National
COVID
Cohort
Collaborative

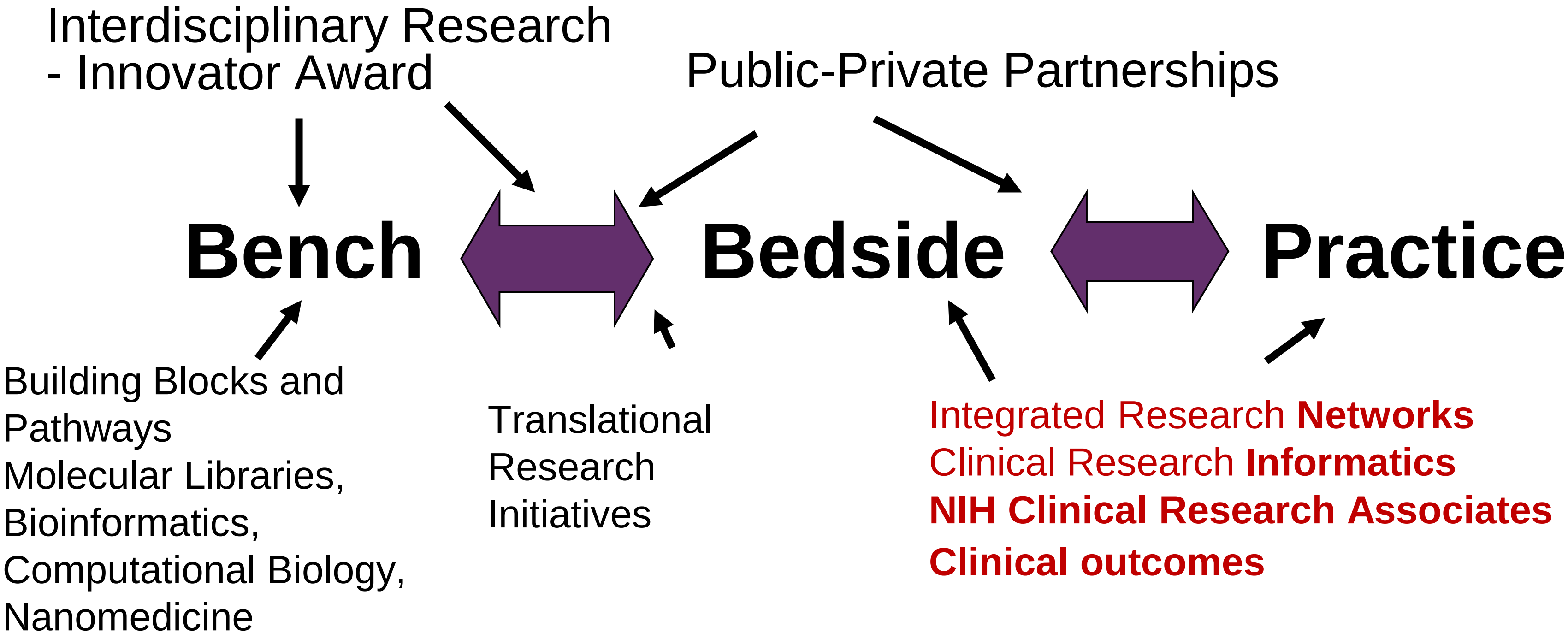
National COVID Cohort Collaborative (N3C)

Data Exchange For Emerging/Novel Diseases (DEFEND)

Internal Team
Rob Star, NIDDK
Ken Gersing, NCATS
Stephen Hewitt, LP, NCI
Michael Kurilla, NCATS
Sam Michael, NCATS
Joni Rutter, NCATS

External Imaging Advisors
Fred Prior, U of Arkansas for Medical Sciences
Joel Saltz, SUNY/Stony Brook

Re-engineering Clinical Research



Interdisciplinary Research
- Innovator Award

Public-Private Partnerships

Bench **Bedside** **Practice**

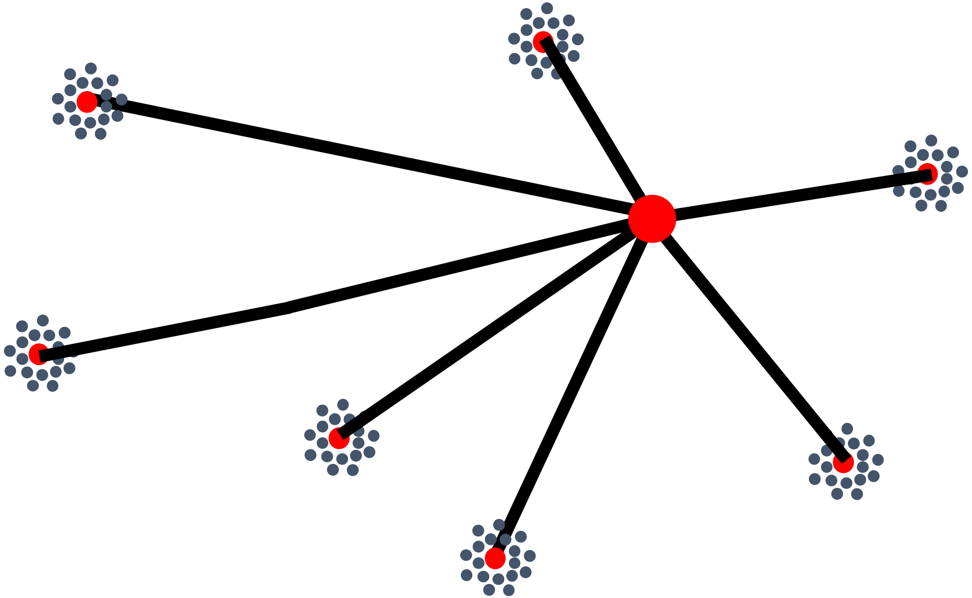
Building Blocks and Pathways
Molecular Libraries,
Bioinformatics,
Computational Biology,
Nanomedicine

Translational
Research
Initiatives

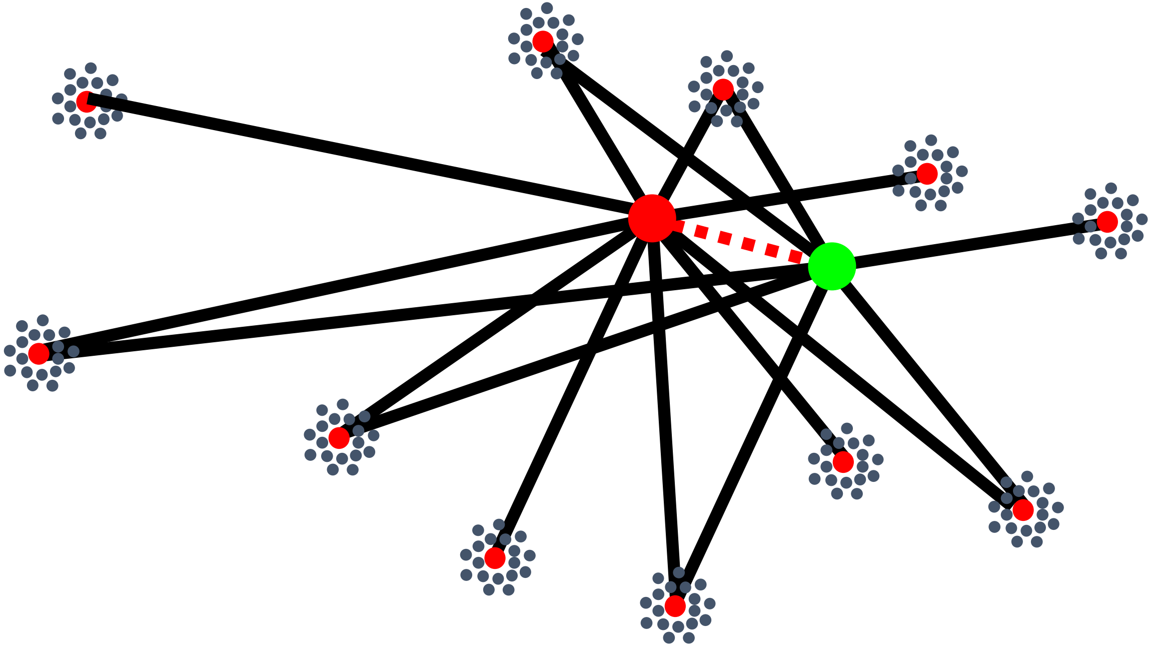
Integrated Research **Networks**
Clinical Research **Informatics**
NIH Clinical Research Associates
Clinical outcomes

Cross cutting: Harmonization, **Training**

Typical NIH Network
Academic Health Center Sites & Data Coordinating Center

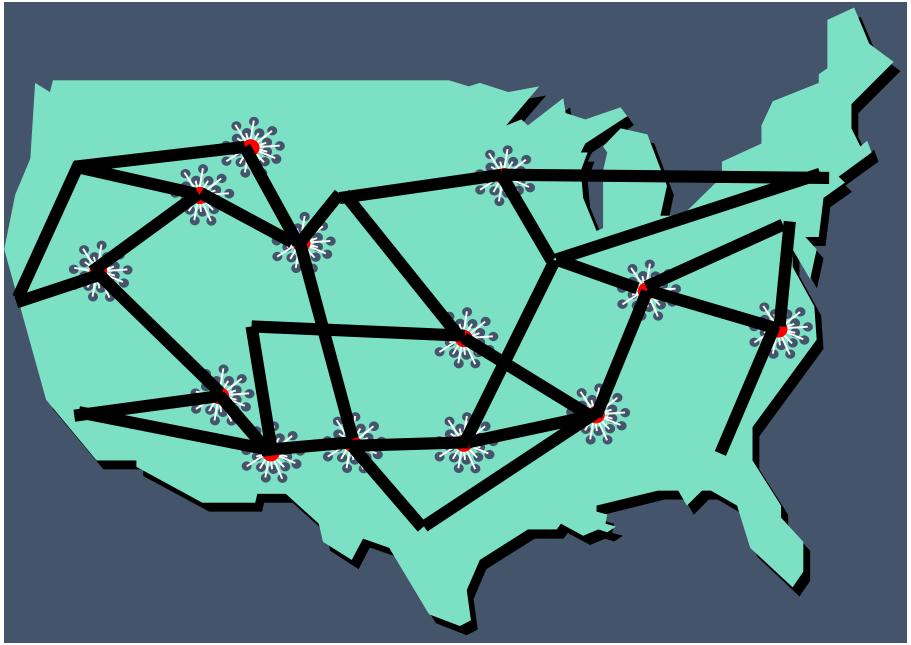


Interoperable Networks Share Sites and Data



Integration of Clinical Research Networks

- Link existing networks so clinical studies and trials can be conducted more effectively
- Ensure that patients, physicians, and scientists form true “Communities of Research”



Re-engineering the Clinical Research Enterprise

Increasing Level of Difficulty

1-3 years	4-7 years	8-10 years
Plan and start a few demonstration networks Simplify complex regulatory systems – demonstration projects Plan for networks in place for all institutes	Funding mechanism to sustain national system through consensus of all constituents ("1% solution") Simplified regulatory system in place for networks	**National Clinical Research System creates effectiveness data that moves rapidly into the community AND data on outcomes and quality of care; sustained efficient infrastructure to rapidly initiate large clinical trials; scientific information for patients, families, advocacy groups**
Establish repositories of biological specimens and standards for collection Standardize nomenclature, data standards, core data, forms for most major diseases Start a library of these elements shared between institutes and NLM Develop efficient network administration infrastructure at NIH Develop standards for capturing images for research	Data standards shared across NIH institutes Funding mechanisms evaluated to determine which are most efficient	ONE medical nomenclature with national data standards (agreed to by NIH, CMS, FDA, DOD, CDC) Data standards updated 'in real time" through networks National repository of images and samples Critical national "problem list" Most efficient network funding mechanisms in place across NIH
Create NIH standards to provide "safe haven" for clinical research Inventory and evaluate existing public-private partnerships, networks, CR institutions, and regulatory systems Establish FORUM(S) of <u>all</u> stakeholders Establish standards for and pilot creation of a National Clinical Research Corps Demonstration/planning grants to enhance/evaluate/develop model networks	NIH standards for safe haven in place Regulations and ethics harmonized with FDA, CMS Public private partnership mechanisms in place 100,000 members of certified "Clinical Research Corps" Standards shared across NIH	Participation in research is a professional standard (taught in all health professions schools) Study, evaluation and training regarding clinical research a part of every medical school, nursing school, pharmacy school Clinical research practices documented and updated regularly to maintain safe haven Networks provide detailed training about network specific issues

Time

2002-3

Re-engineering the Clinical Research Enterprise

Plan and start a few demonstration networks Simplify complex regulatory systems – demonstration projects Plan for networks in place for all institutes	Funding mechanism to sustain national system through consensus of all constituents (“1% solution”) Simplified regulatory system in place for networks	**National Clinical Research System creates effectiveness data that moves rapidly into the community AND data on outcomes and quality of care; sustained efficient infrastructure to rapidly initiate large clinical trials; scientific information for patients, families, advocacy groups**
Establish repositories of biological specimens and standards for collection Standardize no core data, form Start a library c between institu Develop efficie infrastructure a Develop standa research	Data standards shared across NIH institutes	ONE medical nomenclature with national data standards (agreed to by NIH, CMS, ed ‘in real time” mages and samples em list” funding mechanisms
Create NIH sta haven” for clini Inventory and evaluate existing public-private partnerships, networks, CR institutions, and regulatory systems Establish FORUM(S) of <u>all</u> stakeholders Establish standards for and pilot creation of a National Clinical Research Corps Demonstration/planning grants to enhance/evaluate/develop model networks	FDA, CMS Public private partnership mechanisms in place 100,000 members of certified “Clinical Research Corps” Standards shared across NIH	ch is a professional health professions schools) Study, evaluation and training regarding clinical research a part of every medical school, nursing school, pharmacy school Clinical research practices documented and updated regularly to maintain safe haven Networks provide detailed training about network specific issues
1-3 years	4-7 years	8-10 years

Increasing Level of Difficulty

Time

National Clinical Research System creates effectiveness data that moves rapidly into the community AND data on outcomes and quality of care; sustained efficient infrastructure to rapidly initiate large clinical trials; scientific information for patients, families, advocacy groupsz

2002-3

Re-engineering the Clinical Research Enterprise

Increasing Level of Difficulty

Plan and start a few demonstration networks Simplify complex regulatory systems – demonstration projects Plan for networks in place for all institutes	Funding mechanism to sustain national system through consensus of all constituents (“1% solution”) Simplified regulatory system in place for networks	**National Clinical Research System creates effectiveness data that moves rapidly into the community AND data on outcomes and quality of care; sustained efficient infrastructure to rapidly initiate large clinical trials; scientific information for patients, families, advocacy groups**
Establish repositories of biological specimens and standards for collection Standardize no core data, form Start a library c between institu Develop efficie infrastructure a Develop standa research	Data standards shared across NIH institutes	ONE medical nomenclature with national data standards (agreed to by NIH, CMS, ed ‘in real time” mages and samples em list” funding mechanisms
Create NIH sta haven” for clini Inventory and e private partners institutions, and regulatory systems Establish FORUM(S) of all stakeholders Establish standards for and pilot creation of a National Clinical Research Corps Demonstration/planning grants to enhance/evaluate/develop model networks	place 100,000 members of certified “Clinical Research Corps” Standards shared across NIH	h is a professional health professions training regarding clinical research a part of every medical school, nursing school, pharmacy school Clinical research practices documented and updated regularly to maintain safe haven Networks provide detailed training about network specific issues
1-3 years	4-7 years	8-10 years

Time

National Clinical Research System creates **effectiveness data** that **moves rapidly into the community** AND **data on outcomes and quality of care**; sustained **efficient infrastructure** to rapidly initiate large clinical trials; **scientific information for patients, families, advocacy groups**

NCATS

COLLABORATE. INNOVATE. ACCELERATE.

National COVID Cohort Collaborative (N3C)

7/2020

National COVID Cohort Collaborative (N3C)

Goals – Version 2.0

Rapidly collect and aggregate clinical, lab, and imaging **data** from **hospitals, *health plans, and CMS*** at the **peak of the pandemic** and as it **evolves**

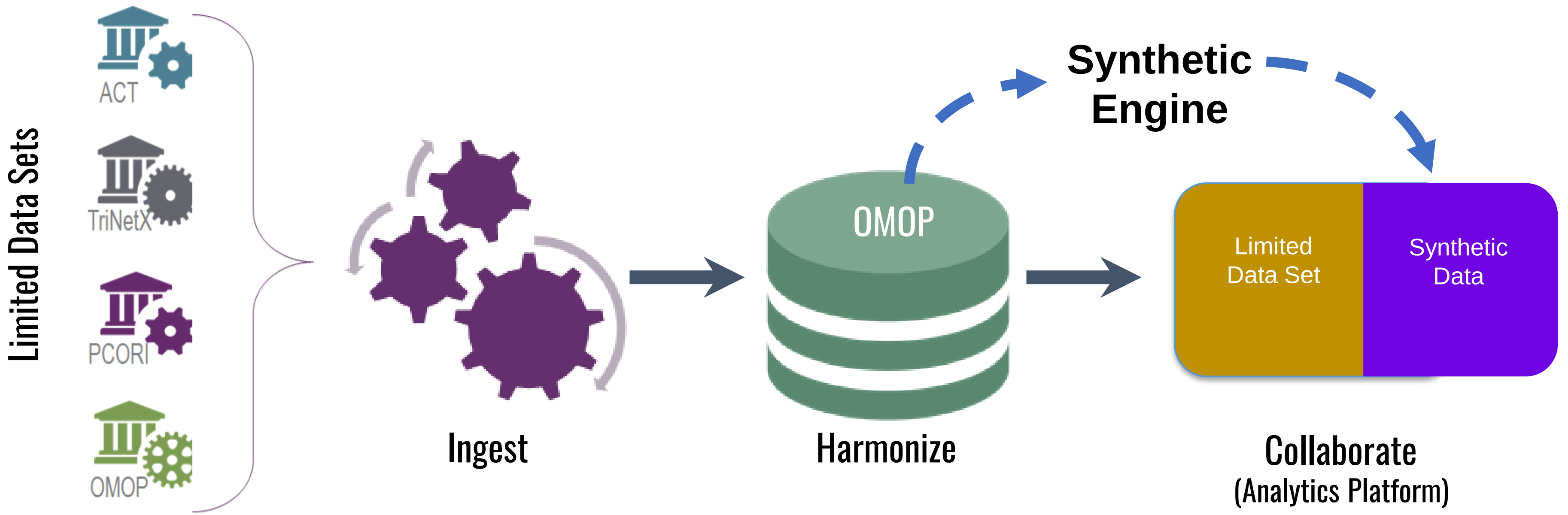
- Provide a **longitudinal dataset** to understand acute **hospital** and **recovery** phases
- Understand **pathophysiology** of disease
- Support **clinical trials** – identify patients who might wish to participate in trials

Develop a **robust, flexible infrastructure** to enable rapid response to COVID-19 and the next emerging threats

- **Speed is critical; leverage existing infrastructure;** poised to collect data immediately
- Analytics platform should be non-proscriptive and easily reconfigurable
- Must be able to interconnect to numerous data streams and analytic resources

National
COVID
Cohort
Collaborative

N3C Overview



Ingest

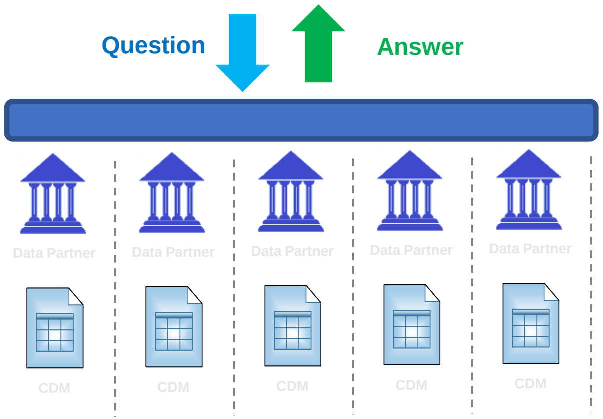
Harmonize

Collaborate
(Analytics Platform)

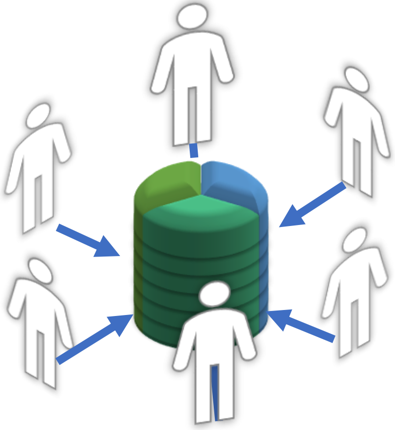
1. Data partnership & governance
2. Data acquisition & Phenotype
3. Data ingest & harmonization
4. Collaborative analytics & FAIR Sharing/Credit

Federated versus Centralized Analytical Models: Characteristics

Federated Model



Centralized Model

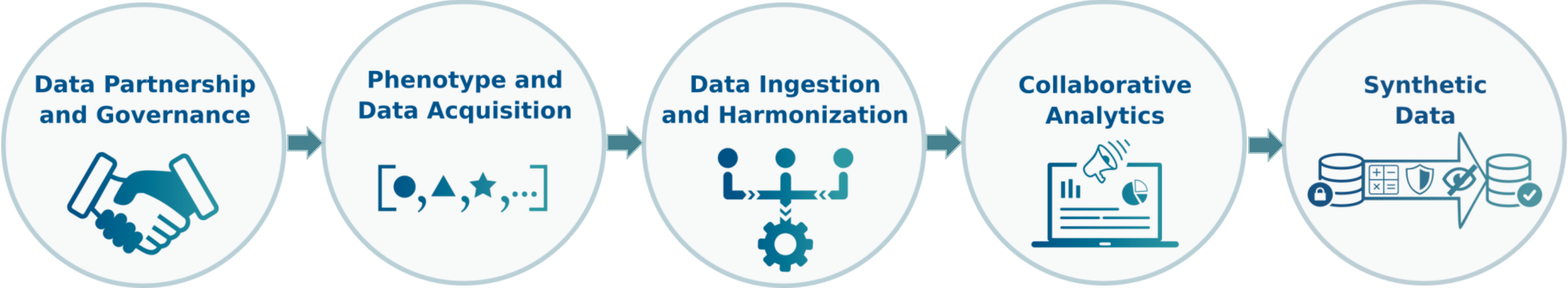


Is **drug X** beneficial to covid-19 patients?
Does **Disease Y** impair course?
Does an **income > $50,000** per year improve outcomes?

What **drugs** help covid-19 patients, and which hinder?
What **Diagnoses** impact outcome?
What **Social Determinants** impact course and outcome?

National
COVID
Cohort
Collaborative

N3C Community Workstreams



Data Partnership and Governance

Phenotype and Data Acquisition

Data Ingestion and Harmonization

Collaborative Analytics

Synthetic Data

NCATS N3C website: ncats.nih.gov/n3c

CD2H N3C website: covid.cd2h.org

Onboarding to N3C: bit.ly/cd2h-onboarding-form

NATIONAL CENTER FOR DATA TO HEALTH

NIH National Center for Advancing Translational Sciences

National
COVID
Cohort
Collaborative

NIH National Center for Advancing Translational Sciences

National
COVID
Cohort
Collaborative

N3C Statistics

7/8/2020
48 DTAs executed
27 IRB protocols approved (23 reliance, 4 local)
24 Regulatory complete (both DTA and IRB)
36 Met with Data Acquisition Group
......9 Deposited data:
..........4 - PCORI
..........3 - OMOP
..........1 - TriNetX
..........1 - ACT

CTSA Organizations	85%
N3C Organizations	105
N3C Individual Members	800

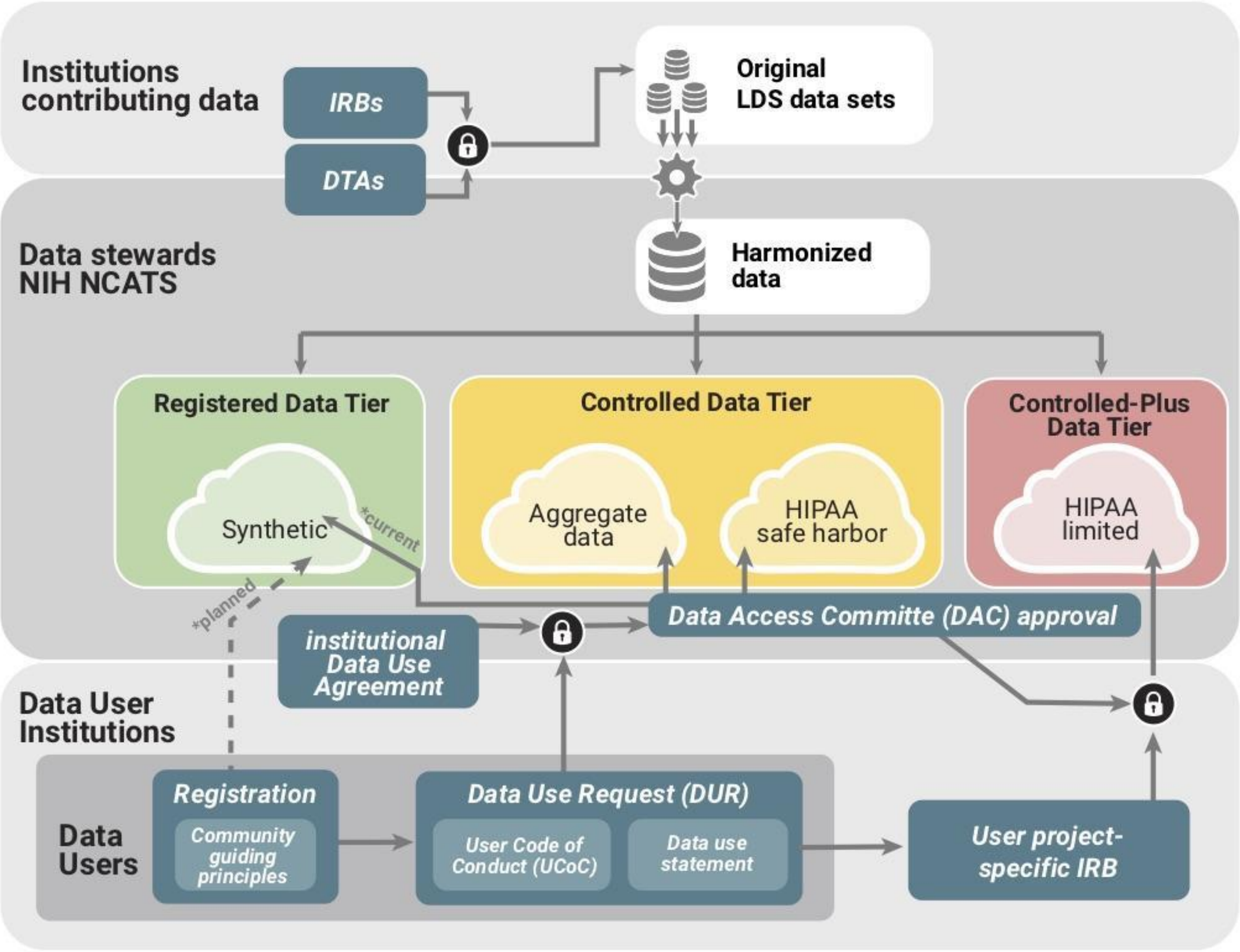
Data Partnership and Governance

Goal of the Data Use Agreement is broad access:

- COVID-Related research only
- **Open platform to all Credentialed 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
- No download of data

National
COVID
Cohort
Collaborative

Regulatory overview



Data Tiers

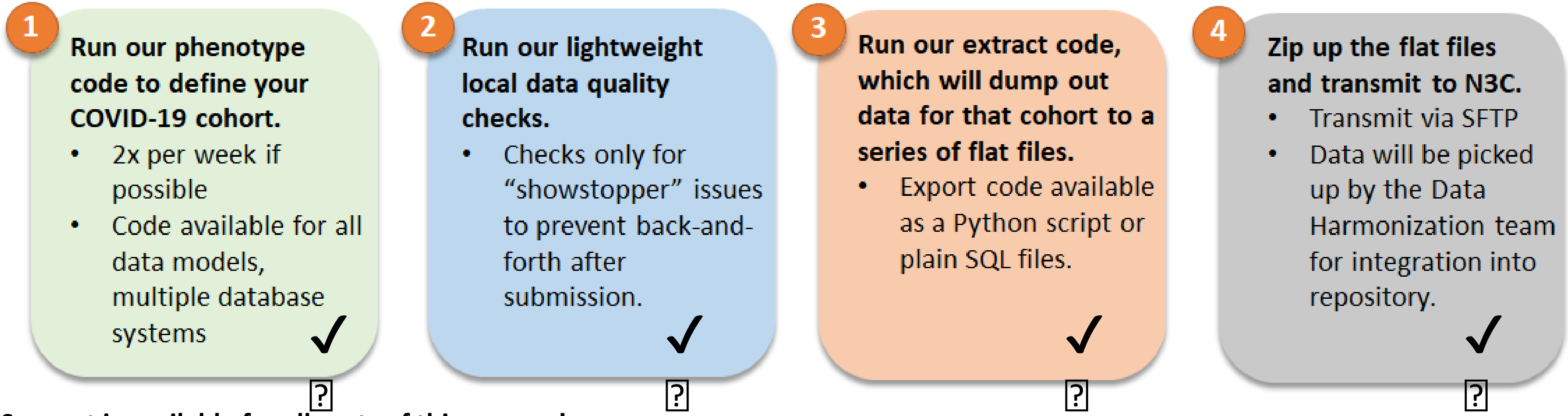
Access Level	**Registered**	**Controlled**		**Controlled-Plus**
Data Type	**Synthetic Data (pending pilot)**	**Aggregate Data (i.e., counts)**	**HIPAA Safe Harbor**	**HIPAA Limited**
Description	Computational data derivative that statistically resembles the original data	Counts and summary statistics representing 10 or more individuals	Data stripped of 18 direct identifiers per HIPAA rules	Data that may contain 3 direct identifiers per HIPAA rules (dates, full zip code, and any age)
Capabilities				
Downloadable data	Planned: pending validation & organizational agreement	Downloadable query results	No	No
Custom software	Yes	Yes - on downloaded query results	Yes with DAC approval	Yes - with independent IRB and DAC approval

Phenotype & Acquisition

Dual-purpose workstream:

1. Work with the community to write and maintain a computable phenotype for COVID-19.
2. Write and maintain a series of scripts to execute the computable phenotype in each of four common data models (CDMs): OMOP, i2b2/ACT, PCORnet, and TriNetX.

What does it look like to run our process locally?



1 **Run our phenotype code to define your COVID-19 cohort.**

- 2x per week if possible
- Code available for all data models, multiple database systems

✓

⍰

2 **Run our lightweight local data quality checks.**

- Checks only for "showstopper" issues to prevent back-and-forth after submission.

✓

⍰

3 **Run our extract code, which will dump out data for that cohort to a series of flat files.**

- Export code available as a Python script or plain SQL files.

✓

⍰

4 **Zip up the flat files and transmit to N3C.**

- Transmit via SFTP
- Data will be picked up by the Data Harmonization team for integration into repository.

✓

⍰

Support is available for all parts of this process!

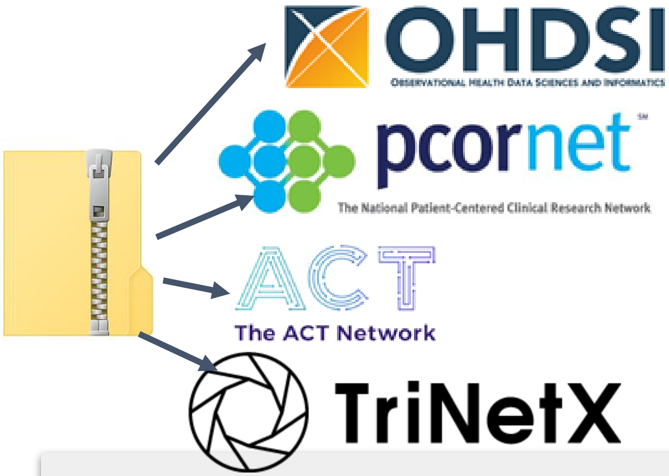
Latest phenotype: covid.cd2h.org/phenotype

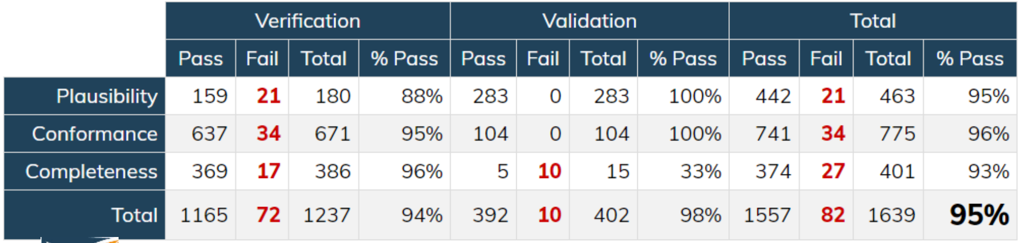
Documentation: covid.cd2h.org/phenotype-wiki

All specifications and software shared on GitHub

Common Data Model Harmonization

ADEPTIA
Workflow


OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS
pcornet
The National Patient-Centered Clinical Research Network
ACT
The ACT Network
TriNetX



	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	159	21	180	88%	283	0	283	100%	442	21	463	95%
Conformance	637	34	671	95%	104	0	104	100%	741	34	775	96%
Completeness	369	17	386	96%	5	10	15	33%	374	27	401	93%
Total	1165	72	1237	94%	392	10	402	98%	1557	82	1639	**95%**

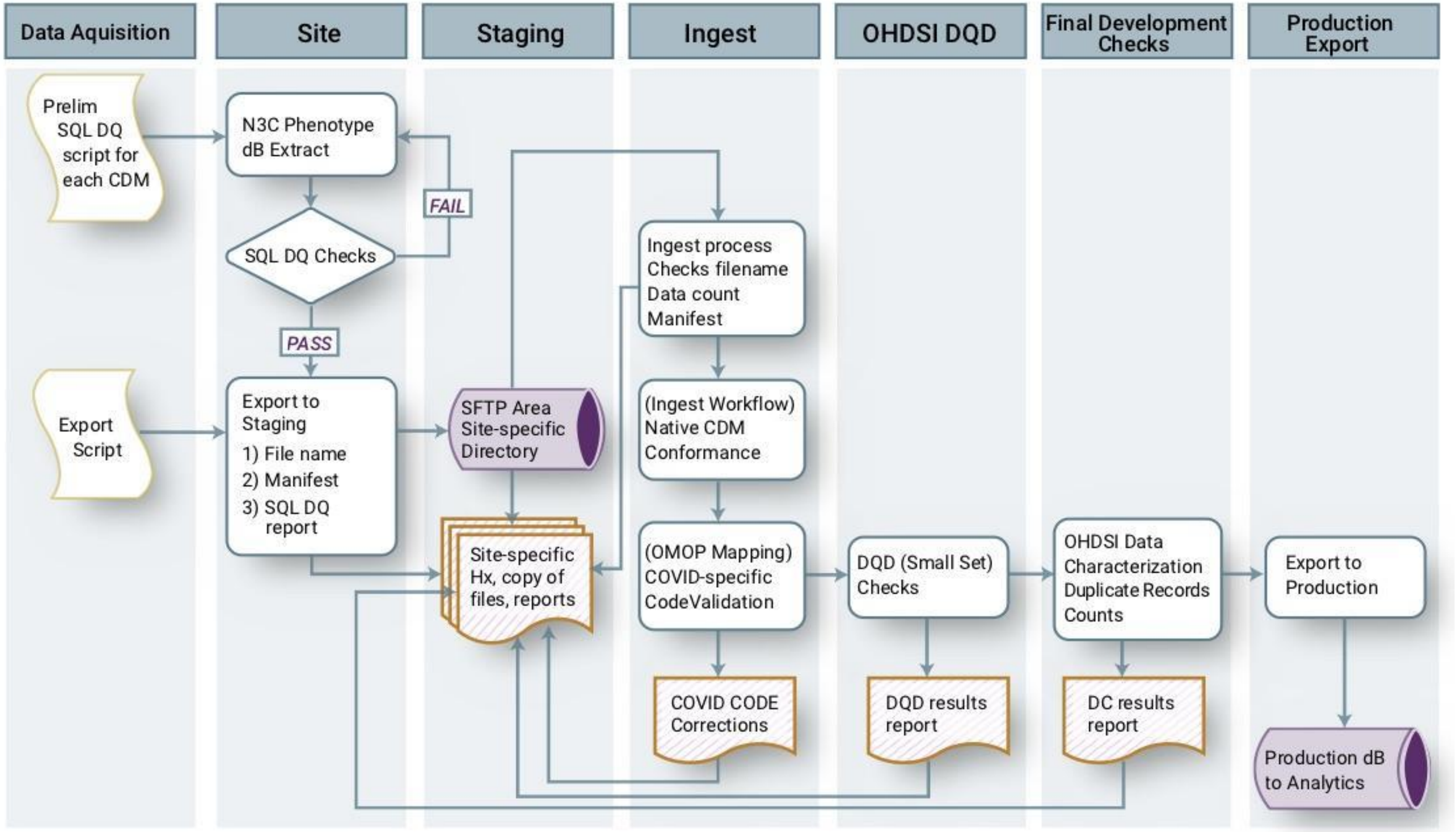
Data Quality Dashboard (shared with site)

First Stage Ingestion

- Unpack Zip’ed csv Files. Check data manifests ✓
- Reconstitute into native CDM formats ✓
- Hybrid Data Quality checks adapting OHDSI Data Quality Dashboard ✓

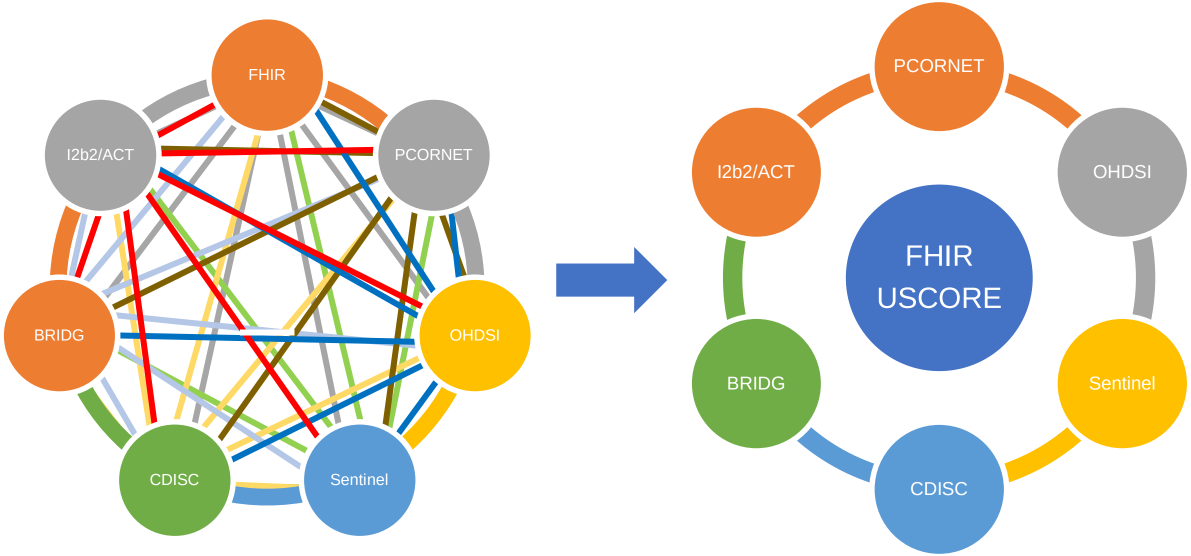
National
COVID
Cohort
Collaborative

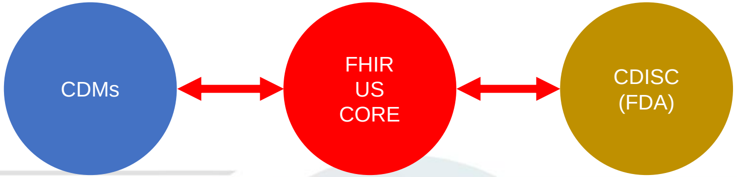
Data Quality Gates


Data Aquisition
Site
Staging
Ingest
OHDSI DQD
Final Development Checks
Production Export
Prelim SQL DQ script for each CDM
N3C Phenotype dB Extract
FAIL
SQL DQ Checks
PASS
Export Script
Export to Staging
1) File name
2) Manifest
3) SQL DQ report
SFTP Area Site-specific Directory
Site-specific Hx, copy of files, reports
Ingest process Checks filename Data count Manifest
(Ingest Workflow) Native CDM Conformance
(OMOP Mapping) COVID-specific CodeValidation
COVID CODE Corrections
DQD (Small Set) Checks
DQD results report
OHDSI Data Characterization Duplicate Records Counts
DC results report
Export to Production
Production dB to Analytics

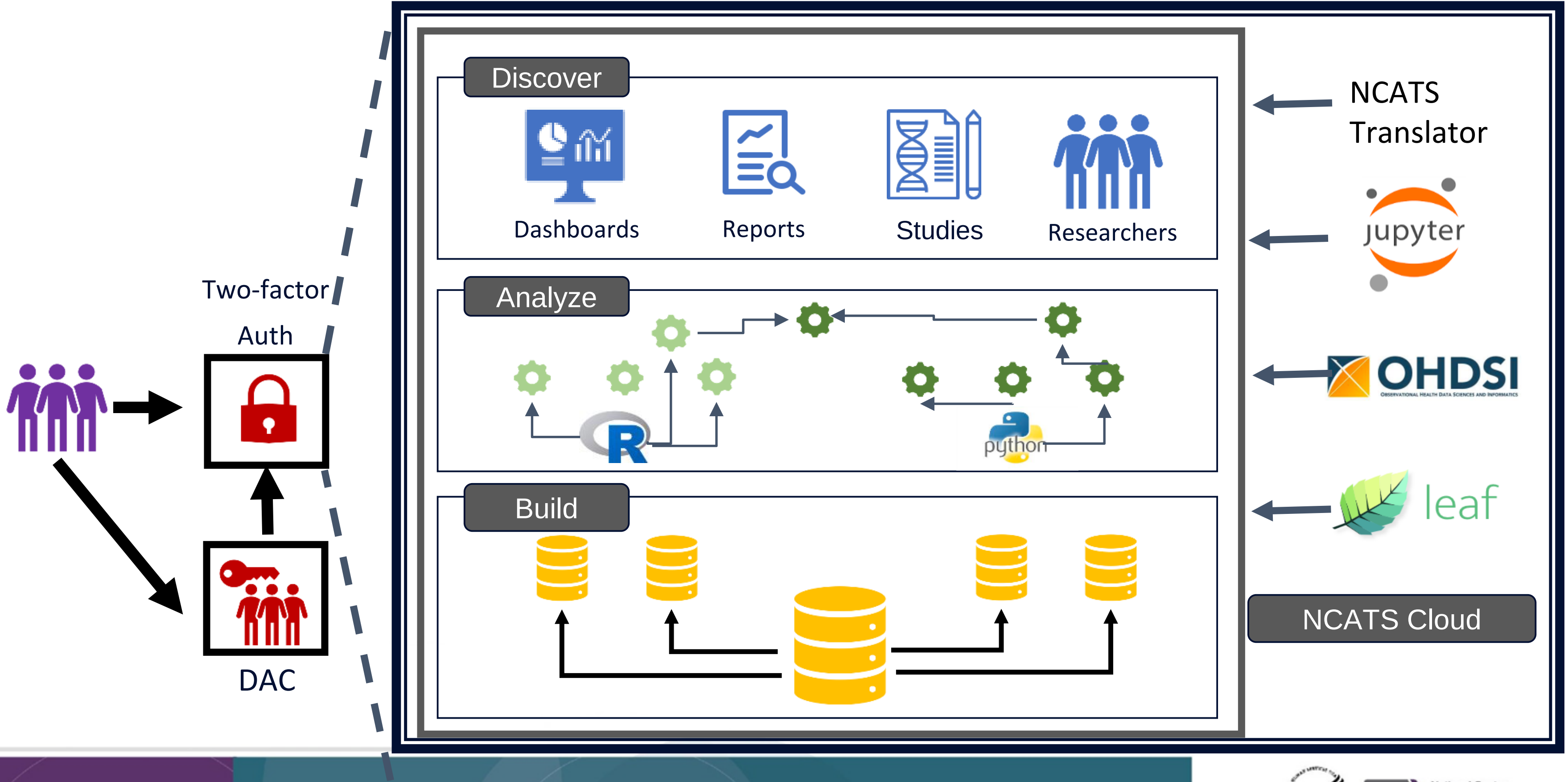
NCATS, FDA, and NCI working together on CDM harmonization

Harmonization of Common data models, (PCORMET, Sentinel, OMOP, ACT) FHIR / USCORE and CDISC
Meta data initiative makes the meaning of data publicly available and reusable in **human and machine-readable**

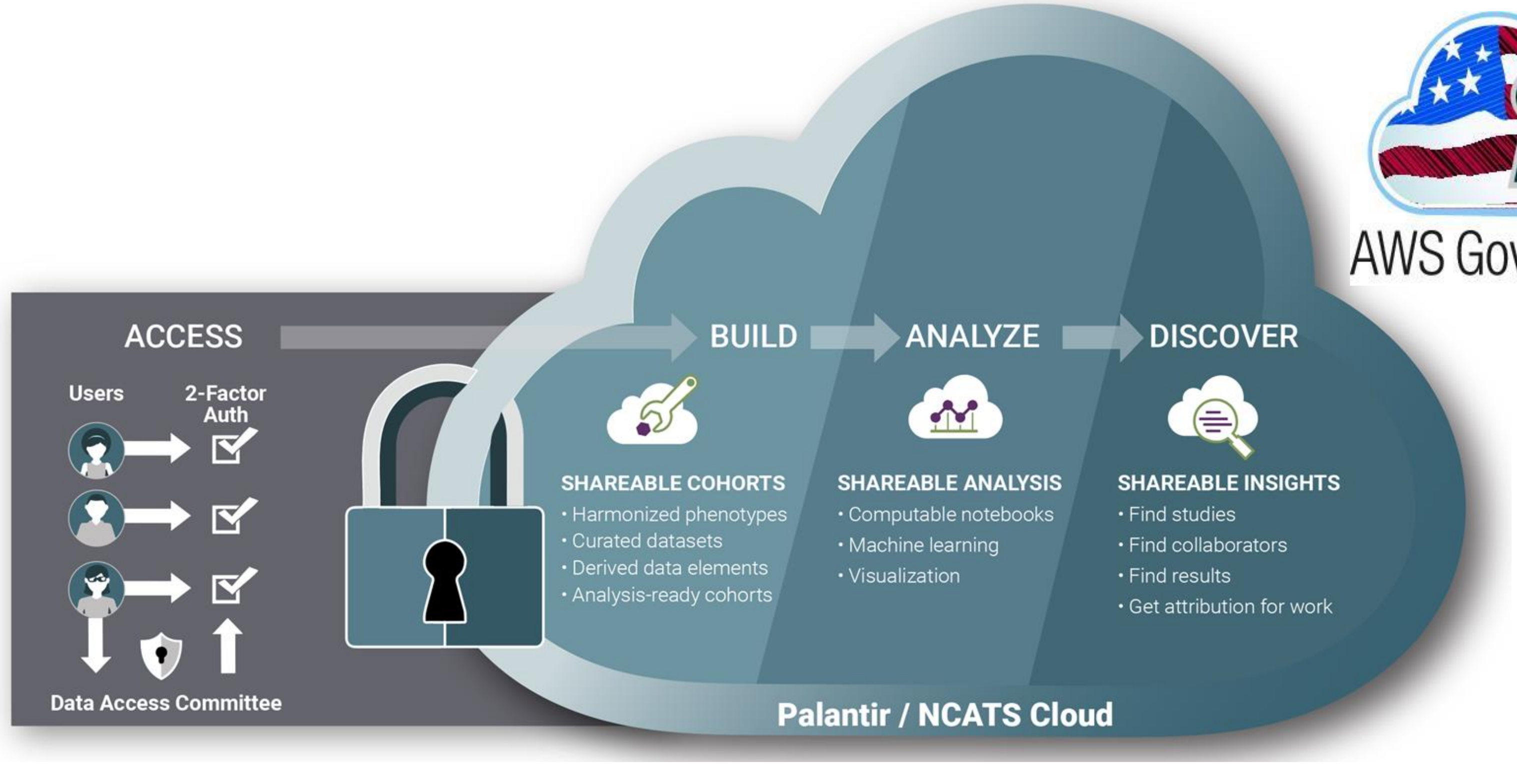




Collaborative Analytics - N3C Secure Data Enclave



Collaborative Analytics - N3C Secure Data Enclave



SHAREABLE COHORTS

- Harmonized phenotypes
- Curated datasets
- Derived data elements
- Analysis-ready cohorts

SHAREABLE ANALYSIS

- Computable notebooks
- Machine learning
- Visualization

SHAREABLE INSIGHTS

- Find studies
- Find collaborators
- Find results
- Get attribution for work

AWS GovCloud (US)

Clinical Scenarios

AKI/ARB/ACE

Critical Care

Short/Long term Complications

Diabetes

Pregnancy

Social Determinants of Health

Immuno-suppressed/ Compromised

Elder Impact

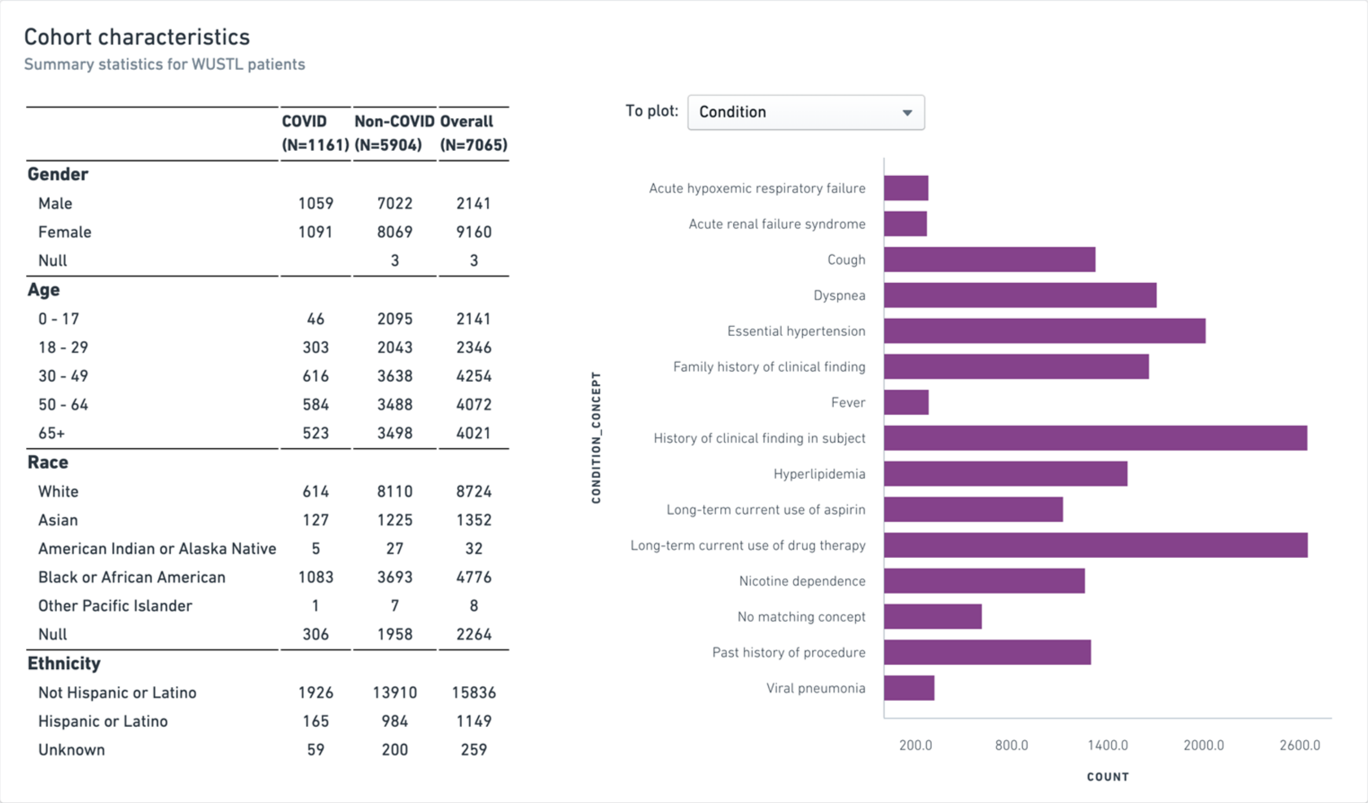
Oncology

Pediatrics

Population Health/Health Policy

Emergency Dept Avoidance Impact

Cohort Characterisation



Cohort characteristics

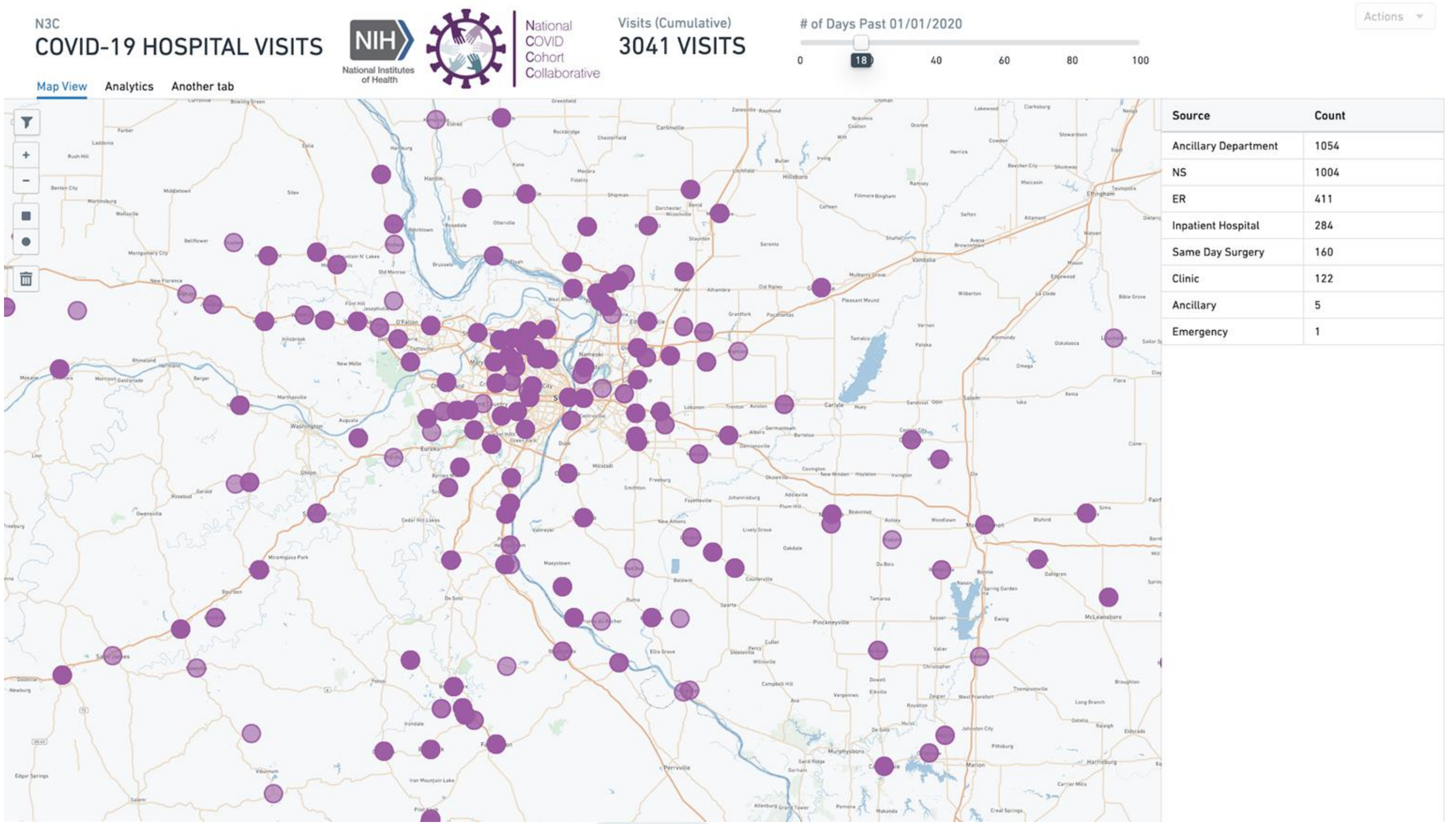
Summary statistics for WUSTL patients

	COVID (N=1161)	Non-COVID (N=5904)	Overall (N=7065)
Gender			
Male	1059	7022	2141
Female	1091	8069	9160
Null		3	3
Age			
0 - 17	46	2095	2141
18 - 29	303	2043	2346
30 - 49	616	3638	4254
50 - 64	584	3488	4072
65+	523	3498	4021
Race			
White	614	8110	8724
Asian	127	1225	1352
American Indian or Alaska Native	5	27	32
Black or African American	1083	3693	4776
Other Pacific Islander	1	7	8
Null	306	1958	2264
Ethnicity			
Not Hispanic or Latino	1926	13910	15836
Hispanic or Latino	165	984	1149
Unknown	59	200	259

To plot: Condition

CONDITION_CONCEPT
Acute hypoxemic respiratory failure
Acute renal failure syndrome
Cough
Dyspnea
Essential hypertension
Family history of clinical finding
Fever
History of clinical finding in subject
Hyperlipidemia
Long-term current use of aspirin
Long-term current use of drug therapy
Nicotine dependence
No matching concept
Past history of procedure
Viral pneumonia
200.0
800.0
1400.0
2000.0
2600.0
COUNT

Time/Space Vector - Live Example



N3C

COVID-19 HOSPITAL VISITS

Visits (Cumulative)

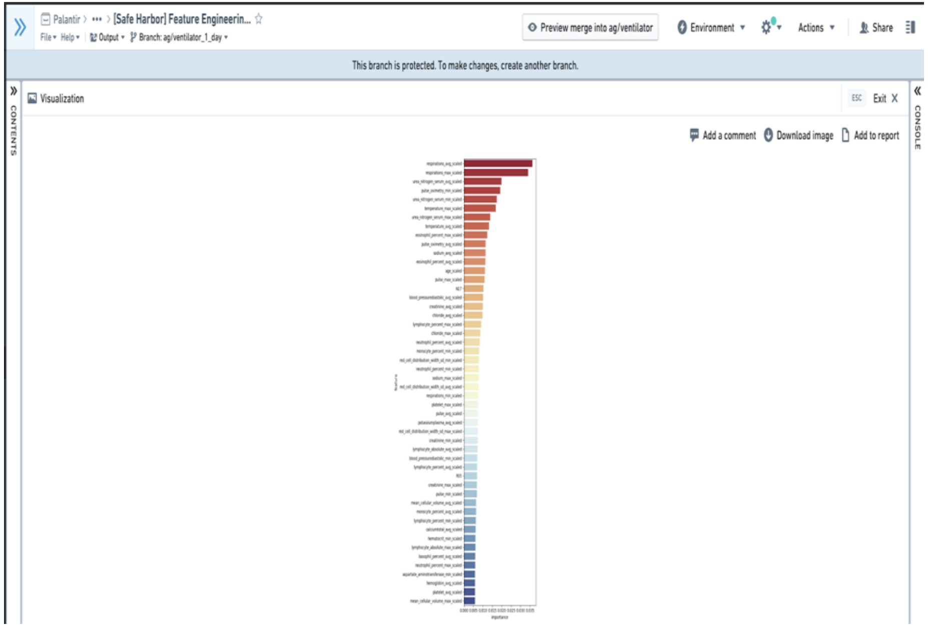
3041 VISITS

# of Days Past 01/01/2020

Source	Count
Ancillary Department	1054
NS	1004
ER	411
Inpatient Hospital	284
Same Day Surgery	160
Clinic	122
Ancillary	5
Emergency	1

National COVID Cohort Collaborative

Predictive Modeling: Risk of Ventilation and AKI

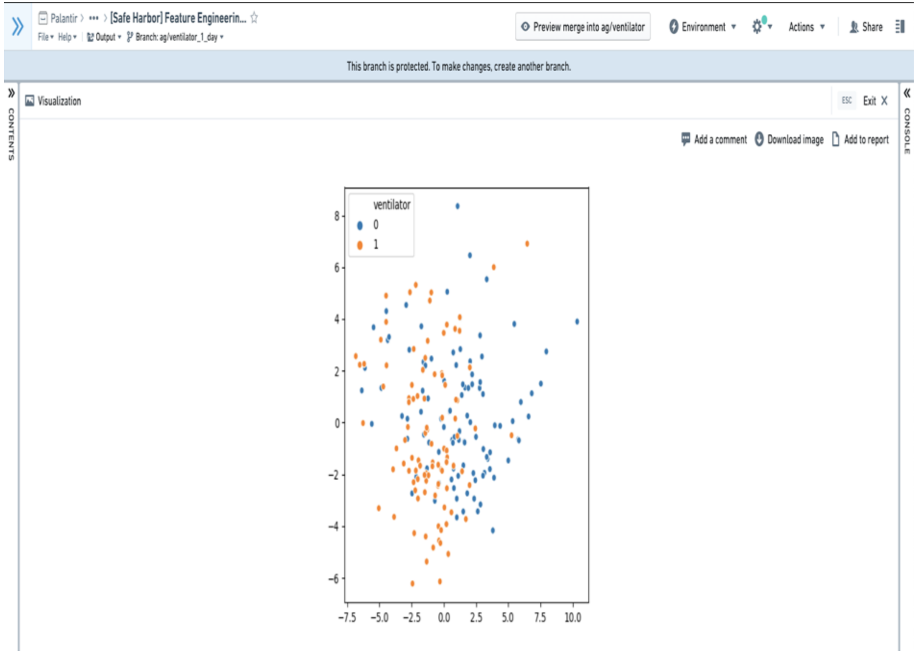

Palantir > ••• > [Safe Harbor] Feature Engineerin... ☆

File ▾ Help ▾ | Output ▾ Branch: ag/ventilator_1_day ▾

This branch is protected. To make changes, create another branch.

Preview merge into ag/ventilator
Environment ▾
Actions ▾
Share
CONTENTS
CONSOLE
Visualization
ESC Exit X
Add a comment
Download image
Add to report
respirations_avg_scaled
respirations_max_scaled
urea_nitrogen_serum_avg_scaled
pulse_oximetry_min_scaled
urea_nitrogen_serum_min_scaled
temperature_max_scaled
urea_nitrogen_serum_max_scaled
temperature_avg_scaled
eosinophil_percent_max_scaled
pulse_oximetry_avg_scaled
sodium_avg_scaled
eosinophil_percent_avg_scaled
age_scaled
pulse_max_scaled
NLR
blood_pressurediastolic_avg_scaled
creatinine_avg_scaled
chloride_avg_scaled
lymphocyte_percent_max_scaled
chloride_max_scaled
neutrophil_percent_avg_scaled
monocyte_percent_min_scaled
red_cell_distribution_width_sd_min_scaled
neutrophil_percent_min_scaled
sodium_max_scaled
red_cell_distribution_width_sd_avg_scaled
respirations_min_scaled
platelet_max_scaled
pulse_avg_scaled
potassiumplasma_avg_scaled
red_cell_distribution_width_sd_max_scaled
creatinine_min_scaled
lymphocyte_absolute_avg_scaled
blood_pressurediastolic_min_scaled
lymphocyte_percent_avg_scaled
RDS
creatinine_max_scaled
pulse_min_scaled
mean_cellular_volume_avg_scaled
monocyte_percent_avg_scaled
lymphocyte_percent_min_scaled
calciumtotal_avg_scaled
hematocrit_min_scaled
lymphocyte_absolute_max_scaled
basophil_percent_avg_scaled
neutrophil_percent_max_scaled
aspartate_aminotransferase_min_scaled
hemoglobin_avg_scaled
platelet_avg_scaled
mean_cellular_volume_max_scaled
feature
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
importance

Random forest model trained on 200 COVID-19 patients, 100 of whom required ventilation, and 100 did not. It performs well, with an AUC of 0.85. Shown are the top features in the model predicting ventilator usage as an outcome.


Palantir > ••• > [Safe Harbor] Feature Engineerin... ☆

File ▾ Help ▾ | Output ▾ Branch: ag/ventilator_1_day ▾

This branch is protected. To make changes, create another branch.

Using these features, we are able to see separation in a PCA plot between the ventilator population in orange and the non-ventilator population in blue.

National Center for Advancing Translational Sciences

Data Sharing Initiative: Synthetic Data

*Computer Derived Synthetic Data: Validation of Sepsis Prediction

Public / Private Partnership

- *Wash University*
- *Microsoft*
- *MDClone*

		Trained on real data Tested on real data	Trained on synthetic data Tested on real data
Train	Accuracy	0.925	0.911
	Precision	0.95	0.925
	Recall	0.817	0.799
	F-Score	0.879	0.858
10-fold cross-validation	Accuracy	0.839	0.816
	Precision	0.802	0.754
	Recall	0.704	0.666
	F-Score	0.745	0.704
Test	Accuracy	0.846	0.841
	Precision	0.836	0.845
	Recall	0.671	0.645
	F-Score	0.745	0.731

ML model performance (random forest)

*Wash. U. Philip Payne

National COVID Cohort Collaborative

Partners, Teams, Collaborators

NCATS
Chris Austin
Joni Rutter
Mike Kurilla
Clare Schmitt
Ken Gersing
Xinzhi Zhang
Erica Rosemond
Sam Bozzette
Lili Portilla
Chris Dillon
Penny Burgoon
Emily Marti
Meredith Temple-O'Connor
Sam Jonson
Christine Cutillo
Nicole Garbarini

NIH & HHS Partners
NCI
Janelle Cortner
Stephen Hewitt
Denise Warzel

FDA
Mitra Rocca
Scott Gideon
Wei Chen

NIDDK
Robert Star

NIGMS
Ming Lee

NCATS ITRB
Sam Michael
Mariam Deacy
Gary Berkson
Josephine Kennedy
Usman Sheikh
Mark Backus
Nam Ngo
Amit Virakatmath
Keats Kirsch
Sulochana Nunna
Rafael Fuentes
Reid Simon
Biju Mathew
Tim Mierzwa
Ke Wang
Kalle Virtaneva

CD2H
OHSU/OSU
Melissa Haendel
Anita Walden
Julie McMurry
Moni Munoz-Torres
Andrea Volz
Connor Cook
Racquel Dietz
Andrew Neumann
Rich Lorimor

Sage Bionetworks
Justin Guinney
James Eddy

U of Iowa:
Dave Eichmann
Alexis Graves

Northwestern:
Kristi Holmes
Justin Starren
Lisa O'Keefe

Washington U.
Philip Payne
Albert Lai
Tom Dillon

CD2H
U. Of Washington
Adam Wilcox
Liz Zampino

Johns Hopkins U
Chris Chute
Tricia Francis

Jax Labs
Peter Robinson

Scripps
Chunlei Wu

Teams
Governance
Sage Bionetworks
John Wilbanks
Christine Suver

Data Harmonization
JHU
Davera Gabriel
Stephanie Hong
Harold Lehmann
Tanner Zhang
Richard Zhu

SAMVIT
Smita Hastak
Charles Yaghmour

NCATS
Raju Hemadri
Nancy Nurthen
Sai Manjula

Adeptia
Sandeep Naredla

Teams
Phenotype & Acquisition
Emily Pfaff, UNC

ACT
Michele Morris, Pitt
Shyam Visweswaran, Pitt
Shawn Murphy HRD

OMOP
Kristin Kostka, IQVIA
Karthik Natarajan, Columbia
Clare Blacketer JNJ

PCORI
Kellie Walters, UNC
Robert Bradford, UNC
Marshall Clark, UNC
Adam Lee, UNC
Evan Colmenares, UNC

TriNetX
Matvey Palchuk
Lora Lingrey

Teams
Analytics
Warren Kibbe, Duke
Heidi Sprait, UTMB
Tell Bennett, U of CO
Andrew Williams, Tufts
Joel Saltz, SBU
Janos Hajagos, SBU
Richard Moffitt, SBU
Tahsin Kurc, SBU

Palantir
Nabeel Qureshi
Andrew Girvin
Amin Manna

Synthetic Data
Regenstrief
Peter Embi

MDClone
Daniel Blumenthal
Hovav Dror
Luz Erez
Josh Rubel

Microsoft
Allison T Rodriguez
Kenji Takeda

NIH National Center for Advancing Translational Sciences

National
COVID
Cohort
Collaborative

Thank you!

National
COVID
Cohort
Collaborative

N3C 2.0: Key Focus Areas

Patient-focused

- Descriptive
  - Epidemiology (in non-hospitalized and hospitalized people)
  - Disparities (racial, ethnic, SES) – identification of risk; spread through communities
  - Disease course of hospitalized disease (subgroups)
  - Drugs – what tried, multiple drugs, association with outcomes
- Pathophysiology (from routinely collected data)
  - Causes of disease (lung injury, hypoxia, cytokine storm, thrombosis, cardiac, renal, etc), and subgroups
  - Which patients with Negative COVID test have COVID19 disease (false negative)?
- Predictors (supervised AI)
  - Predictors of hospitalization, prolonged hospitalization, mortality
  - Scoring systems for intervention (ventilation, dialysis)
  - How does imaging influence subgroups and predictions
- Special populations (subgroups; Latent class analysis; unsupervised AI)
  - Do poorly, different pathophys, respond differently to treatments, etc.
- Long term sequala (Post COVI19 syndromes: weakness, lung, brain, heart, kidney)

System-focused

- Hospital responses to COVID
- Effect of COVID on hospitals
- Economics

Patient Portal: Future studies, Track Recovery

Patient autonomy

- Opt in for future data synch (to show to other care givers)
- Opt in to get information about related clinical trials
- Once enrolled in a study, can Opt in to synch information for research studies
- Opt in to share information back

Track recovery

- Overall: how do you feel?
- Degree of return to usual activities (Physical, Mental)
- Degree of recovery to pre-baseline state of health
  - Subscales (strength, lung, ADL)
- Major symptoms
  - Smell, Breathing (SONG COVID scale); Cough
  - Pain (where), Thinking, Weakness,

Green button: Synergize Care and Research

Taken from SONG COVID outcomes consortium measures
COVID-19sympoms app (http://www.monganinstitute.org/cope-consortium)