NORTHERN NEW ENGLAND
CLINICAL & TRANSLATIONAL
RESEARCH NETWORK

Introduction to the National COVID Cohort Collaborative (N3C)

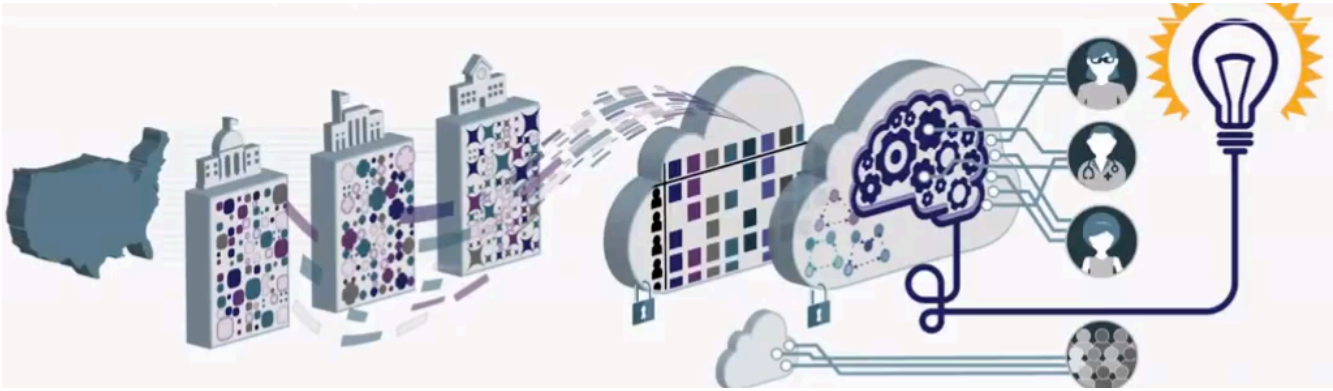
Ivette F. Emery, PhD
N3C Point Person for Outreach and Education

25 January 2021

N3C – An Analytics Platform

National COVID Cohort Collaborative

- The N3C is a clinical data analytics platform that leverages vast amounts of data from collaborating health care providers nationwide to accelerate advances in COVID-19 research and clinical care.

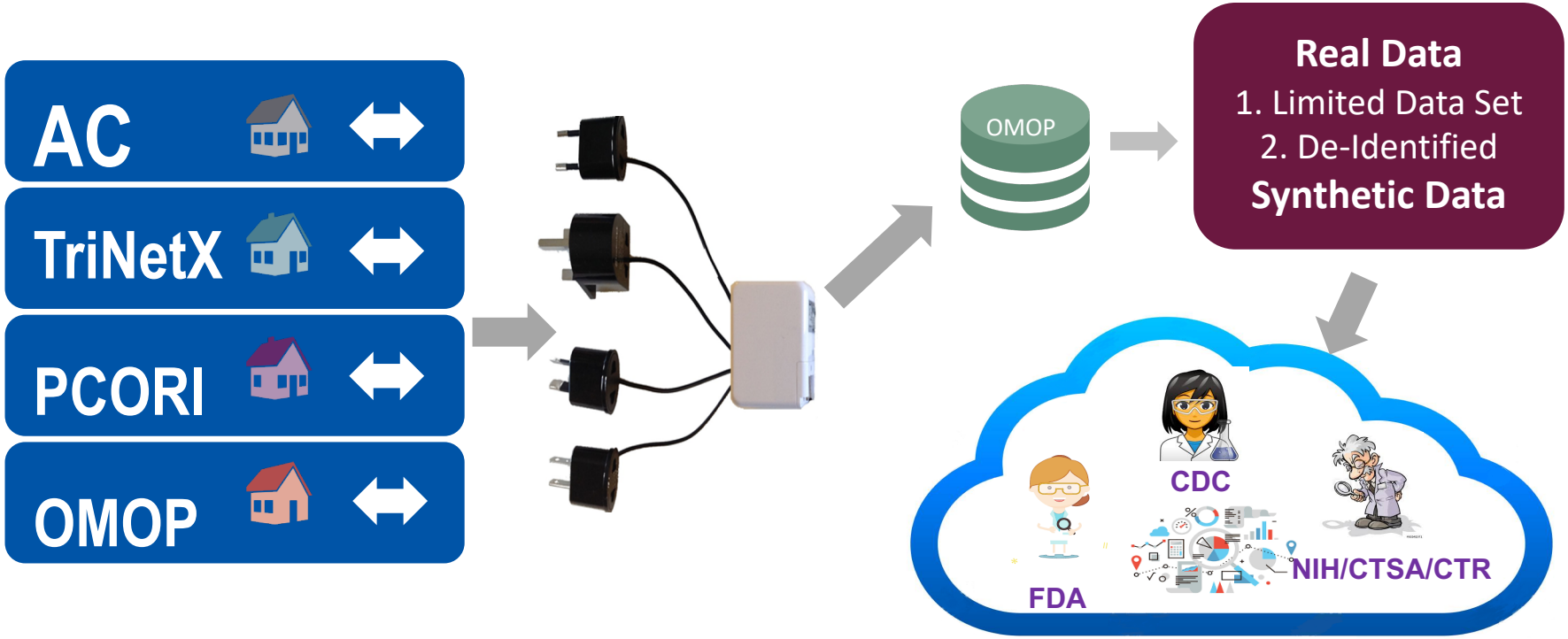


- The N3C is a partnership among the Clinical and Translational Science (CTSA) awardees, the Clinical and Translational Research (CTR) awardees*, and the National Center for Data to Health (CD2H) with overall stewardship by NIH's National Center for Advancing Translational Science (NCATS).

8 CTRs: Northern New England, West Virginia, Rhode Island, Mississippi, Oklahoma, Nebraska, Louisiana, Delaware

NCATS N3C website: ncats.nih.gov/n3c CD2H N3C website: covid.cd2h.org

N3C Analytics Platform in a nutshell



Data Acquisition → Harmonization → Collaborative Analytics

From clinical data networks common data models

Into an OMOP analytics data set

Within the N3C platform

N3C Benefits to Investigators

* Access to large scale COVID-19 data to answer important questions

    1. Do some therapies work better than others? By region? By demographics?
    2. Can we compare local rare clinical observations with national occurrences?
    3. Can we predict who might have severe outcomes if they have COVID-19?
    4. Why are some people asymptomatic?
    5. What social determinants of health are risk factors for mortality?

* Pilot data for grant proposals

* Team science opportunities

    Access to clinical domain, statistics, machine learning, and informatics expertise.

* Participating is what is expected to be sea-change in healthcare research

N3C Data Transfer and Utilization Status

Image courtesy of NCATS

As of Jan 14th 2021

75 sites have executed data transfer agreements (DTAs)

39 sites have data released in the N3C data enclave

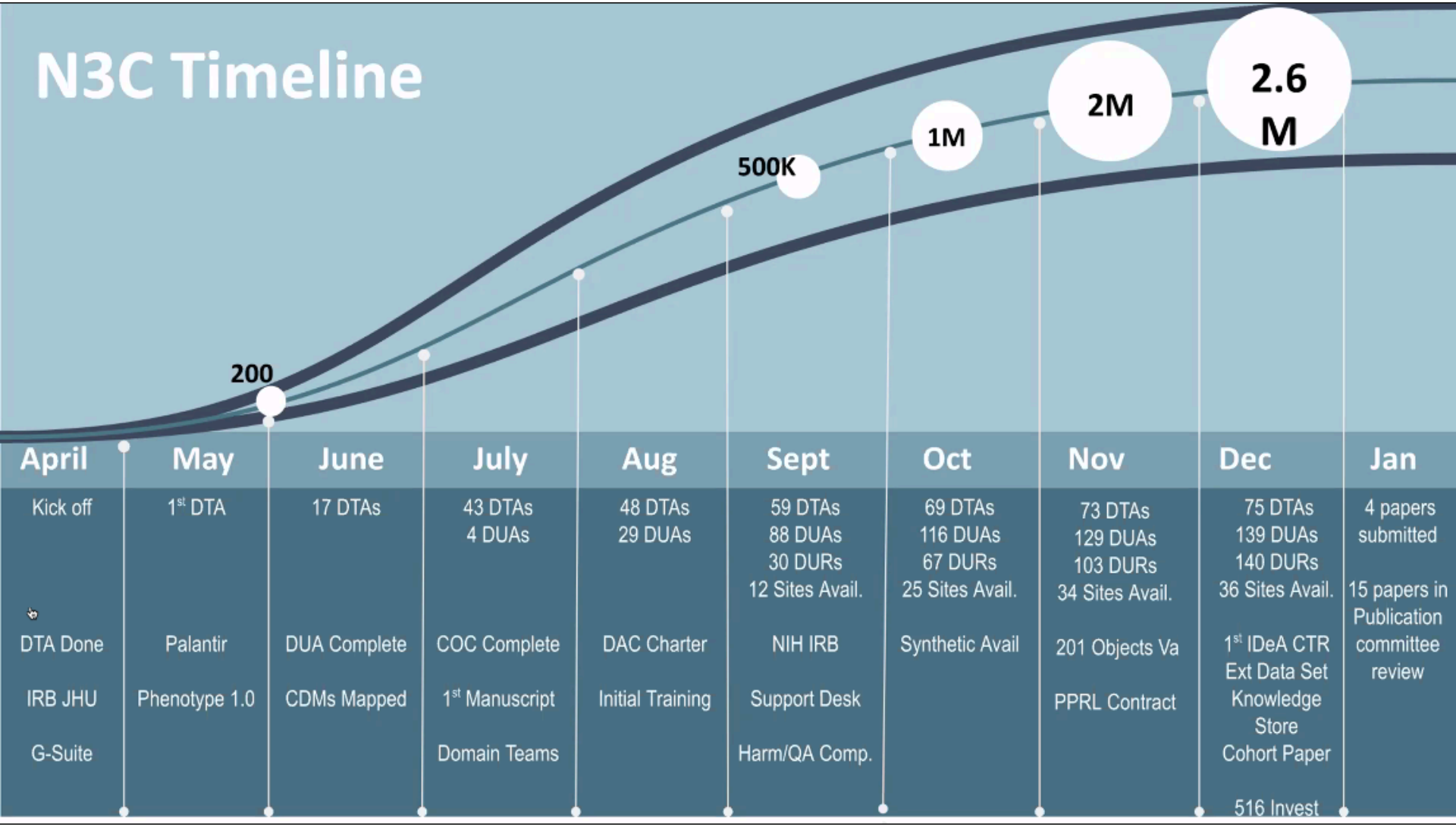
140 sites have executed data use agreements (DUAs)

106 approved projects

1231 participants

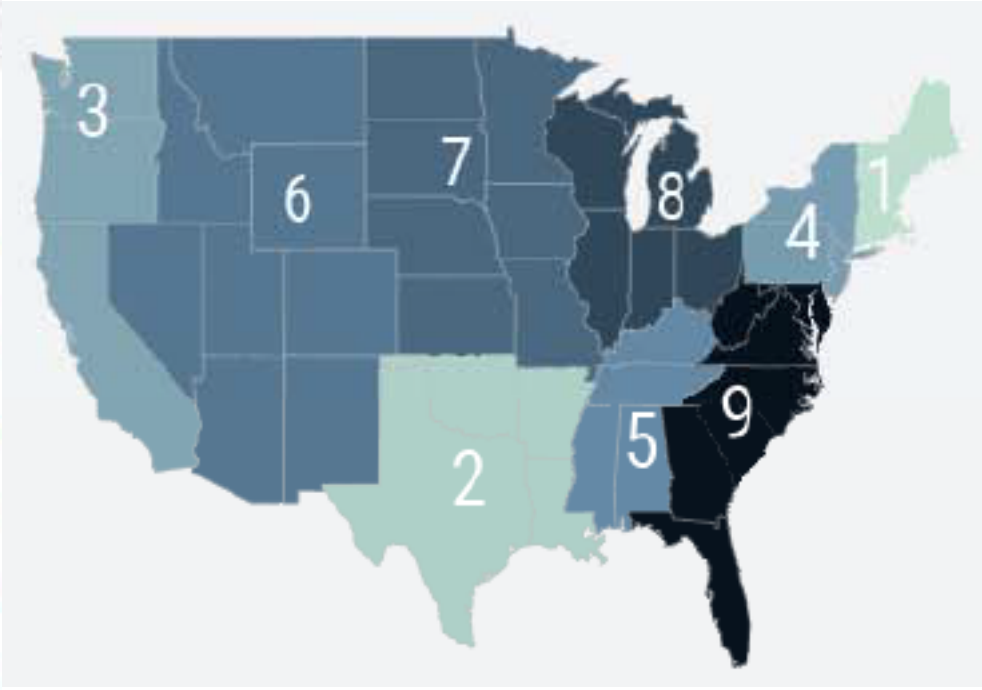
15 manuscripts under review

N3C Data Transfer and Utilization Status

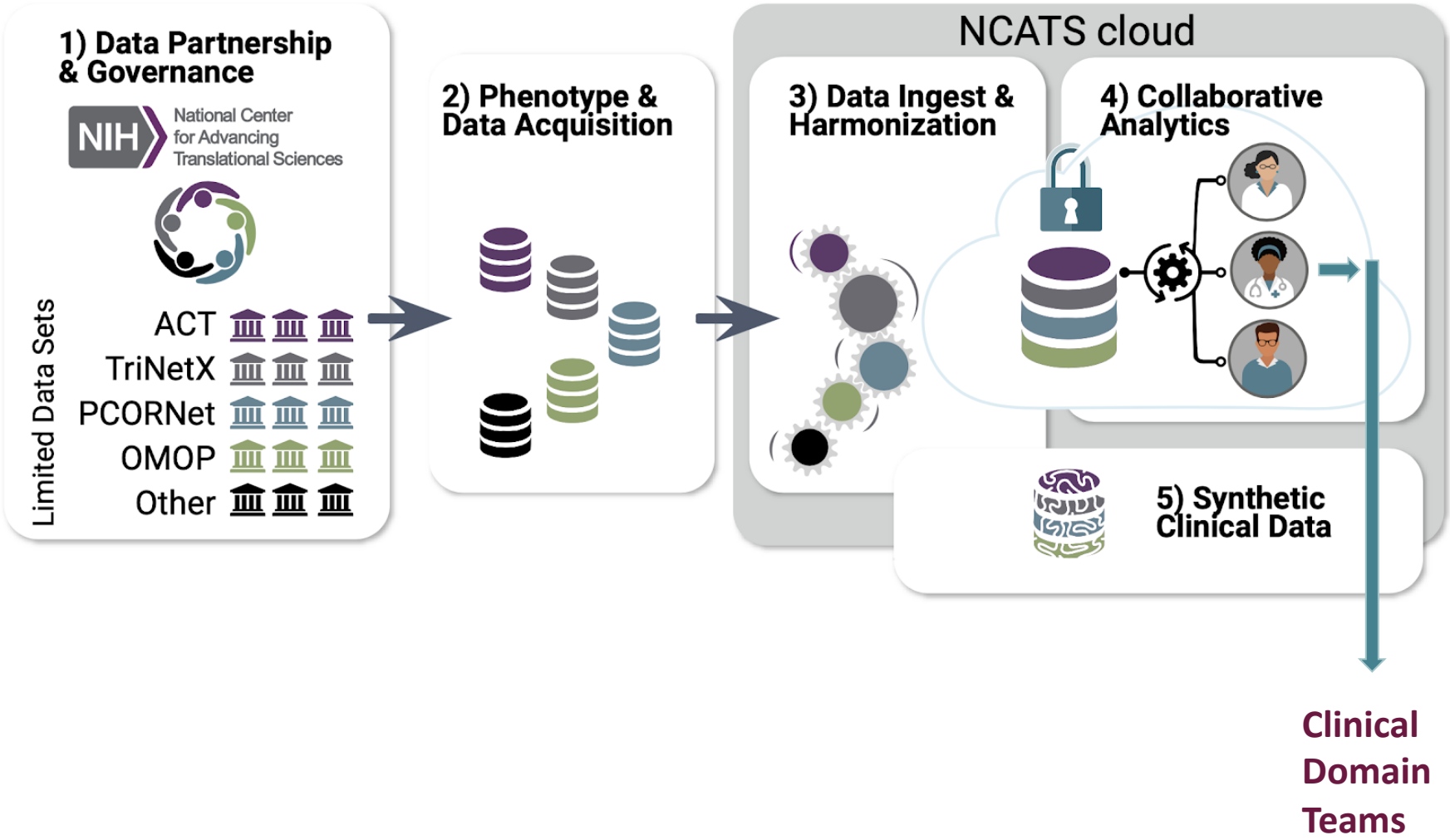


A. Geographic Distribution of N3C Cohort

Region Name	#	Cumulative in N3C Cohort
New England	1	5,560
West South Central	2	23,003
Pacific	3	66,974
Middle Atlantic	4	75,208
East South Central	5	95,746
Mountain	6	193,343
West North Central	7	277,347
East North Central	8	452,921
South Atlantic	9	736,424



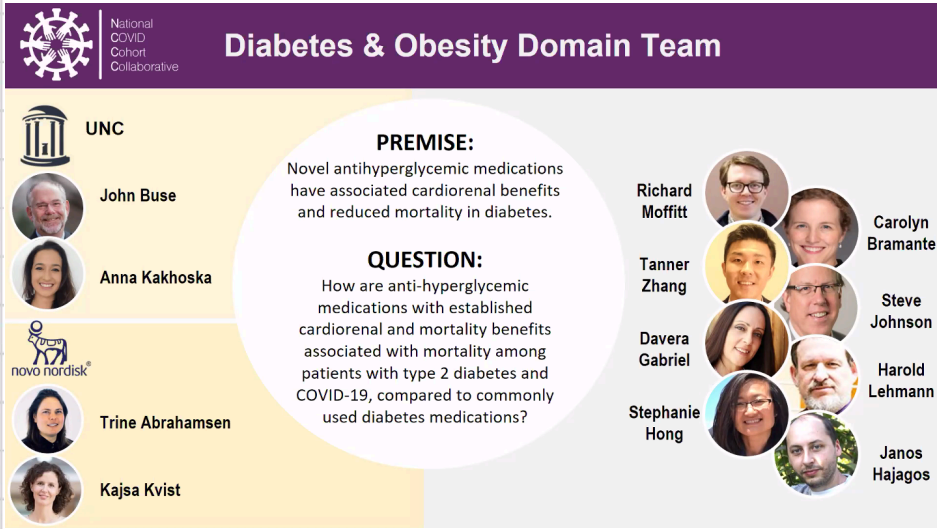
N3C Community Workstreams



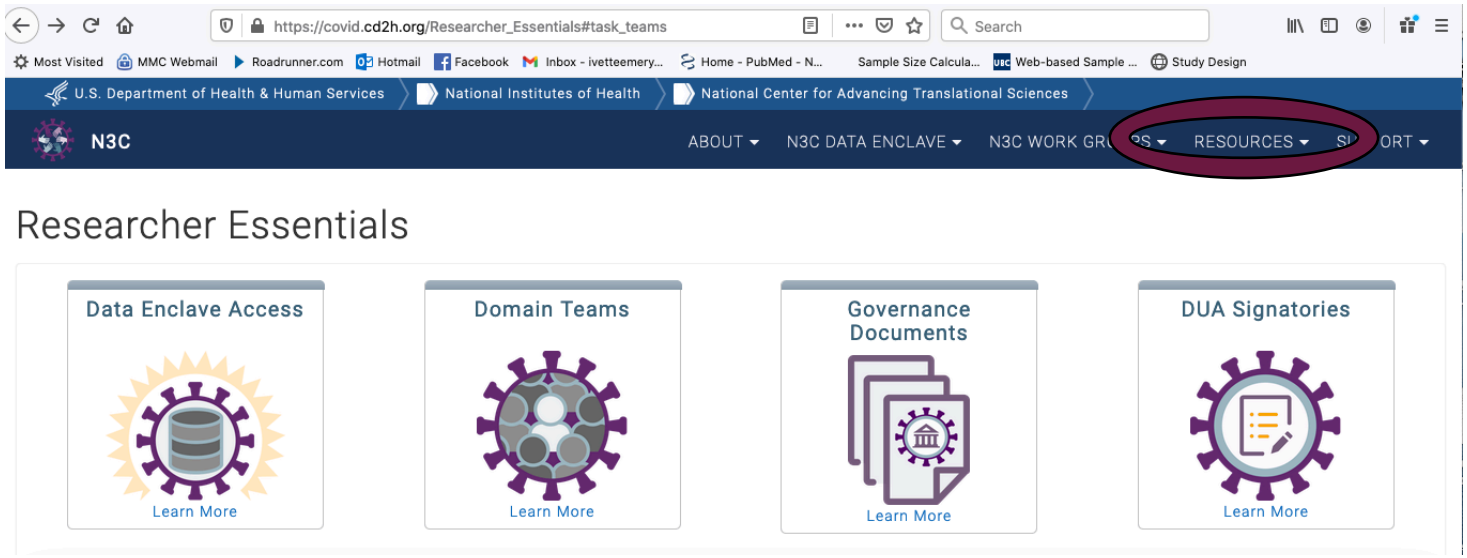
N3C Clinical Domain Teams

Group Name	Mailing List
AKI/ARB/ACE	Join
Applicable Data Methods Standards	Join
Critical Care	Join
Decompensation/multi-organ dysfunction	Join
Diabetes	Join
Elder impact	Join
Emergency Department Avoidance Impact	Join
Imaging	Join
Immunosuppressed/compromised	Join
Oncology	Join
Pediatrics	Join
Pharmacoepidemiology	Join
Pharma-commercial	Join
Population Health/Health Policy	Join
Pregnancy	Join
Short/long-term complications	Join
Social Determinants of Health (SDoH)	Join

Each domain team enables researchers with shared clinical questions to analyze data collectively and collaborate in a team science environment.



N3C Clinical Domain Teams



- To join a Domain Team - Go to Researcher Essentials
  - Google Drive Folder
  - Slack Channel
  - Email Listserv (Google Groups)
- Meetings Published on N3C Calendar
- Common DUR Formats
- For Limited Dataset access - Common IRB format
- Shared Codesets and Resources
- Domain Team Web pages- *Coming soon!*
- Domain Team Project Dashboard-*Coming soon!*

N3C Data

INCLUSION CRITERIA

- All ages
- Inclusion criteria start date of 1/1/2020, lookback period to 1/1/2018.

Lab Confirmed Positive

- LOINC codes Positive result

Lab Confirmed Negative

- LOINC codes Negative result
- Asymptomatic negatives excluded

Suspected Positive

- COVID Dx Code (other strong positive) with no lab result

Possible Positive

- Two or more suggestive ICD codes

DATA

- Observations
- Specimens
- Visit
- Procedures
- Drugs
- Devices
- Conditions
- Measurements
- Location
- Provider

N3C Data

COVID-19 Clinical Data Warehouse Data Dictionary

Based on

OMOP Common Data Model Specifications Version 5.3.1

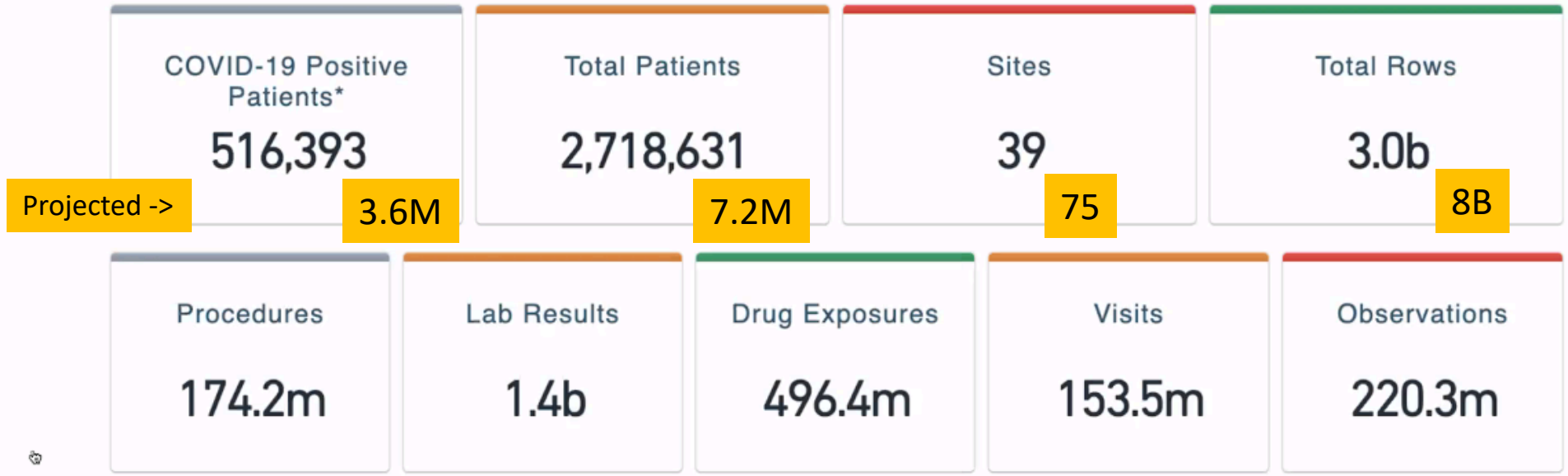
Contents

1 PERSON 2
2 OBSERVATION_PERIOD 4
3 SPECIMEN 4
4 DEATH 6
5 VISIT_OCCURRENCE 6
6 PROCEDURE_OCCURRENCE 8
7 DRUG_EXPOSURE 9
8 DEVICE_EXPOSURE 11
9 CONDITION_OCCURRENCE 12
10 MEASUREMENT 14
11 OBSERVATION 16
Standardized Health System Data Tables 17
12 LOCATION 19
13 CARE_SITE 19
14 PROVIDER 20
OHDSI Release notes v5.3 21

N3C Data

Three Billion Rows!

Over Two and Half Million Patients

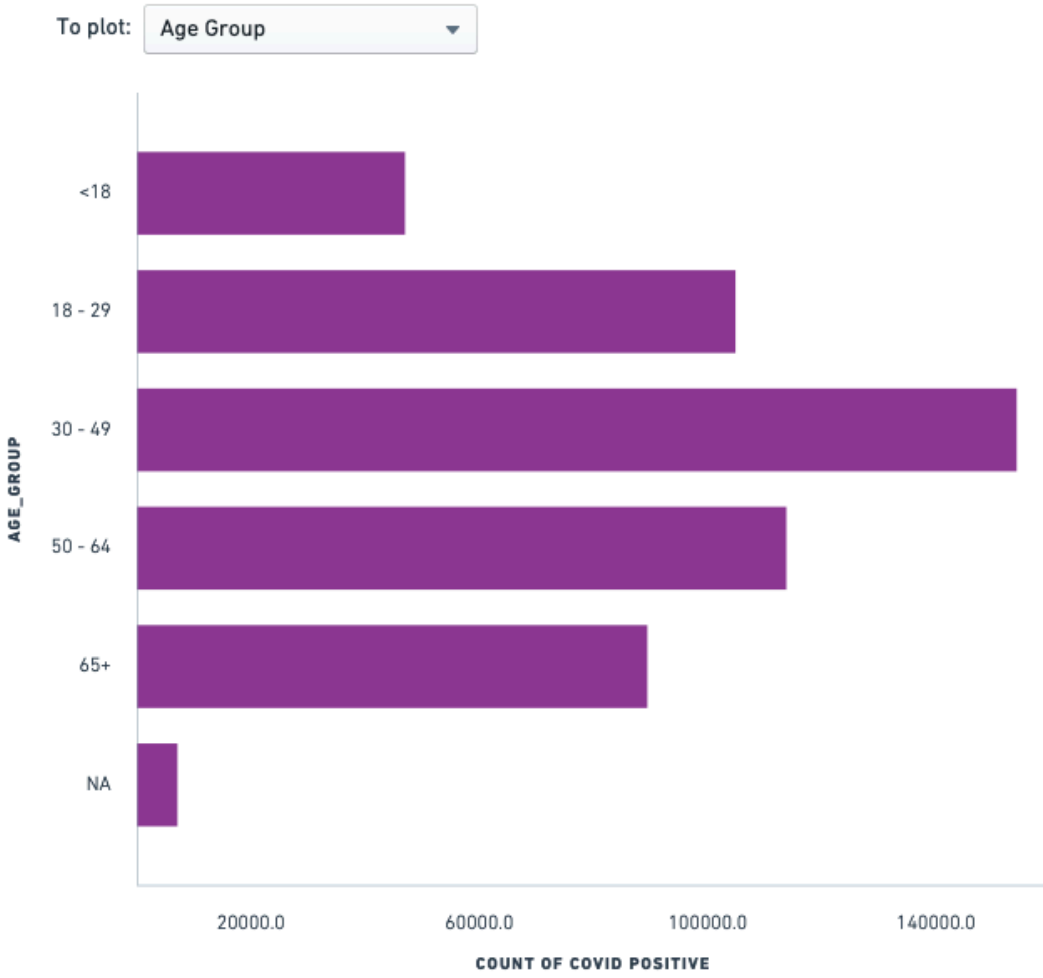


N3C Data

Cohort characteristics

Summary statistics for all patients

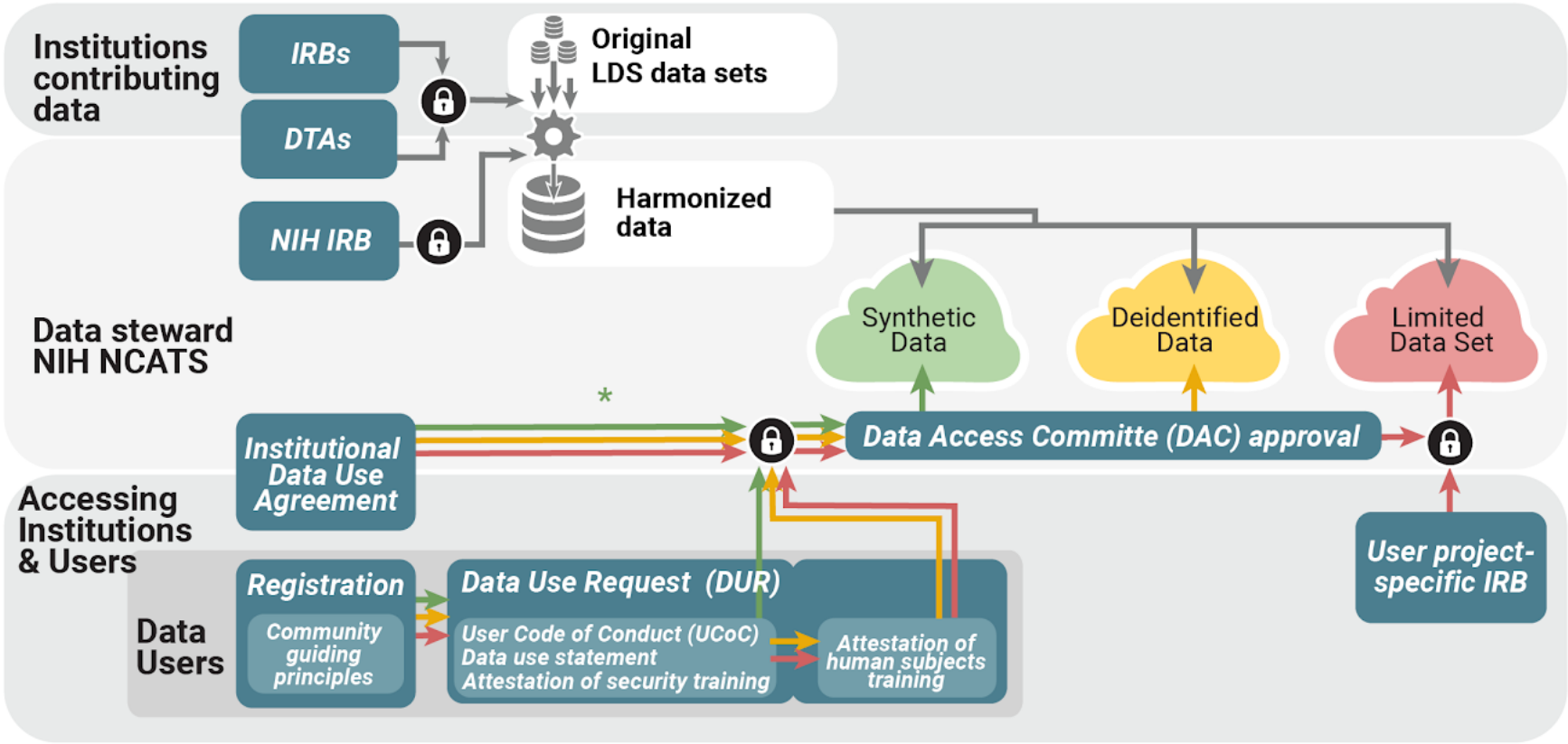
	COVID (N=516393)	Non-COVID (N=2202238)	Overall (N=2,718,631)
Gender			
Male	236203	974100	1210303
Female	279792	1227010	1506802
Null		1119	1119
Age			
0 - 17	46988	329079	376067
18 - 29	104884	322274	427158
30 - 49	154149	570979	725128
50 - 64	113803	470147	583950
65+	89444	479250	568694
Unknown	7125	30509	37634
Race			
White	308964	1536227	1845191
Other	8148	31678	39826
Black or African American	82944	337389	420333
Asian	11867	57118	68985
Pacific Islander	1729	4029	5758
Null	93142	206027	299169
Ethnicity			
Not Hispanic or Latino	353916	1756840	2110756
Hispanic or Latino	78470	184965	263435
Unknown	82965	255305	338270



N3C Data Levels

Data Level	Data Description	Eligible Users	Access Requirements*
Synthetic Data Set	Artificial, statistically-comparable, computational derivative of the original data; it does not contain individually identifiable health information, also known as protected health information (PHI) as defined by the Health Insurance Portability and Accountability Act (HIPAA)	• Researchers from U.S.-based institutions • Researchers from foreign institutions • Citizen scientists	• N3C registration • N3C Data Enclave account • Data Use Agreement (DUA) executed with NCATS • NIH IT training completion • Approved Data Use Request (DUR)
De-identified Data Set	Patient data that has been stripped of PHI identifiers as defined by HIPAA	• Researchers from U.S.-based institutions • Researchers from foreign institutions	• N3C registration • N3C Data Enclave account • DUA executed with NCATS • NIH IT training completion • Approved DUR • Human Subjects Research Protection training completion
Limited Data Set	Patient data that includes only two of the 18 elements defined as PHI by HIPAA (dates of service and patient zip code)	• Researchers from U.S.-based institutions	• N3C registration • N3C Data Enclave account • DUA executed with NCATS • NIH IT training completion • Approved DUR • Human Subjects Research Protection training completion • Local Human Research Protection Program IRB determination letter

N3C Data Oversight



N3C Data Access

First Time Users

Register at N3C

HSP / Security Training

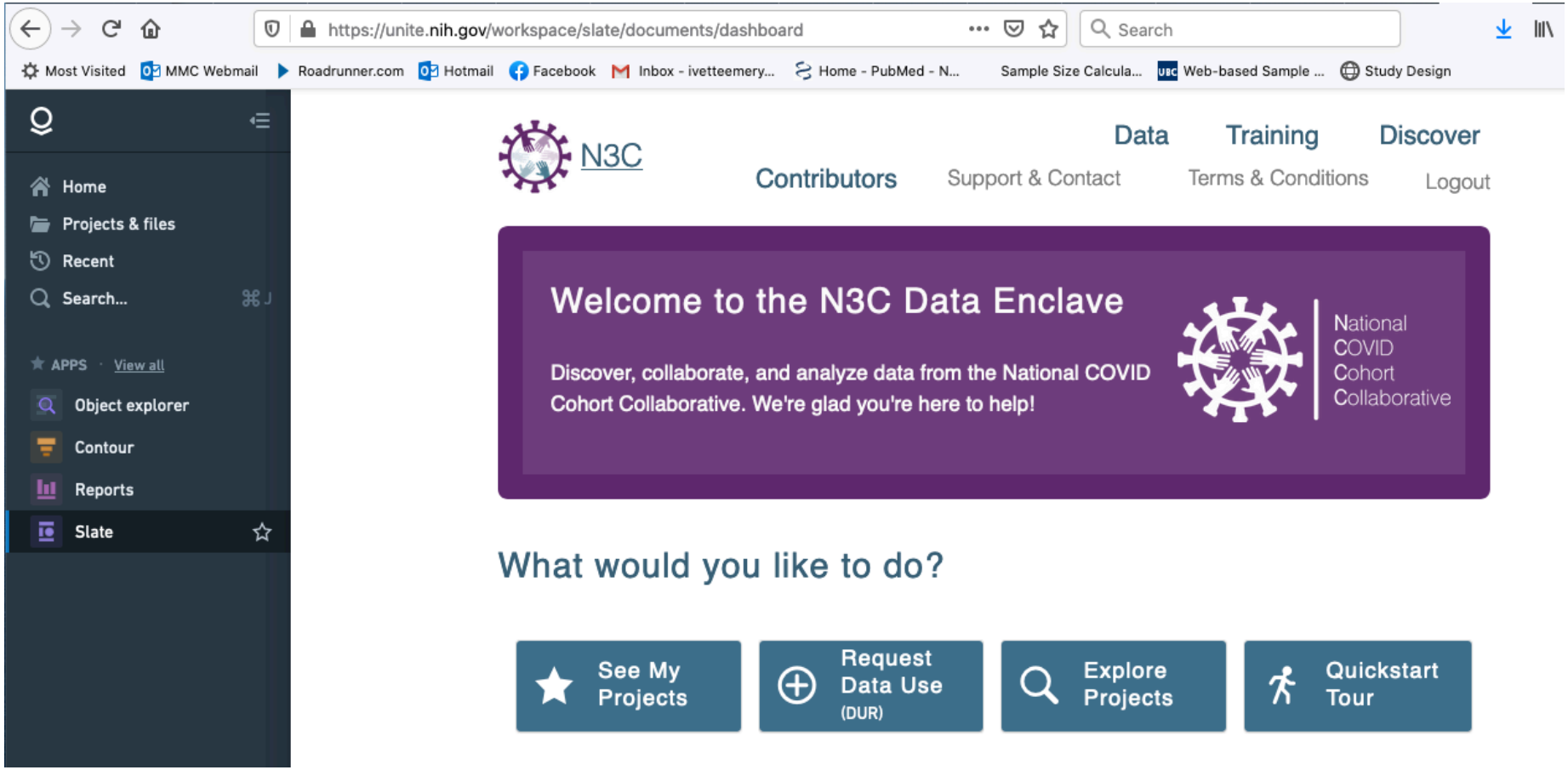
Returning Users

Submit Data Use Request (DUR)

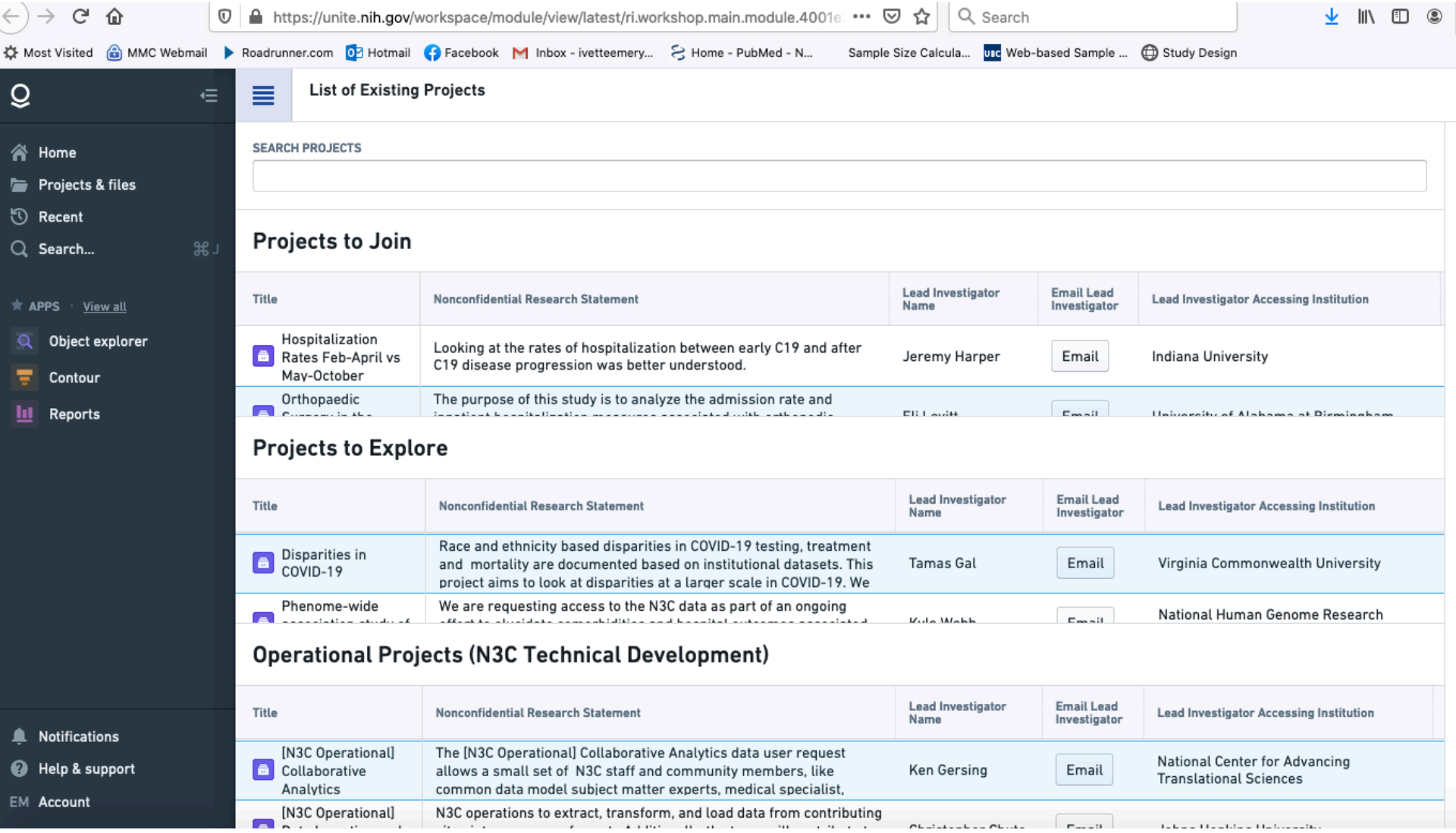
Review by Data Access Committee (DAC)

Your research space within the enclave!

N3C Data Enclave

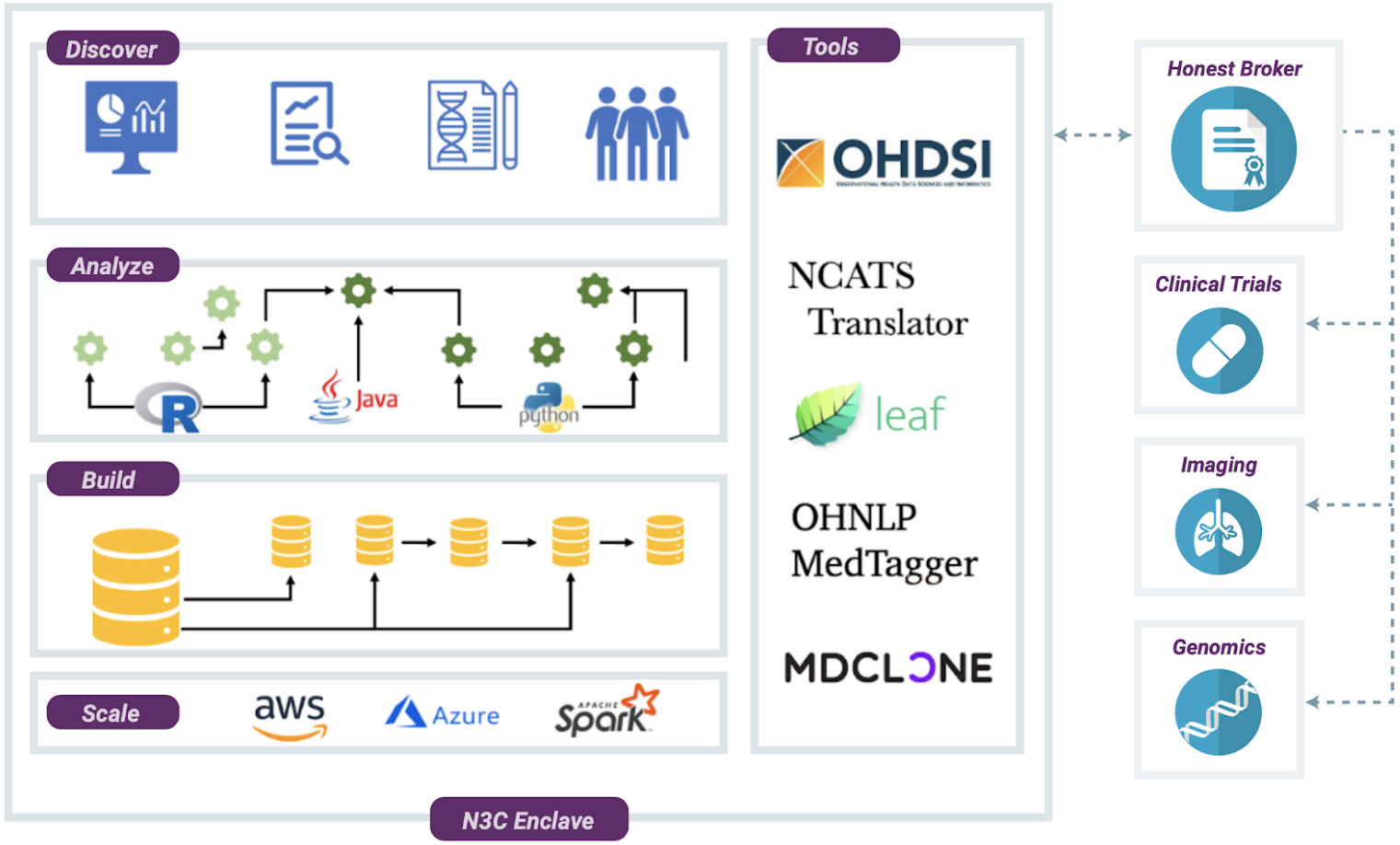


N3C Projects to Explore



N3C Data Enclave

- Collaborative analytics within the N3C secure data enclave: reproducible, transparent, versioned, provenanced, attributed, and shareable analytics on patient-level EHR data
- Analytic tools include Python, R Studio, Jupyter Notebook, and others. N3C also comes with a full set of visualization tools and machine learning tools.



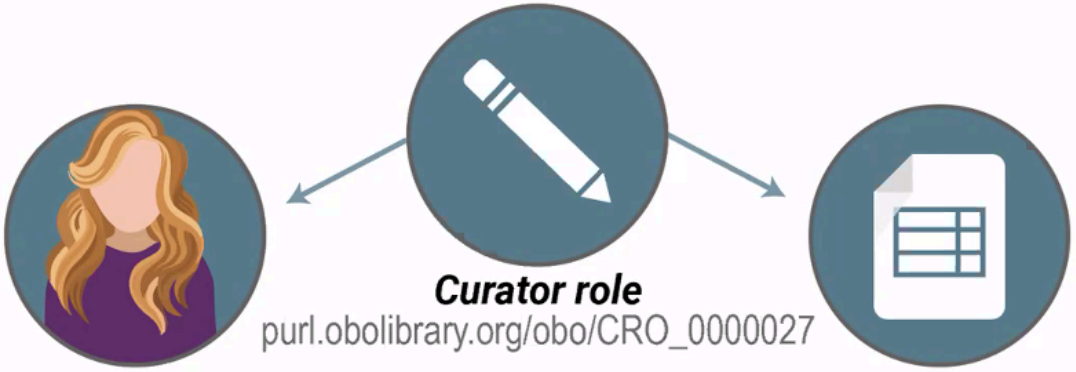
N3C Data Enclave Rules

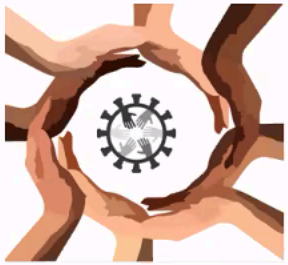
Mission is privacy protection to promote broad access

- COVID-related research only
- No re-identification of individuals or data source
- No download or capture of raw data
  - Results of analyses such as tables, figures, and graphs can be downloaded, as long as they do not include PHI.
- Open platform to all researchers
- Security: activities in the N3C Enclave are recorded and can be audited
- Disclosure of research results to the N3C Enclave for the public good
- Analytics provenance
- Contributor attribution tracking

N3C Provenance and Attribution

Researchers, projects, and artifacts are all linked together in the enclave using the Contributor Attribution Model (CAM).





ACCEPTED MANUSCRIPT

The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment

Journal of the American Medical Informatics Association, ocaa196,
https://doi.org/10.1093/jamia/ocaa196
Published: 17 August 2020 **Article history** ▾

OPEN ACCESS bit.ly/n3c-methods-jamia

Melissa A. Haendel,[1,4,7,8,10,13,14,52,78,101] Christopher G. Chute,[1,4,8,10,13,14,52,78,100,101] Tellen D. Bennett,[9,10,13,14,52,100,101] David A. Eichmann,[4,9,10,13,78,101] Justin Guinney,[4,9,10,14,78,101] Warren A. Kibbe,[9,10,52,78,101] Philip R.O. Payne,[4,9,10,78,101] Emily R. Pfaff,[9,10,13,15,52,78] Peter N. Robinson,[4,9,10,15,52,78,100] Joel H. Saltz,[10,13,14,15,52,78,101] Heidi Spratt,[9,10,100] Christine Suver,[10,78,101] John Wilbanks,[10,78,101] Adam B. Wilcox,[10,101] Andrew E. Williams,[10,13,78] Chunlei Wu,[9,13,14,78] Clair Blacketer,[15,52] Robert L. Bradford,[9,52] James J. Cimino,[10,14,101] Marshall Clark,[9,15,52] Evan W. Colmenares,[9,15,52] Patricia A. Francis,[78] Davera Gabriel,[9,10,13,14,15,52] Alexis Graves,[7,9,78] Raju Hemadri,[9,15,52] Stephanie S. Hong,[9,15,52] George Hripscak,[10,52] Dazhi Jiao,[9,15,52] Jeffrey G. Klann,[14,52,101] Kristin Kostka,[9,15,52] Adam M. Lee,[9,15,52] Harold P. Lehmann,[9,15,52] Lora Lingrey,[9,15,52] Robert T. Miller,[9,15,52] Michele Morris,[9,15,52] Shawn N. Murphy,[9,15,52] Karthik Natarajan,[9,15,52] Matvey B. Palchuk,[9,15,52] Usman Sheikh,[9,78] Harold Solbrig,[9,15,52] Shyam Visweswaran,[10,15,52,101] Anita Walden,[7,10,13,14,52,101] Kellie M. Walters,[10,14,101] Griffin M. Weber,[10,101] Xiaohan Tanner Zhang,[9,15,52] Richard L. Zhu,[9,15,52] Benjamin Amor,[78] Andrew T. Girvin,[15,78] Amin Manna,[78] Nabeel Qureshi,[15,78] Michael G. Kurilla,[10,78] Sam G. Michael,[10,78] Lili M. Portilla,[101] Joni L. Rutter,[1,101] Christopher P. Austin,[101] Ken R. Gersing,[78,101]

Shaymaa Al-Shukri,[4,15] Adil Alaoui,[101] Ahmad Baghal,[15] Pamela D. Banning,[15,100] Edward M. Barbour,[8,15] Michael J. Becich,[15,52,101] Afshin Beheshti,[14] Gordon R. Bernard,[8,15] Sharmodeep Bhattacharyya,[100] Mark M. Bissell,[9,15] L. Ebony Boulware,[14,100] Samuel Bozzette,[100,101] Donald E. Brown,[101] John B. Buse,[14] Brian J. Bush,[8,101] Tiffany J. Callahan,[14,52] Thomas R. Campion,[8,15] Elena Casiraghi,[9,15] Ammar A. Chaudhry,[13,14] Guanhua Chen,[9] Anjun Chen,[13] Gari D. Clifford,[8,15] Megan P. Coffee,[14,100] Tom Conlin,[14] Connor Cook,[7,78] Keith A. Crandall,[9,14,101] Mariam Deacy,[78] Racquel R. Dietz,[78] Nicholas J. Dobbins,[8,9] Peter L. Elkin,[15,52,100] Peter J. Embi,[52,101] Julio C. Facelli,[8,15] Karamarie Fecho,[13] Xue Feng,[9] Randi E. Foraker,[8,13,15] Tamas S. Gal,[8,15] Linqiang Ge,[14] George Golovko,[15,101] Ramkiran Gouripeddi,[14,15] Casey S. Greene,[13,14] Sangeeta Gupta,[52,101] Ashish Gupta,[13,101] Janos G. Hajagos,[9,15] David A. Hanauer,[15,52] Jeremy Richard Harper,[9,14,52] Nomi L. Harris,[14] Paul A. Harris,[101] Mehadi R. Hassan,[9] Yongqun He,[15,52,100] Elaine L. Hill,[9,14] Maureen E. Hoatlin,[14] Kristi L. Holmes,[4,101] LaRon Hughes,[14] Randeep S. Jawa,[14] Guoqian Jiang,[14] Xia Jing,[7,14] Marcin P. Joachimiak,[8,15] Steven G. Johnson,[9,14,101] Rishikesan Kamaleswaran,[9,15,78] Thomas George Kannampallil,[15,101] Andrew S. Kanter,[15,52] Ramakanth Kavuluru,[9,13,14] Kamil Khanipov,[8,14] Hadi Kharrazi,[9,14] Dongkyu Kim,[15,52] Boyd M. Knosp,[8,15] Arunkumar Krishnan,[9] Tahsin Kurc,[9,15] Albert M. Lai,[101] Christophe G. Lambert,[52,101] Michael Larionov,[14] Stephen B. Lee,[1,14] Michael D. Lesh,[9] Olivier Lichtarge,[14] John Liu,[9] Sijia Liu,[8,9,101] Hongfang Liu,[9,15] Johanna J. Loomba,[1,15,78,101] Sandeep K. Mallipattu,[9,14,15] Chaitanya K. Mamillapalli,[14] Christopher E. Mason,[15] Jomol P. Mathew,[8,15,52] James C. McClay,[101] Julie A. McMurry,[1,4,7,9,13,14,78] Paras P. Mehta,[14] Ofer Mendelevitch,[9] Stephane Meystre,[8,14,15] Richard A. Moffitt,[9,13,15] Jason H. Moore,[8,9] Hiroki Morizono,[13,14,15,52] Christopher J. Mungall,[15,52] Monica C. Munoz-Torres,[7,10,78] Andrew J. Neumann,[78] Xia Ning,[14] Jennifer E. Nyland,[13,14] Lisa O'Keefe,[78] Anna O'Malley,[78] Shawn T. O'Neil,[78] Jihad S. Obeid,[10,14,15] Elizabeth L. Ogburn,[13] Jimmy Phuong,[9,15,52,100,101] Jose D Posada,[8,15] Prateek Prasanna,[14,52] Fred Prior,[9,14,15] Justin Prosser,[9,78] Amanda Lienau Purnell,[101] Ali Rahnavard,[9,52] Harish Ramadas,[9,52,78] Justin T. Reese,[9,10] Jennifer L. Robinson,[14,100] Daniel L. Rubin,[101] Cody D. Rutherford,[9,101] Eugene M. Sadhu,[8,15] Amit Saha,[9] Mary Morrison Saltz,[15,52,101] Thomas Schaffter,[78] Titus KL Schleyer,[14] Soko Setoguchi,[8,14,15] Nigam H. Shah,[8,14] Noha Sharafeldin,[14] Evan Sholle,[15,52] Jonathan C. Silverstein,[15,52,101] Anthony Solomonides,[101] Julian Solway,[14,101] Jing Su,[101] Vignesh Subbian,[9,52,101] Hyo Jung Tak,[15] Bradley W. Taylor,[9,14] Anne E. Thessen,[14,101] Jason A. Thomas,[15] Umit Topaloglu,[15,52] Deepak R. Unni,[8,9,15,52] Joshua T. Vogelstein,[14] Andréa M. Volz,[7] David A. Williams,[14,15] Kelli M. Wilson,[9,78] Clark B. Xu,[8,9,15] Hua Xu,[9,10,14] Yao Yan,[9,15,52] Elizabeth Zak,[8,15] Lanjing Zhang,[101] Chengda Zhang,[14] Jingyi Zheng,[14]

[1]CREDIT_00000001 (Conceptualization) [4]CREDIT_00000004 (Funding acquisition) [7]CRO_0000007 (Marketing and Communications) [8]CREDIT_00000008 (Resources) [9]CREDIT_00000009 (Software role) [10]CREDIT_00000010 (Supervision role) [13]CREDIT_00000013 (Original draft) [14]CREDIT_00000014 (Review and editing) [15]CRO_0000015 (Data role) [52]CRO_0000052 (Standards role) [78]CRO_0000078 (Infrastructure role) [100]Clinical Use Cases [101]Governance

Journal List › medRxiv › PMC7814838

NIH National Institutes of Health
Turning Discovery Into Health

Preprint
Made publicly accessible
prior to peer review

Version 1. medRxiv. Preprint. 2021 Jan 13.
doi: 10.1101/2021.01.12.21249511

PMCID: PMC7814838
PMID: 33469592

This article is a preprint.
Preprints have not been peer reviewed.
To learn more about preprints in PMC see: NIH Preprint Pilot.

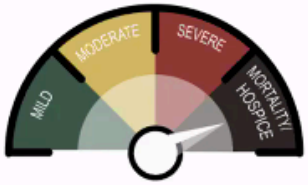
Preprint

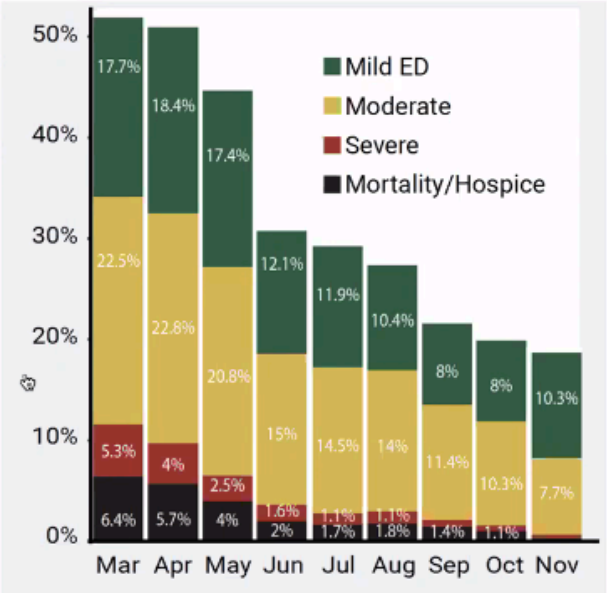
The National COVID Cohort Collaborative: Clinical Characterization and Early Severity Prediction

Tellen D. Bennett, Richard A. Moffitt, Janos G. Hajagos, Benjamin Amor, Adit Anand, Mark M. Bissell, Katie Rebecca Bradwell, Carolyn Bremer, James Brian Byrd, Alina Denham, Peter E. DeWitt, Davera Gabriel, Brian T. Garibaldi, Andrew T. Girvin, Justin Guinney, Elaine L. Hill, Stephanie S. Hong, Hunter Jimenez, Ramakanth Kavuluru, Kristin Kostka, Harold P. Lehmann, Eli Levitt, Sandeep K. Mallipattu, Amin Manna, Julie A. McMurry, Michele Morris, John Muschelli, Andrew J. Neumann, Matvey B. Palchuk, Emily R. Pfaff, Jingjing Qian, Nabeel Qureshi, Seth Russell, Heidi Spratt, Anita Walden, Andrew E. Williams, Jacob T. Wooldridge, Yun Jae Yoo, Xiaohan Tanner Zhang, Richard L. Zhu, Christopher P. Austin, Joel H. Saltz, Ken R. Gersing, Melissa A. Haendel, and Christopher G. Chute

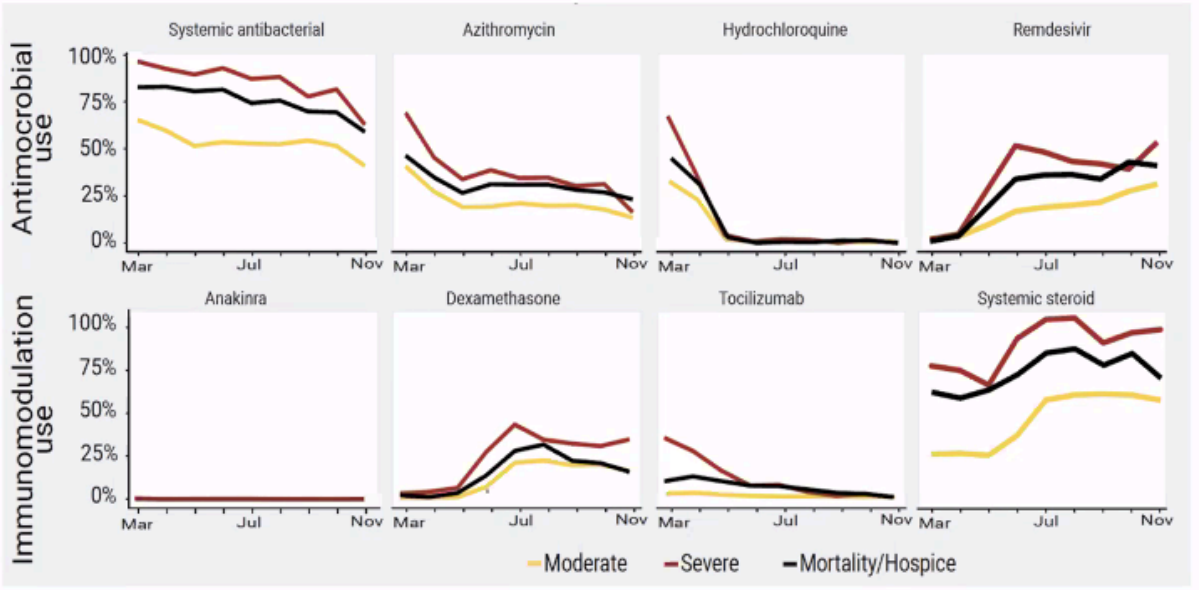
N3C Cohort Characterization: Severity has decreased & use of antimicrobials and immunomodulation have shifted

Severity





Medications



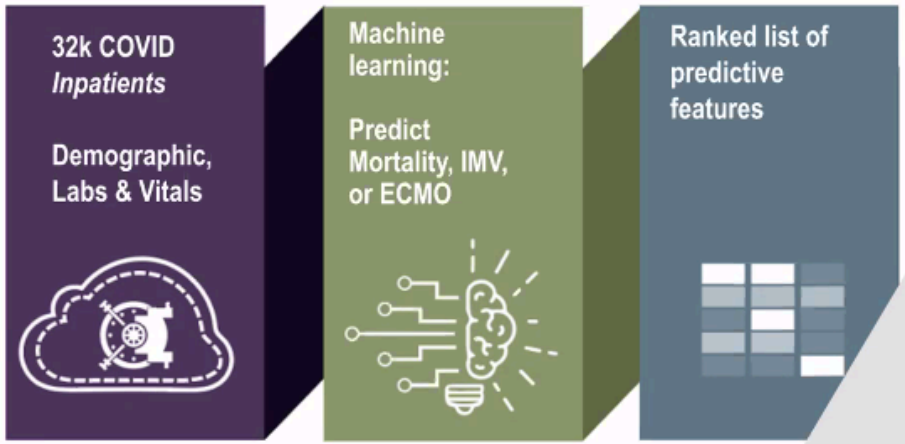
National COVID Cohort Collaborative

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

First-Day ML Models to Predict Patient Severity

bit.ly/cohort-preprint

Tellen Bennett
U Colorado



Random Forest Top 10 Predictors

- Age
- AST
- BUN
- Cr
- Glucose
- pH
- RR
- SBP
- SpO2
- WBC

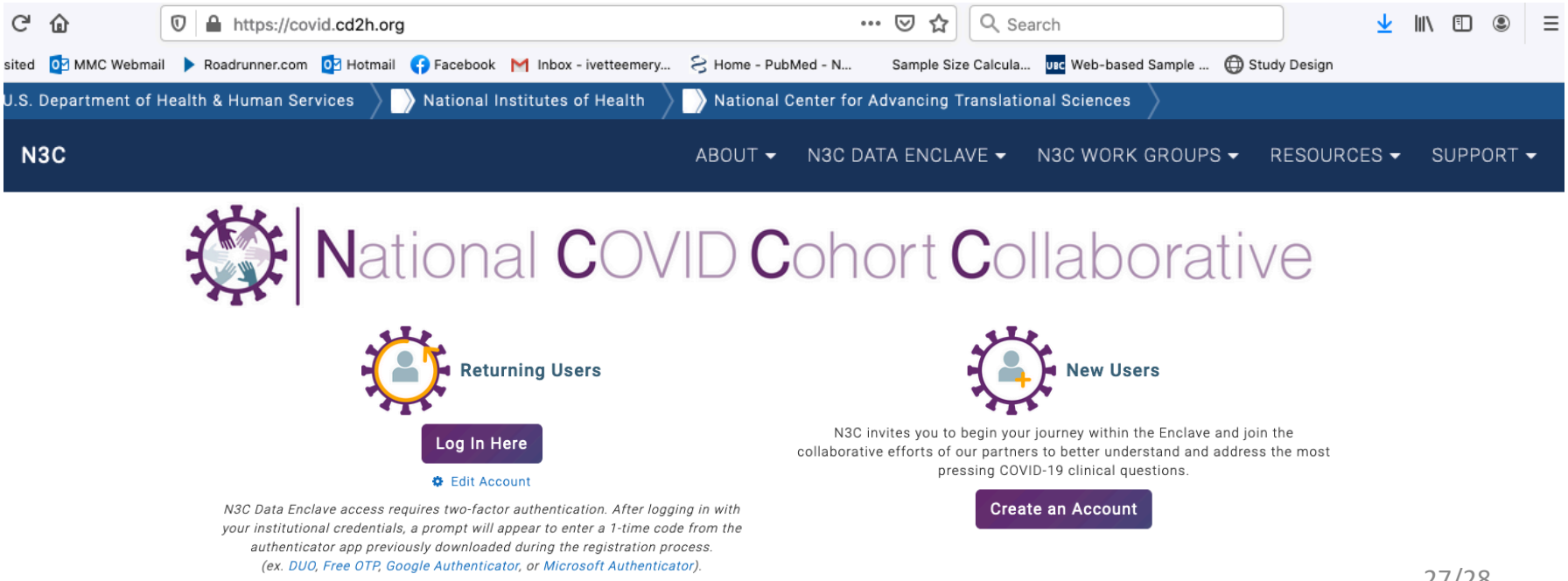
✔ **High confidence.**
✔ **Repeatable.**
✔ **Clinically relevant.**
✔ **Readily deployable.**
✔ **Easily refined.**

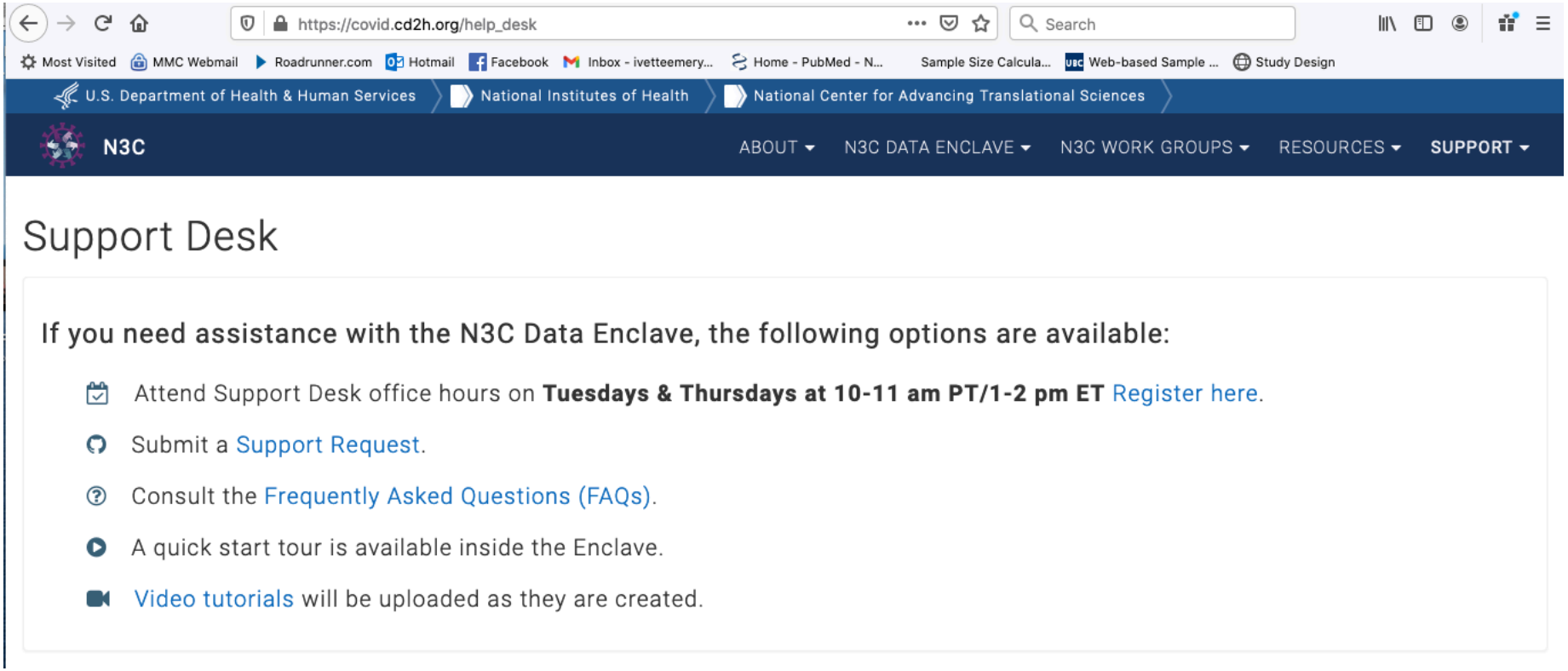
RF **Mar-May** AUROC: 0.864
RF **Jun-Oct** AUROC: 0.865

Variable	Random Forest	XG Boost	Logistic Regression: None	Logistic Regression: L1	Logistic Regression: L2	Mean
pH	0	0	1	1	1	0.6
Age at visit start (years)	3	4	0	0	0	1.4
Respiratory rate	7	3	2	2	2	3.2
Oxygen saturation (SpO2)	2	2	5	4	3	3.2
Systolic blood pressure (SBP)	9	8	4	3	5	5.8
Blood urea nitrogen (BUN)	1	1	11	11	10	6.8
Albumin	21	6	6	5	4	8.4
Lactate	16	9	7	7	7	9.6
C-reactive protein (CRP)	16	11	8	8	8	10.2
Aspartate aminotransferase (AST)	6	5	12	22	18	12.6
Absolute neutrophil count	12	25	9	9	9	12.8
Ethnicity = missing or unknown	45	7	3	6	6	13.4
Glucose	5	18	16	15	14	13.6
Platelet count	10	22	14	13	13	14.4
Diastolic blood pressure (DBP)	13	15	21	17	16	16.4
B-type natriuretic peptide (BNP)	32	12	18	14	15	18.2
Sodium	15	27	15	18	17	18.4
Troponin	14	10	30	21	23	19.6
Erythrocyte sedimentation rate (ESR)	36	30	13	12	12	20.6
Sex = male	38	38	10	10	11	21
Body weight	19	49	19	16	19	24.4
Hemoglobin	17	51	24	20	21	26.6
Charlson Diabetes mellitus	37	24	25	27	24	27.4
Creatinine	8	46	26	32	26	27.6
D-dimer	25	42	31	23	20	28.2
Ferritin	20	38	34	25	25	28.4
Bilirubin total	25	50	29	19	22	29.6
Charlson Dementia	42	53	37	26	26	29.6
Temperature	26	43	32	26	29	31.2
Potassium	23	53	33	29	27	33
Body mass index	27	41	38	31	31	33.6
Bilirubin conjugated	34	28	40	35	34	34.2
Charlson Q Score	30	20	48	38	40	35.2
Charlson Metastases	55	31	36	24	30	35.4
Absolute lymphocyte count	11	14	56	48	51	36
Charlson Renal Disease	44	23	39	39	36	36.2
Alanine aminotransferase (ALT)	24	48	23	57	43	37.4
Heart rate	31	33	51	36	37	37.6
Race = missing or unknown	48	32	17	58	42	39
Race = Asian	54	25	55	30	32	39.4
Chloride	22	57	28	60	39	41.2
Charlson Diabetes mellitus with complications	47	19	42	52	47	41.4
White blood cell count	4	17	62	61	63	41.4
Charlson Peripheral vascular disease	46	45	41	37	38	41.8
Hemoglobin - glycosylated (A1C)	33	48	45	43	41	42
Charlson Congestive heart failure	40	29	46	53	50	43.6
Race = Native Hawaiian or Pacific Islander	60	55	44	33	33	45
Race = Black or African-American	43	54	22	59	49	45.4
Charlson Peptic Ulcer Disease	57	59	43	34	35	45.6
Charlson Stroke	49	44	49	42	45	45.8
Procalcitonin	29	21	61	62	57	46
NTproBNP	35	39	53	50	55	46.4
Charlson Liver Disease (mild)	53	34	52	41	53	46.6
Charlson hemiplegia or paralysis	56	36	58	46	56	46.8
Ethnicity = not Hispanic or Latino	41	52	27	63	52	47
Race = white	39	60	20	54	62	47
Charlson Myocardial Infarction	52	47	50	45	48	48.4
Race = other	61	63	35	47	46	50.4
Sex = other	63	62	47	40	44	51.2
Charlson Cancer	51	56	54	48	54	51.8
Charlson Pulmonary disease	59	37	59	56	59	52
Charlson Rheumatologic disease	55	33	57	56	58	52.2
Charlson Liver Disease (severe)	59	58	63	49	60	57.8
Charlson HIV	62	61	60	51	61	59

N3C Participation – Getting Started

- ❑ Complete NIH Information Security and Information Management Training Course.
- ❑ Prepare to document CITI Human Subjects Research Protection Training.
- ❑ Get ORCID ID.
- ❑ Download 2-factor authentication app to your cell phone or tablet.
  - For example DUO, Free OTP, Google Authenticator, and Microsoft Authenticator.



N3C Support

To get started: Ivette Emery (emeryi@mmc.org)
To analyze data: Kim Murray
PLUS: The entire N3C community