inference/UHI-tool-for-sites

For help adapting this code to your site, please contact the N3C Helpdesk (https://covid.cd2h.org/support) or Kate Bradwell (kbradwell@palantir.com).

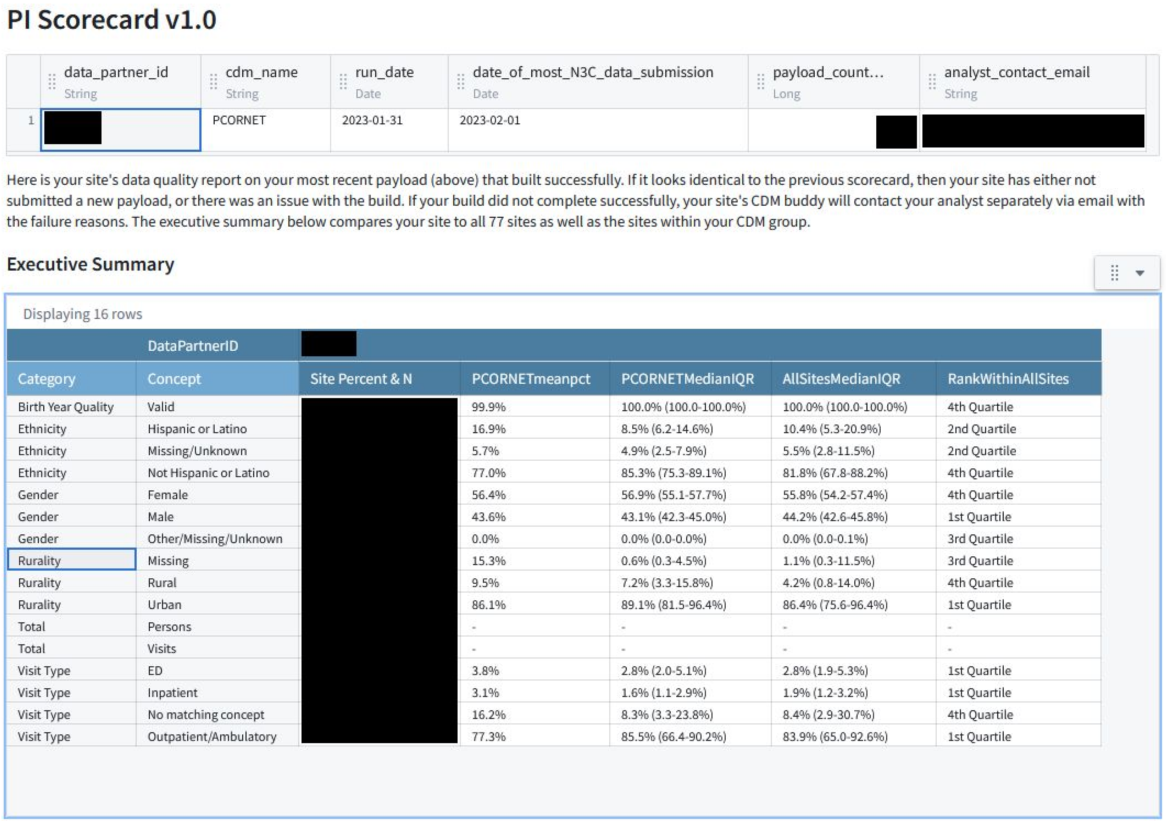
During ingestion, N3C uses machine learning to "rescue" units of measure that are invalid or null (e.g., body weight measured in mmHg).

We aim to provide this rescued data back to sites if they wish to have it, and would welcome discussions from interested sites.

National COVID Cohort Collaborative

NATIONAL CENTER FOR DATA TO HEALTH

PI Scorecard Preview



PI Scorecard v1.0

	data_partner_id String	cdm_name String	run_date Date	date_of_most_N3C_data_submission Date	payload_count... Long	analyst_contact_email String
1	[redacted]	PCORNET	2023-01-31	2023-02-01	[redacted]	[redacted]

Here is your site's data quality report on your most recent payload (above) that built successfully. If it looks identical to the previous scorecard, then your site has either not submitted a new payload, or there was an issue with the build. If your build did not complete successfully, your site's CDM buddy will contact your analyst separately via email with the failure reasons. The executive summary below compares your site to all 77 sites as well as the sites within your CDM group.

Executive Summary

Displaying 16 rows

	DataPartnerID	[redacted]				
Category	Concept	Site Percent & N	PCORNETmeanpct	PCORNETMedianIQR	AllSitesMedianIQR	RankWithinAllSites
Birth Year Quality	Valid	[redacted]	99.9%	100.0% (100.0-100.0%)	100.0% (100.0-100.0%)	4th Quartile
Ethnicity	Hispanic or Latino	[redacted]	16.9%	8.5% (6.2-14.6%)	10.4% (5.3-20.9%)	2nd Quartile
Ethnicity	Missing/Unknown	[redacted]	5.7%	4.9% (2.5-7.9%)	5.5% (2.8-11.5%)	2nd Quartile
Ethnicity	Not Hispanic or Latino	[redacted]	77.0%	85.3% (75.3-89.1%)	81.8% (67.8-88.2%)	4th Quartile
Gender	Female	[redacted]	56.4%	56.9% (55.1-57.7%)	55.8% (54.2-57.4%)	4th Quartile
Gender	Male	[redacted]	43.6%	43.1% (42.3-45.0%)	44.2% (42.6-45.8%)	1st Quartile
Gender	Other/Missing/Unknown	[redacted]	0.0%	0.0% (0.0-0.0%)	0.0% (0.0-0.1%)	3rd Quartile
Rurality	Missing	[redacted]	15.3%	0.6% (0.3-4.5%)	1.1% (0.3-11.5%)	3rd Quartile
Rurality	Rural	[redacted]	9.5%	7.2% (3.3-15.8%)	4.2% (0.8-14.0%)	4th Quartile
Rurality	Urban	[redacted]	86.1%	89.1% (81.5-96.4%)	86.4% (75.6-96.4%)	1st Quartile
Total	Persons	[redacted]	-	-	-	-
Total	Visits	[redacted]	-	-	-	-
Visit Type	ED	[redacted]	3.8%	2.8% (2.0-5.1%)	2.8% (1.9-5.3%)	1st Quartile
Visit Type	Inpatient	[redacted]	3.1%	1.6% (1.1-2.9%)	1.9% (1.2-3.2%)	1st Quartile
Visit Type	No matching concept	[redacted]	16.2%	8.3% (3.3-23.8%)	8.4% (2.9-30.7%)	4th Quartile
Visit Type	Outpatient/Ambulatory	[redacted]	77.3%	85.5% (66.4-90.2%)	83.9% (65.0-92.6%)	1st Quartile

NIH National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

National
COVID
Cohort
Collaborative

What else would you like to see on your scorecard?

What is interesting, useful, or actionable at your site?

- How demographically representative is your site, compared with your region, your state, or the nation?
- N3C enclave user statistics–who's using the enclave, and what contributions have been made by your site?
- What social determinants data is being collected by your site, and in what volumes?
- Other ideas?

Let's discuss!

National
COVID
Cohort
Collaborative

Takeaways

- There has never been more of an opportunity for CTSAs to exist as a harmonized data network.
    - ***Why harmonize?*** With harmonized data, multi-site data-driven research is more feasible, more reproducible, and higher quality.
- Scorecards allow hubs to see where they stand against their peers in an informative, low-pressure exercise.
- Leveraging centralized data quality processes like scorecards means less DQ work at each site and shared decision-making about improvements

National
COVID
Cohort
Collaborative

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

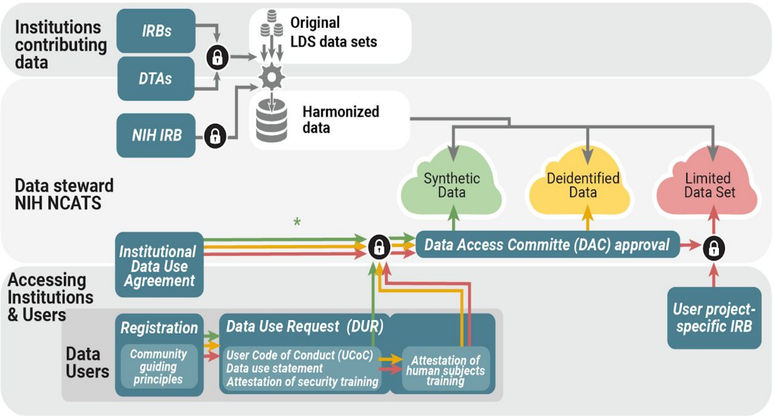
Launching the National "Clinical" Cohort Collaborative (N3C)

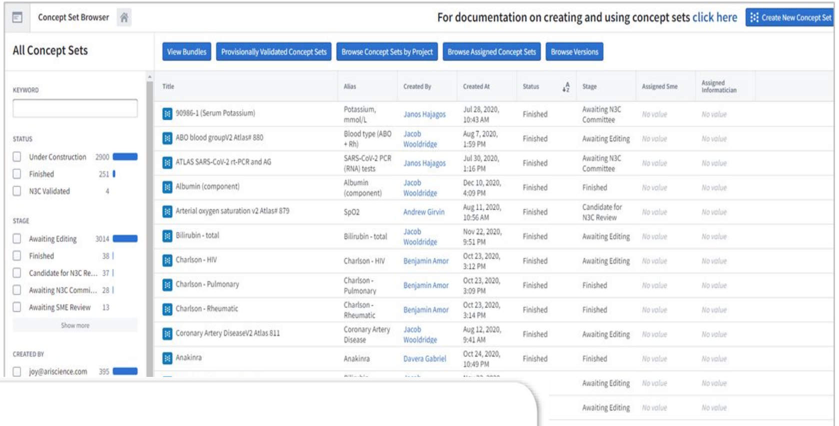
Melissa Haendel, PhD, FACMI

CRIO UC Anschutz
N3C Co-Lead; Center for Data to Health

National COVID Cohort Collaborative

CD2H Remit: Community Governance, Data sharing, Collaborative analytics


Institutions contributing data
IRBs
DTAs
Original LDS data sets
NIH IRB
Harmonized data
Data steward NIH NCATS
Synthetic Data
Deidentified Data
Limited Data Set
*
Institutional Data Use Agreement
Data Access Committe (DAC) approval
User project-specific IRB
Accessing Institutions & Users
Data Users
Registration
Community guiding principles
Data Use Request (DUR)
User Code of Conduct (UCoC)
Data use statement
Attestation of security training
Attestation of human subjects training


Concept Set Browser
For documentation on creating and using concept sets click here
Create New Concept Set
All Concept Sets
View Bundles
Provisionally Validated Concept Sets
Browse Concept Sets by Project
Browse Assigned Concept Sets
Browse Versions
KEYWORD
Show more

STATUS

□ Under Construction 2900
□ Finished 251
□ N3C Validated 4

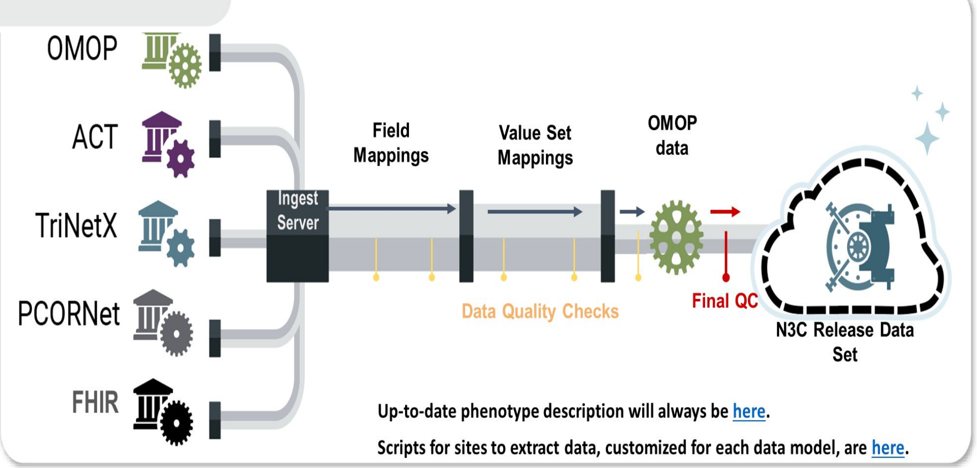
STAGE

□ Awaiting Editing 3014
□ Finished 38
□ Candidate for N3C Re... 37
□ Awaiting N3C Commi... 28
□ Awaiting SME Review 13

CREATED BY

□ joy@ariscience.com 395

Title	Alias	Created By	Created At	Status	Stage	Assigned Sme	Assigned Informatician
90986-1 (Serum Potassium)	Potassium, mmol/L	Janos Hajagos	Jul 28, 2020, 10:43 AM	Finished	Awaiting N3C Committee	No value	No value
ABO blood groupV2 Atlas# 880	Blood type (ABO + Rh)	Jacob Wooldridge	Aug 7, 2020, 1:59 PM	Finished	Awaiting Editing	No value	No value
ATLAS SARS-CoV-2 rt-PCR and AG	SARS-CoV-2 PCR (RNA) tests	Janos Hajagos	Jul 30, 2020, 1:16 PM	Finished	Awaiting N3C Committee	No value	No value
Albumin (component)	Albumin (component)	Jacob Wooldridge	Dec 10, 2020, 4:09 PM	Finished	Finished	No value	No value
Arterial oxygen saturation v2 Atlas# 879	SpO2	Andrew Girvin	Aug 11, 2020, 10:56 AM	Finished	Candidate for N3C Review	No value	No value
Bilirubin - total	Bilirubin - total	Jacob Wooldridge	Nov 22, 2020, 9:51 PM	Finished	Awaiting Editing	No value	No value
Charlson - HIV	Charlson - HIV	Benjamin Amor	Oct 23, 2020, 3:12 PM	Finished	Awaiting Editing	No value	No value
Charlson - Pulmonary	Charlson - Pulmonary	Benjamin Amor	Oct 23, 2020, 3:09 PM	Finished	Finished	No value	No value
Charlson - Rheumatic	Charlson - Rheumatic	Benjamin Amor	Oct 23, 2020, 3:14 PM	Finished	Finished	No value	No value
Coronary Artery DiseaseV2 Atlas 811	Coronary Artery Disease	Jacob Wooldridge	Aug 12, 2020, 9:41 AM	Finished	Awaiting Editing	No value	No value
Anakinra	Anakinra	Davera Gabriel	Oct 24, 2020, 10:49 PM	Finished	Finished	No value	No value
					Awaiting Editing	No value	No value
					Awaiting Editing	No value	No value
					Finished	No value	No value
					Finished	No value	No value
					Finished	No value	No value
					Finished	No value	No value

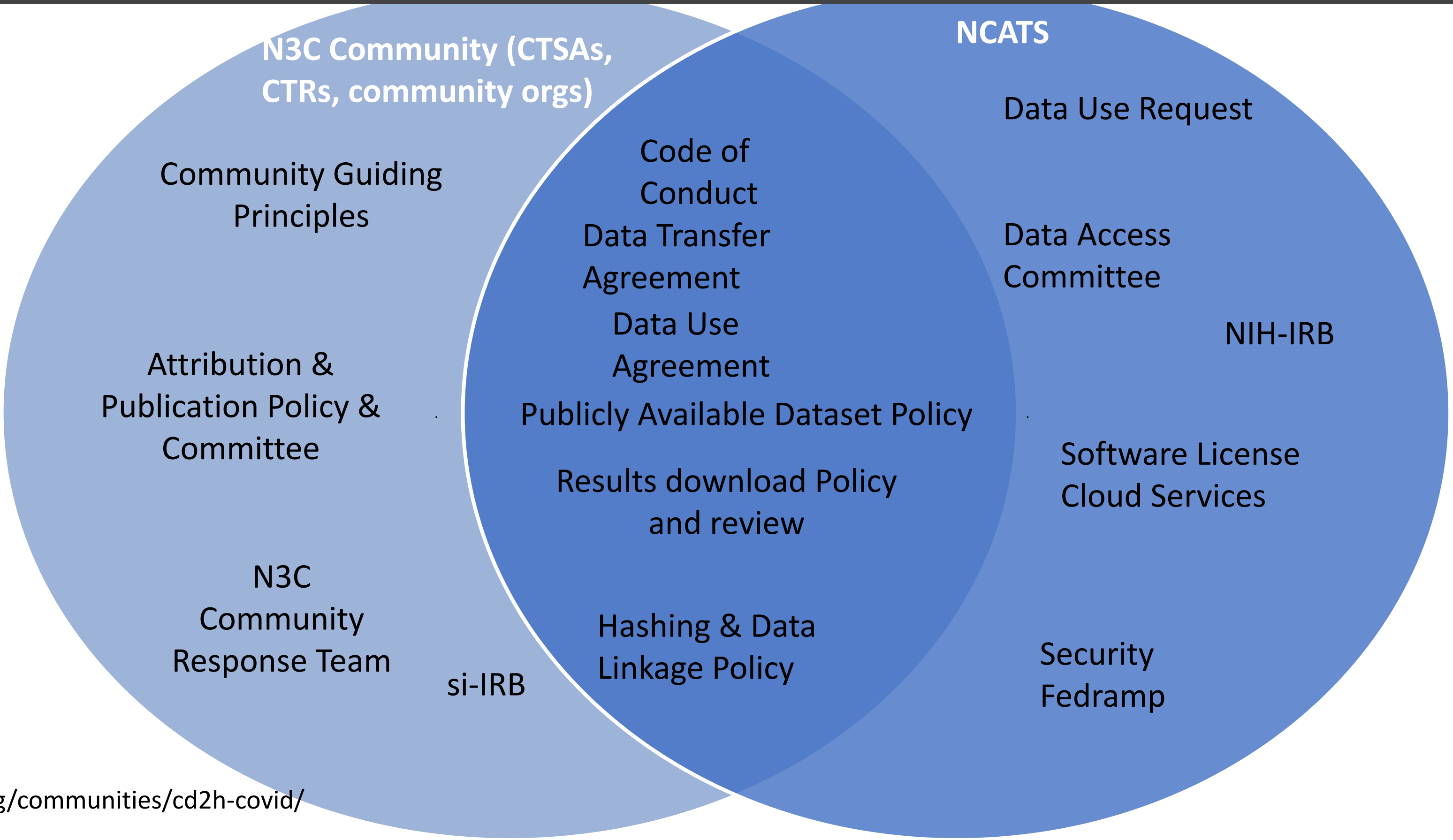


Up-to-date phenotype description will always be here.

Scripts for sites to extract data, customized for each data model, are here.

Historical N3C Shared Governance

NATIONAL CENTER FOR DATA TO HEALTH



N3C Community (CTSAs, CTRs, community orgs)

Community Guiding Principles

Attribution & Publication Policy & Committee

N3C Community Response Team

si-IRB

Code of Conduct

Data Transfer Agreement

Data Use Agreement

Publicly Available Dataset Policy

Results download Policy and review

Hashing & Data Linkage Policy

NCATS

Data Use Request

Data Access Committee

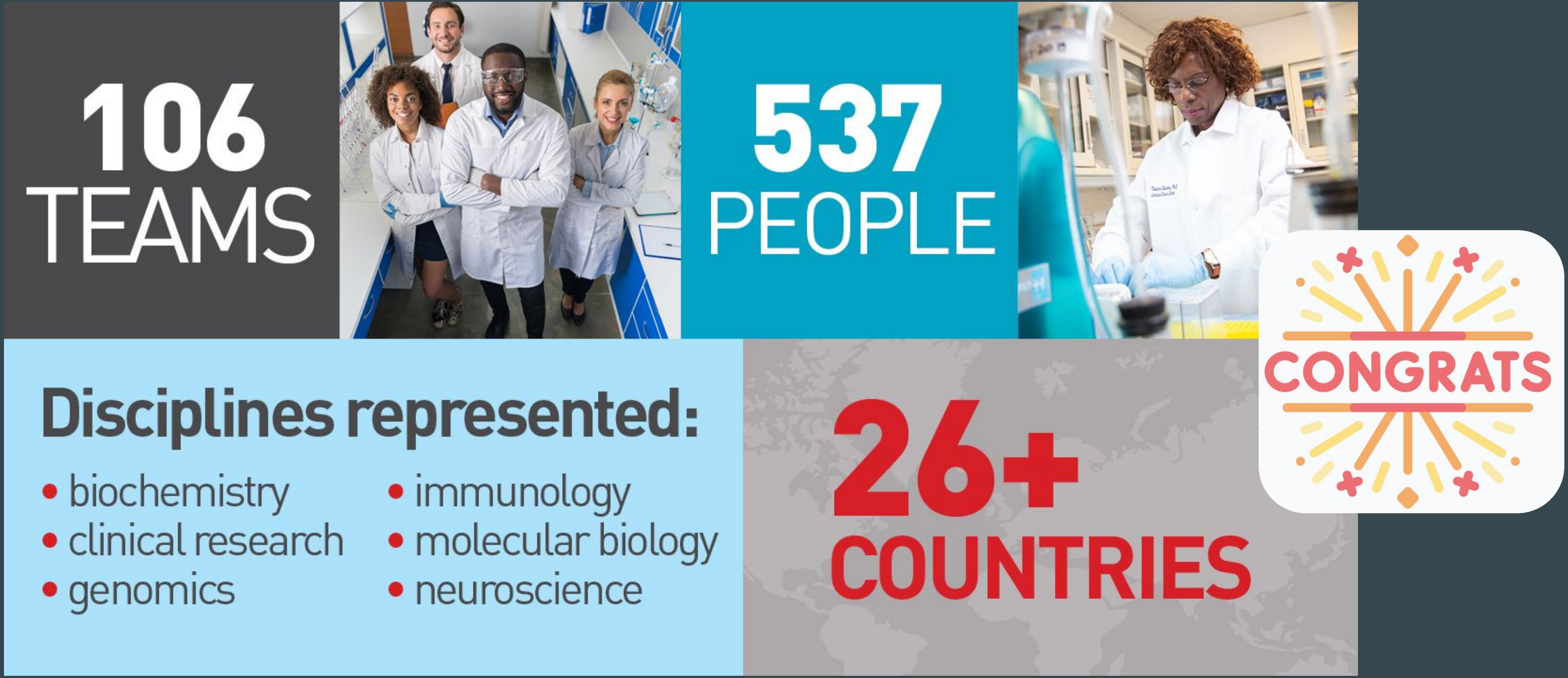
NIH-IRB

Software License
Cloud Services

Security
Fedramp

https://zenodo.org/communities/cd2h-covid/

N3C won grand prize in the Dataworks! Competition!
Democratizing access to sensitive clinical data



106 TEAMS

537 PEOPLE

Disciplines represented:

- biochemistry
- clinical research
- genomics
- immunology
- molecular biology
- neuroscience

26+ COUNTRIES

CONGRATS

https://datascience.nih.gov/director/directors-blog-dataworks-winners-2023

Diverse impact of N3C collaborative analytics

RESULTED IN SIGNIFICANT SCHOLARLY PRODUCTIVITY

ATTRIBUTED AT SCALE AND INCENTIVIZED COLLABORATION

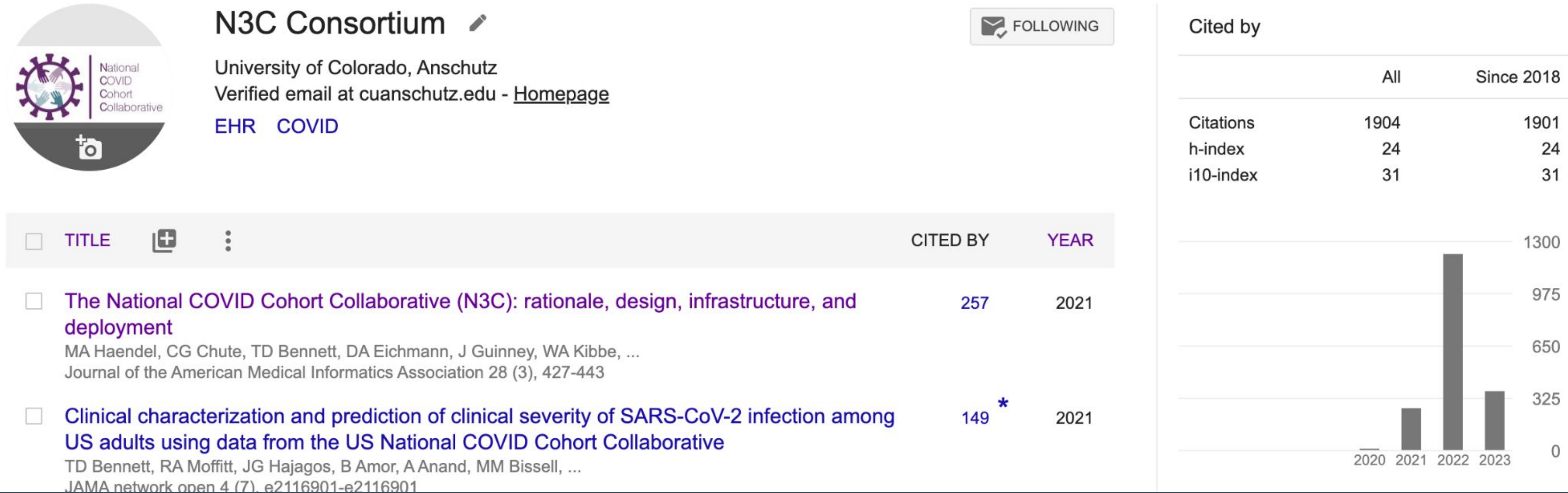
TRANSFORMED CARE GUIDELINES

DEVELOPED EVIDENCE-BASED DISEASE DEFINITIONS

DEVELOPED COMPLEX RISK PREDICTION MODELS

How can we bring these successes to bear on all the other disease areas of interest to the CTSA program?

Impact: Across the program, >1900 citations, H index of 24



N3C Consortium

FOLLOWING

University of Colorado, Anschutz
Verified email at cuanschutz.edu - Homepage

EHR COVID

Cited by

	All	Since 2018
Citations	1904	1901
h-index	24	24
i10-index	31	31

□ TITLE	CITED BY	YEAR
□ The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment MA Haendel, CG Chute, TD Bennett, DA Eichmann, J Guinney, WA Kibbe, ... Journal of the American Medical Informatics Association 28 (3), 427-443	257	2021
□ Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID Cohort Collaborative TD Bennett, RA Moffitt, JG Hajagos, B Amor, A Anand, MM Bissell, ... JAMA network open 4 (7), e2116901-e2116901	149 *	2021

bit.ly/n3c-google-scholar

Julie McMurry, Carolyn Bramante, Swaroop Vedula

How do you measure Success

NATIONAL CENTER
FOR DATA TO HEALTH

Team Science: > 3400 Users, >430 studies, N3C Leadership is predominantly Women and Minority Leadership

Citations: 1904 Citations, h-index 24, i10 index of 31, the 2023 “article of the year” by The Journal of Rural Health.

Largess: Largest COVID repository in the USA >18 million patient, 22 Billion rows of data, 77 health systems

Data Quality: Score Card, Data Quality Checks, Unit harmination

Inclusive Networks: Only Network that includes: PCORNET, OMOP, ACT, TriNetX

Education/support: 763 training resources, personal help, office hours, best practice, tickets, website, news letter, video, office hours, Domain Team, Forum

Recognition: Biden administration, senate, and governor requests; Dataworks! Grand prize, NIH director's blog, NPR

SDoH: AI/AN, 60+ public data sets, CMS medicare and medicaid data

Organizational Users and Partners: ONC, FDA, NCI, ASPE, ASPR, AHRQ, NIBIB, All of Us, NHLBI

How you define Success

Team Science: > 3400 Users, >430 studies, N3C Leadership is predominantly Women and Minority Leadership

Citations: 1904 Citations, h-index 24, i10 index of 31, the 2023 “article of the year” by The Journal of Rural Health.

Scalability: Largest COVID repository in the USA >18 million patient, 22 Billion rows of data, 77 health systems

Data Quality: Score Card, Data Quality Checks, Unit harmination

Inclusive Networks: Only Network that includes: PCORNET, OMOP, ACT, TriNetX

Education/support: 763 training resources, personal help, office hours, best practice, tickets, website, news letter, video, office hours, Domain Team, Forum

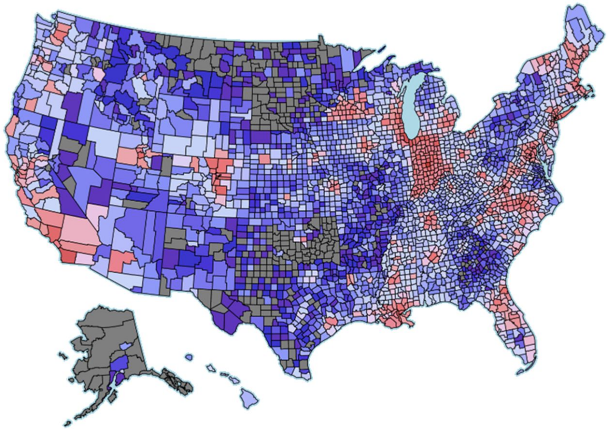
Recognition: Biden administration, senate, and governor requests; Dataworks! Grand prize, NIH director's blog, NPR

SDoH: AI/AN, 60+ public data sets, CMS medicare and medicaid data

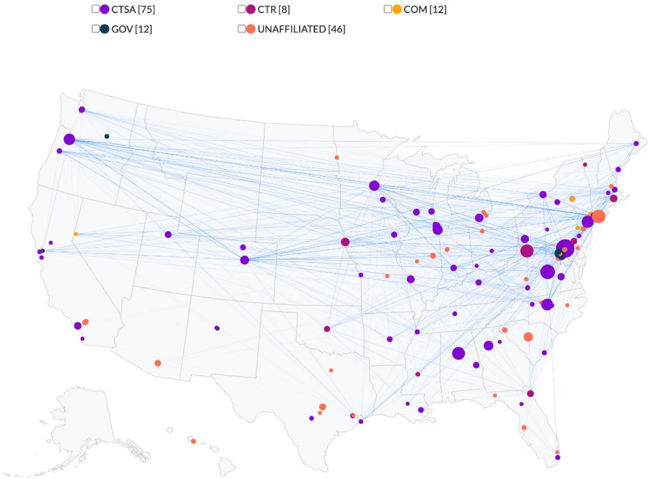
Organizational Users and Partners: ONC, FDA, NCI, ASPE, ASPR, AHRQ, NIBIB, All of Us, NHLBI

N3C has harmonized EHR data from >2800 counties from all fifty states

NATIONAL CENTER FOR DATA TO HEALTH



Jerrod Anzalone



<u>All </u>CTSAs are using N3C data

Data from 56 CTSAs is available, 26 CTSA/affiliated sites are pending/submitted

>230 individual organizations have submitted data

84 CTSA hubs/affiliates have active users, even those that did not submit data

<= 75 CTSA orgs are collaborating on projects

Addressing Bias

- N3C is representative demographically (Race, Age, Ethnicity, Sex, Rurality, Geographic location) and socio-economically compared to CDC, JHU, NYT and other data sources
- There are biases in the fact that many of the sites in N3C are academic medical centers, however with OCHIN, linkage, and CMS/medicare/medicaid data, we have a representative set of patients and data types that cover more outpatient/other patient activities
- Additional biases include the fact that health is not all about clinical encounters - mobile health data, education data, etc, all provide a different suite of perspectives
- A number of methods and data sources (e.g. N3C has patient-level source of SDoH survey data) aim to reduce analytical bias

Consortia for Collaborative Clinical Analyses are Inevitable

- Integrating and harmonizing data across medical centers
  - Enhances power, reduces bias, enables rare disease
  - Support sub-phenotype analysis (precision medicine)
- Inevitable trajectory for American medicine
- The only question is who controls data and knowledge
  - Insurance industry
  - EHR Vendors
  - Hi-tech companies
  - Academic community

N3C Beyond COVID

NATIONAL CENTER
FOR DATA TO HEALTH

- N3C is the largest and most successful public repository of longitudinal EHR data in the US to date and is a testament to the CTSA program
- The robust machinery can be generalized to be disease agnostic
  - Synergy of federated data repositories at contributing sites
  - Phenotype queries and scripts for data transfer and ingestion
  - Multiple model to OMOP transformation and harmonization pipeline
  - FedRAMP secure data repository
  - Analytic environment that has generated >100 published artifacts

N3C Clinical seeks to leverage this infrastructure beyond COVID

This pilot is to explore logistics and governance options for doing so

Characteristics of N3C Clinical that are similar to N3C COVID

- Continuing to leverage common data model (CDM) repositories such as OMOP, PCORNet, TriNetX, or ACT already in place at your CTSA
- Continuing to provide executable queries for your specific CDM that will extract patients with a longitudinal connection to your clinical organization.
- Continuing centralized harmonization and data quality enhancements such as imputing missing units of measure.
- Continuing to provide a highly secure, FISMA-moderate compliant, data analytic environment that blocks any data exfiltration
- Continuing to operate on a federally managed cloud

Fundamental differences between N3C Clinical and N3C COVID

- Contributing organizations have complete agency and access control over how their data is used. They may participate in none, all, or selected project proposals.
- Data access proposals will be reviewed and must be approved by a community managed panel with membership from all data-contributing pilot organizations.
- Completely new central IRB and NIH IRB. Completely new Data Transfer Agreement, and Data Use Agreement have been drafted to reflect the broader scope and more constrained access to these resources; these are subject to revision by the Pilot group.
- Program governance will evolve and change over the pilot to reflect closely the needs and expectations of data contributing organizations.

Advantages of participating: N3C Clinical Data Contributors Retain Agency over Data

NATIONAL CENTER FOR DATA TO HEALTH

- All pilot institutions agree to have their submitted data harmonized through the data extraction and transformation pipeline from their submitted model to OMOP.
    - This will include centralized data quality benchmarking, and units of measure corrections.
- Contributing sites have unrestricted privilege to securely analyze and share their own data on the enclave.
- Pilot members can and will shape the governance of N3C Clinical when it expands beyond the pilot.
- Data access options range from large pan-CTSA projects to smaller projects with a limited set of data contributors. Contributing sites will be able to choose one of these options that range from broad to narrow data sharing:
    - Limiting site data access to defined categories of domains or projects
    - Agreeing to site data access for all projects approved by the community data access review process
    - Limiting site data access to collaborations and projects of interest to the site

Volunteer pilot sites and project domains as voted on by the community

NATIONAL CENTER FOR DATA TO HEALTH

University of North Carolina
University of Colorado/Children’s
JHU
University of Chicago
University of Washington
Stanford
OHSU
University of Virginia
OCHIN
University of Nebraska (UNMC)

Chosen based on N3C activity and interest, and diversity

Alzheimer's
Renal
Pulmonary
Cancer

All with healthcare utilization focus

Chosen based on feasibility and interest

Community governance calls are Fridays 7am PT/10am ET
https://covid.cd2h.org/n3c-calendar

SC questions and requests

NATIONAL CENTER FOR DATA TO HEALTH

- How can N3C best support each CTSA? How can we further advance the N3C network effect?
- We have been requested to meet with each pod to answer questions and solicit feedback. What makes most sense in terms of process?
- How can we best address your sites’ data improvement needs? What kind of training or additional support would be helpful?
- What does your site need to increase participation or otherwise best take advantage of these data assets and collaborative opportunities?
  - E.g. align with RCTs and prep-to-research
  - Training and alignment with K awards
  - etc.