NCATS

COLLABORATE. INNOVATE. ACCELERATE.

National COVID Cohort Collaborative (N3C)

5/26/2020

This pandemic highlights urgent needs

- ML algorithms (diagnosis, triage, predictive, etc.)
- Best practices for resource allocation
- Drug discovery
- Reduced disease severity
- Coordinate our efforts to maximize efficiency

All these things require the creation of a comprehensive clinical data set

Introduction

- **Rapid**, collection of clinical, laboratory, and diagnostic data from hospitals and healthcare plans, at the peak of the pandemic, and as the pandemic evolves to understand COVID-19

- **Critical design elements:**
  - **Speed is critical.** Need to collect data now, before the pandemic abates
  - **Make access to the data fast and easy**, and do not prescribe the analysis
  - As data models are developed, test/validate with ongoing data collection
  - **Evolve** to **support clinical trials**

Introducing the National COVID Cohort Collaborative (N3C)

- A **centralized**, secure portal for hosting patient-level COVID-19 clinical data and deploying and evaluating methods and tools for clinicians, researchers, and healthcare

- A **partnership** among CTSA program institutions, distributed clinical data networks (e.g. PCORnet, OHDSI, ACT/i2b2, and TriNetX), and many other clinical partners and collaborators

National
COVID
Cohort
Collaborative

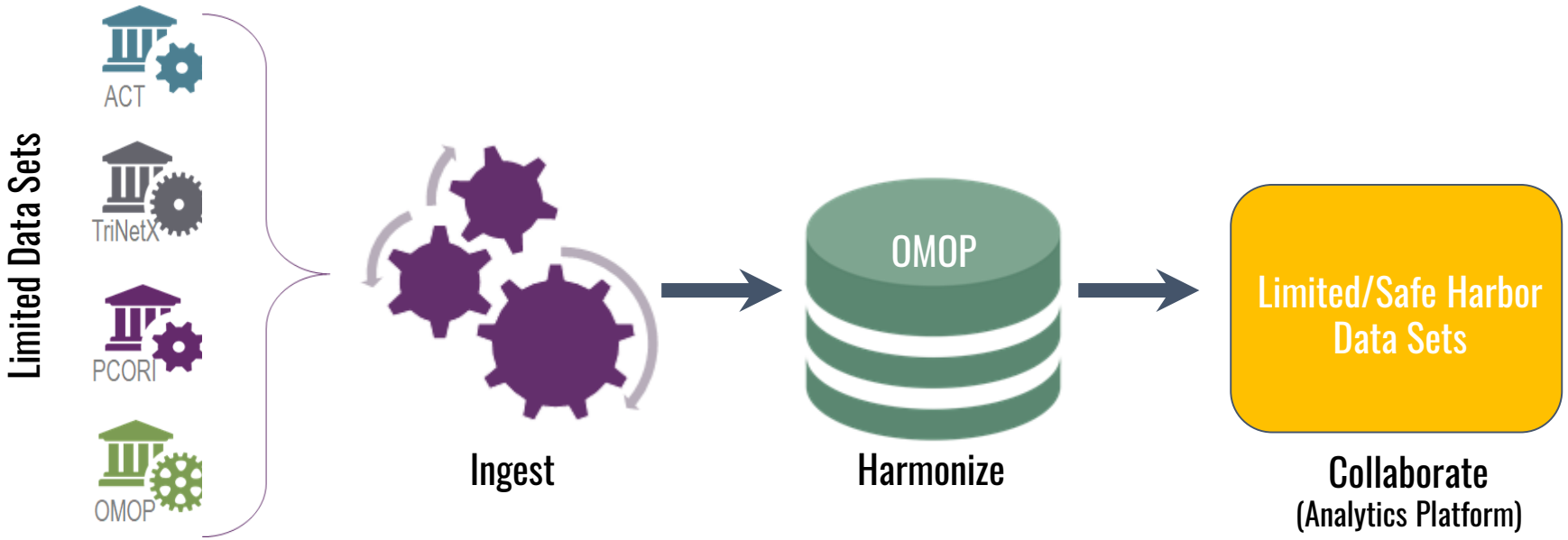
It is being (rapidly) organized:

Five community workstreams:

- Data Partnership & Governance
- Phenotype & Data Acquisition
- Data Ingestion & Harmonization
- Collaborative Analytics
- Synthetic Data

National COVID Cohort Collaborative

N3C Overview



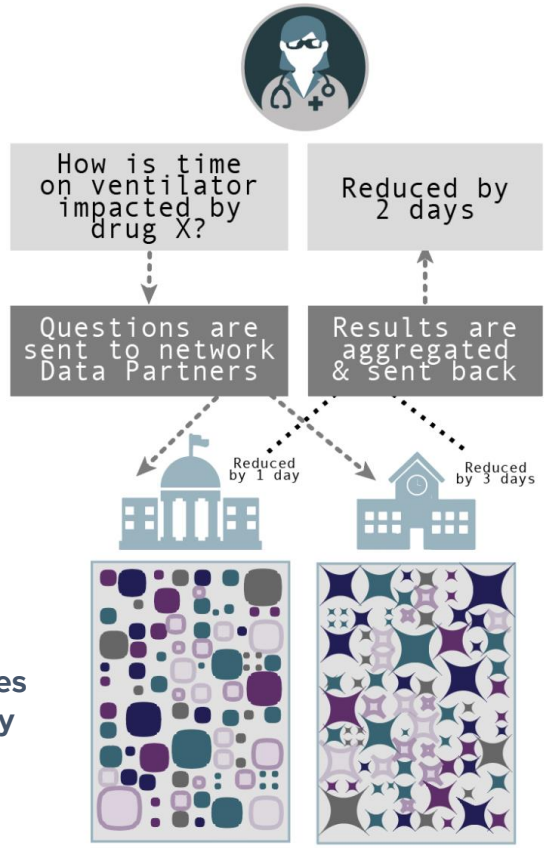
National COVID Cohort Collaborative

"But, am I not already sending data?"

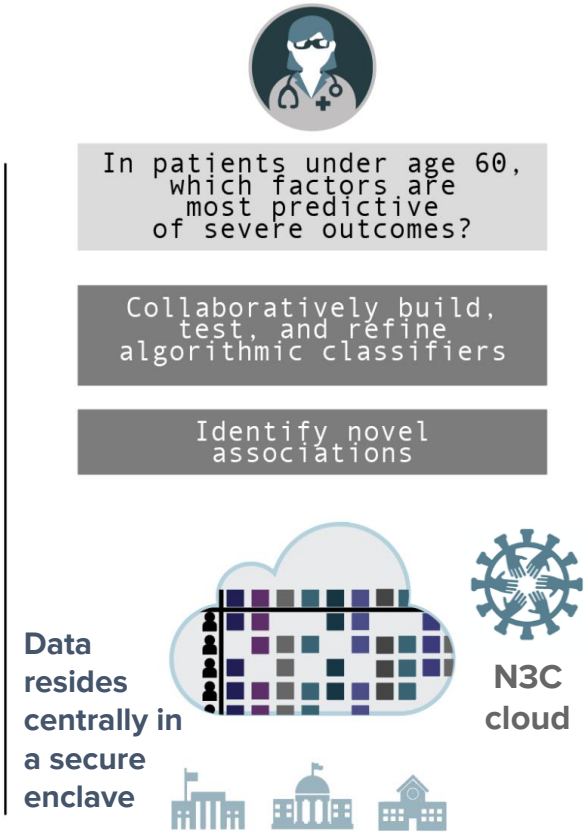
N3C is synergistic with distributed data networks!

Centralizing patient-level data makes it possible to ask qualitatively different and more powerful questions, but is only possible due to each institution having their data in a common data model.

Federated querying



Centralized analytics



National COVID Cohort Collaborative

N3C Progress

AMIA Kickoff
APRIL 13

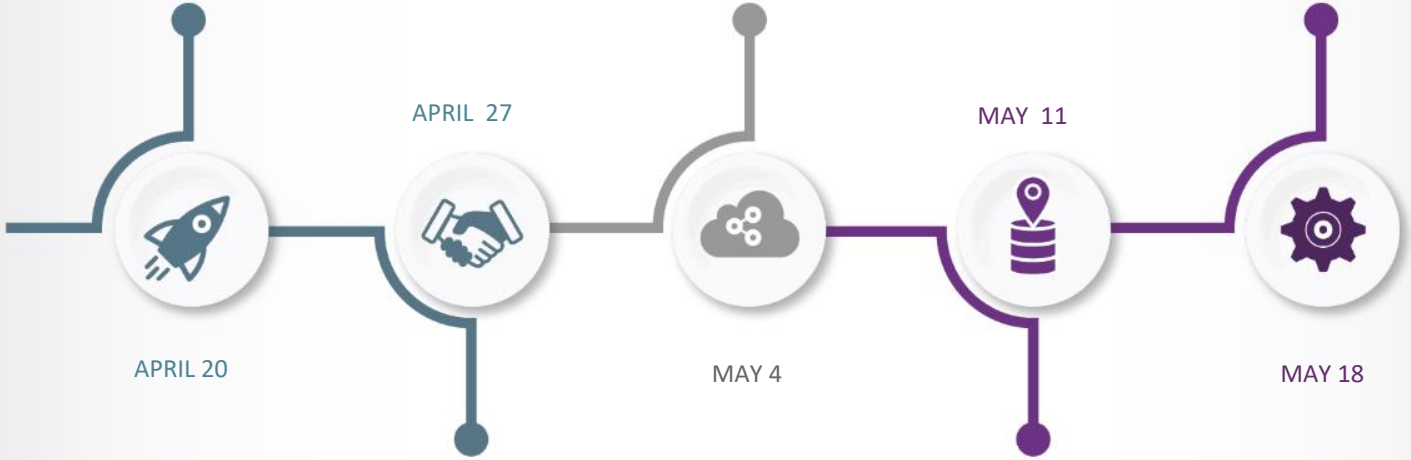
APRIL 20
Workstreams launched
JHU IRB established

APRIL 27
DTA finalized by NIH
Phenotype v1.0 published

MAY 4
N3C platform provisioned in Cloud
1st DTA signed
Platform training initiated

MAY 11
NCATS C.Austin community mtg
Phenotype & extract scripts published
Data Harmonization maps created

MAY 18
500+ members
71 requested DTAs; **17** signed
4 sites submitted data
49 people trained
1st ML models: intubation & AKI
Data Harmonization pipeline built



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
Clair 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

National COVID Cohort Collaborative

N3C Community Workstreams

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

Data Partnerships and Governance

Workstream GOAL

- Develop partnerships with organizations and their IRBs.
- Execute a common data use agreement for contributing to and accessing the COVID-19 dataset.
- Establish a Data Access Committee for reviewing access requests.

John Wilbanks,
Sage Bionetworks

National COVID Cohort Collaborative

Data Partnership and Governance

Data Transfer Agreement

- Facilitates the transfer of Limited Data Set into the NCATS cloud
- 71 have been sent, 14 have been signed, 4 data sets transferred, 2 data sets ingested

Data Use Agreement: Goal 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
- Contributor Attribution
- No download of data

Data Access Committee: [in formation]

Central IRB

National COVID Cohort Collaborative

Phenotype and Data Acquisition

Emily Pfaff, UNC

Workstream GOAL

- Establish a common COVID-19 phenotype that will define the data pull for the limited access dataset
- Create a “white glove” service to obtain data from each site by building easily adaptable scripts for each clinical data model
- Ingest data into a secure location as per approved institutional agreement

National COVID Cohort Collaborative

Phenotype and Data 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

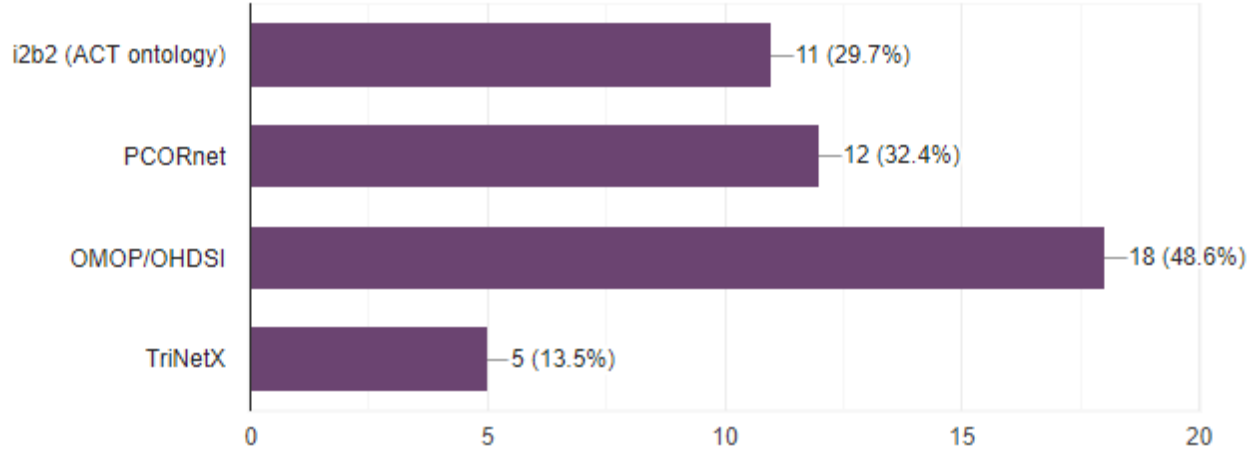
NIH National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

Why Choose OMOP for N3C?

If your institution was to submit a set of covid-specific patient data to a central repository using one of these data models, which data model would be optimal? (Check multiple only if there is a "tie" for first place.)

37 responses



Note: Respondents may support more than one common data model in their environment.

NIH National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

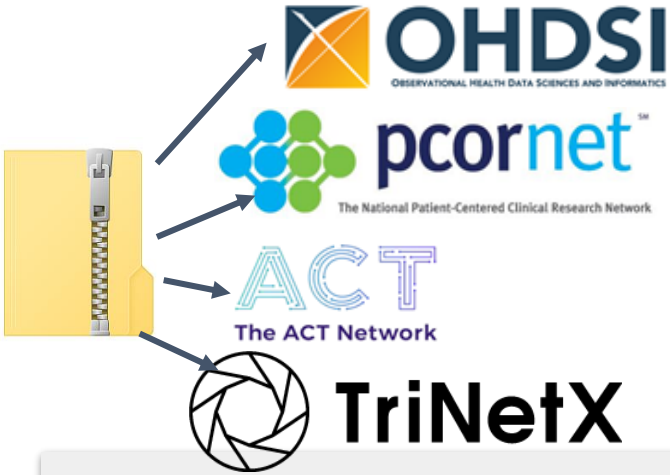
Data Ingestion and Harmonization

Workstream GOAL

- Ingest limited data sets that are available in their native data formats such as PCORnet, ACT and OMOP and harmonize them into common data model based on OMOP standard

Christopher Chute,
Johns Hopkins University

Common Data Model Harmonization



ADEPTIA Workflow

	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%**

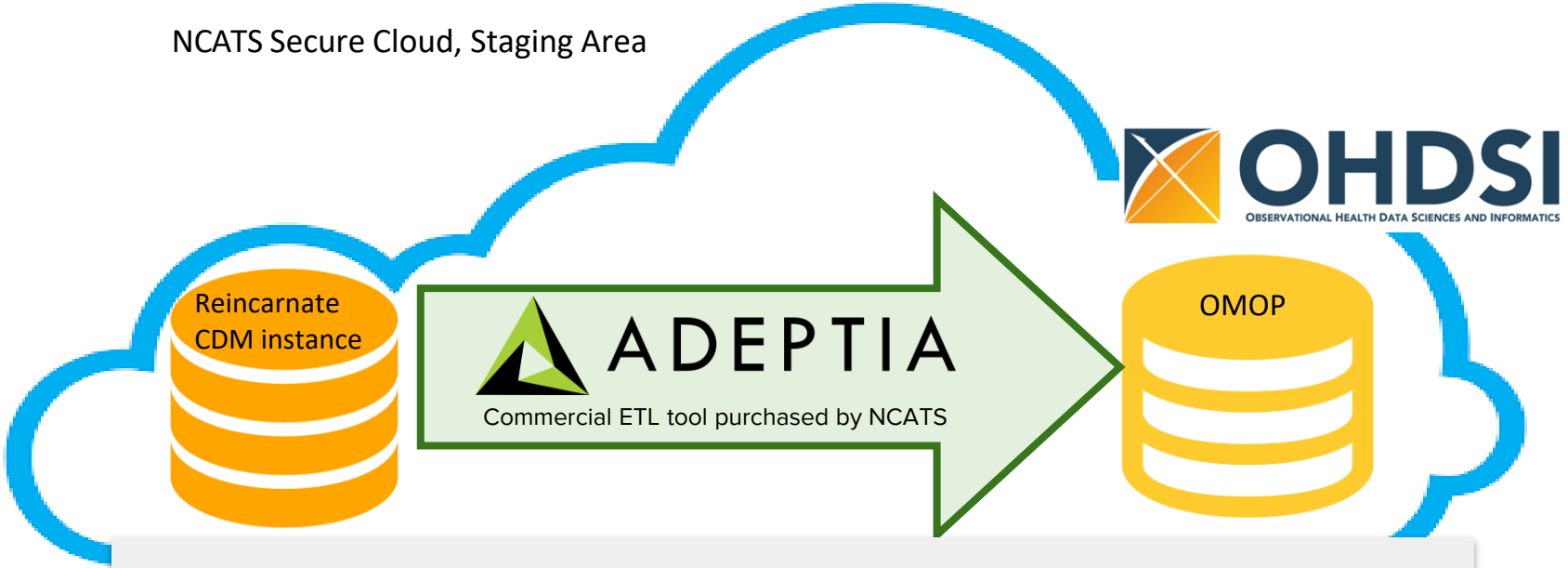
Data Quality Dashboard (shared with site)

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 Harmonization: Transformation



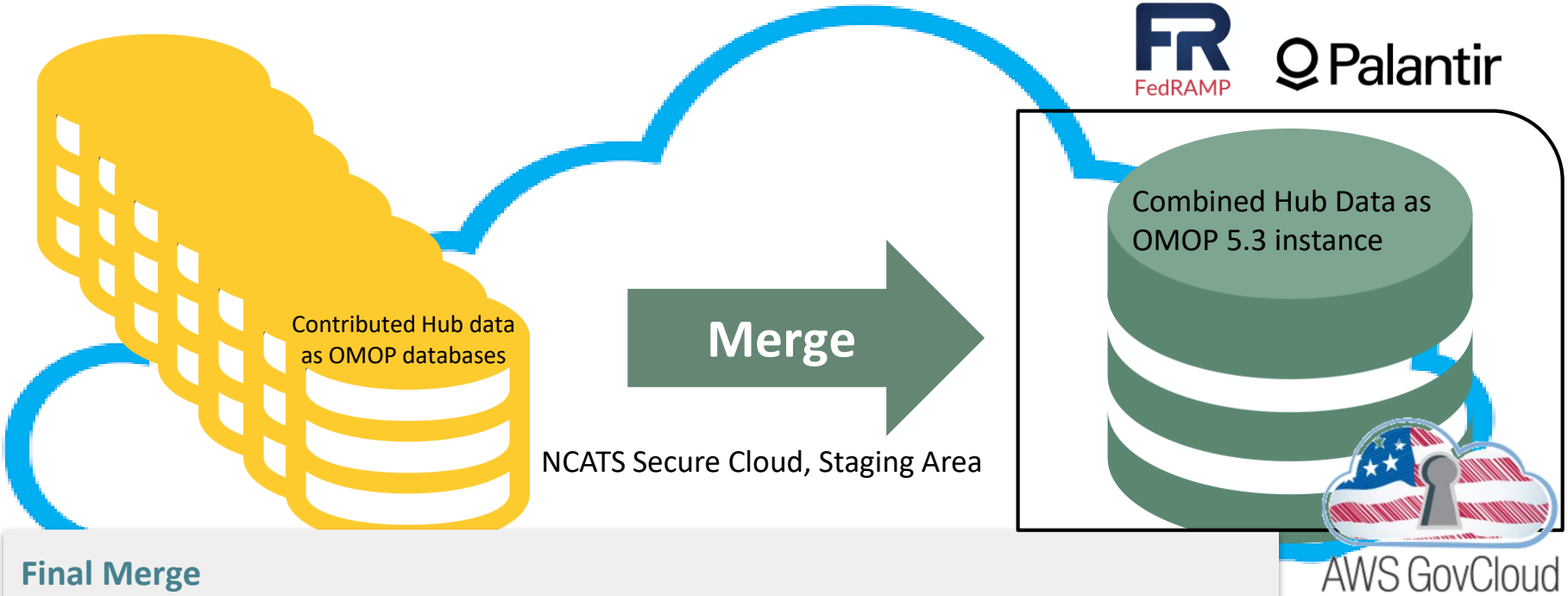
Second Stage Ingestion

- Repair or encode aberrant data (COVID LOINC codes) ✔
- Transform source CDM into OMOP 5.3 ✔
- Leverage library of validated CDM to OMOP maps

National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

Data Harmonization: Secure Integration



Final Merge

- OMOP versioned data from all sources combined into analytic database
- Analytic database will transfer to Palantir Analytic Platform

National COVID Cohort Collaborative

Collaborative Analytics

Justin Guinney, PhD

Sage Bionetworks

Joel Saltz, MD, PhD

Stony Brook

Workstream GOAL

- Work collaboratively to generate insights related to COVID-19 from the harmonized limited access dataset
- Experts in AI, ML, and other technologies will assist in reviewing and iterating on portal architecture to ensure fit-for-purpose implementation
- Design UX and apps for diverse analytical users (researchers, informaticians, clinicians)

National COVID Cohort Collaborative

Collaborative Analytics

To join a workstream or sub-group, click here

PM: rafael.fuentes@nih.gov

Portals & Dashboards

Dave Eichmann, U of Iowa
Warren Kibe, Duke University
Meeting time: Tues, 10am ET

Tools & Resources Resources

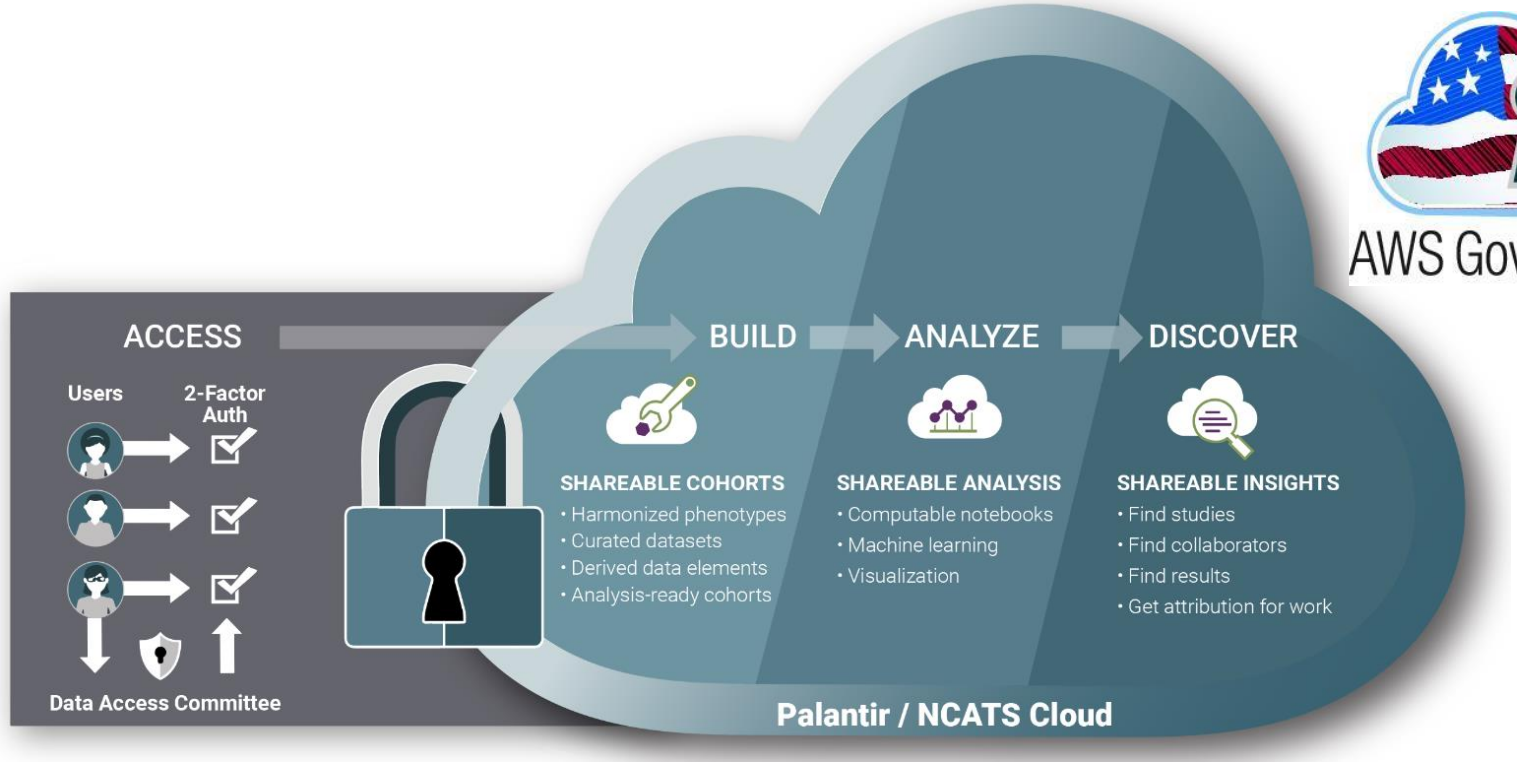
Chunlei Wu, Scripps
Andrew Williams, Tufts
Meeting time: Fri, 4pm ET

Clinical Scenarios & Data Analytics

Peter Robinson, JAX
Heidi Spratt, UofTexas;
Tell Bennett, UofColorado
Meeting time: Monday, 11am ET

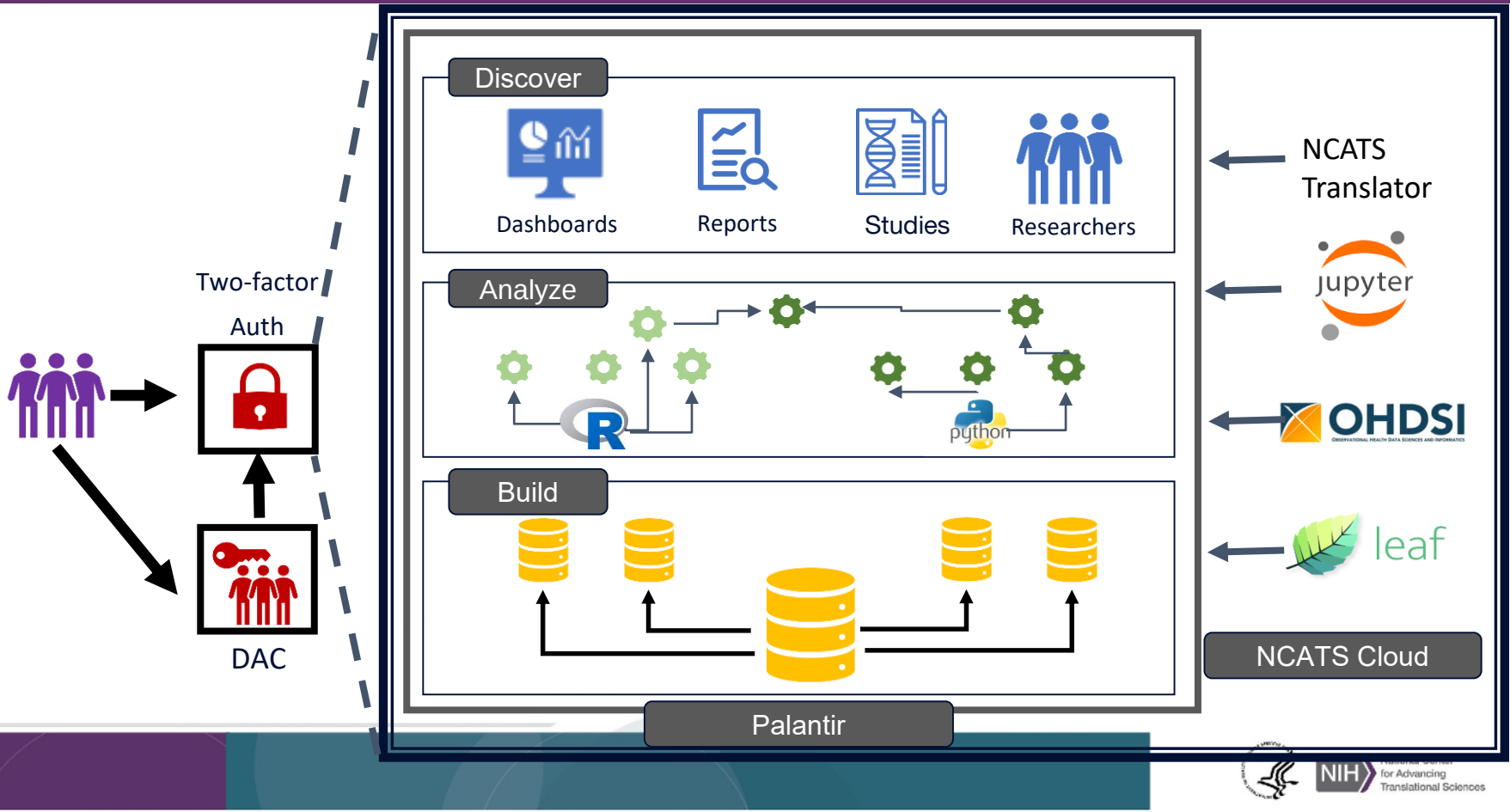
National COVID Cohort Collaborative

Collaborative Analytics - N3C Secure Data Enclave



National COVID Cohort Collaborative

Collaborative Analytics - N3C Secure Data Enclave



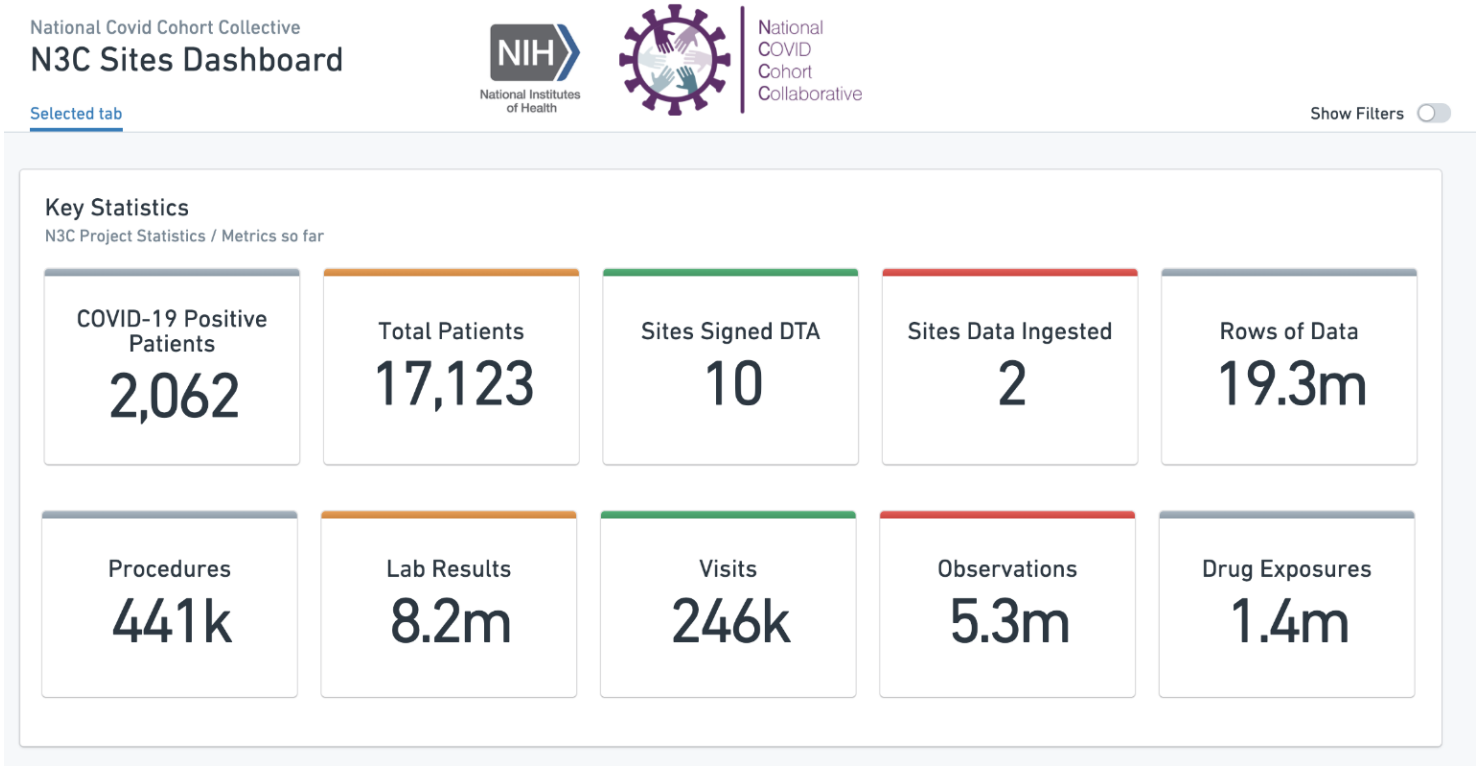
Clinical Phenotype & Predictive Analytics

Recently Achieved Milestone: Predictive Analytics Demo on N3C Platform

- Demonstration of N3C/Palantir Foundry Analytic Platform using **real world data**
- Limited Data Set from Wash U. ingested into the N3C platform
- N3C Platform is hosted in AWS GovCloud and is FedRAMP Moderate certified
- N3C Platform preserves attribution, reproducibility and provenance
- Clinical early warning/clinical decision support
- Machine learning demo: Random Forest Model trained on WashU data predicting
    - Invasive ventilation (intubation) - one day heads up
    - AKI 5 days out

National COVID Cohort Collaborative

N3C Analytics Platform



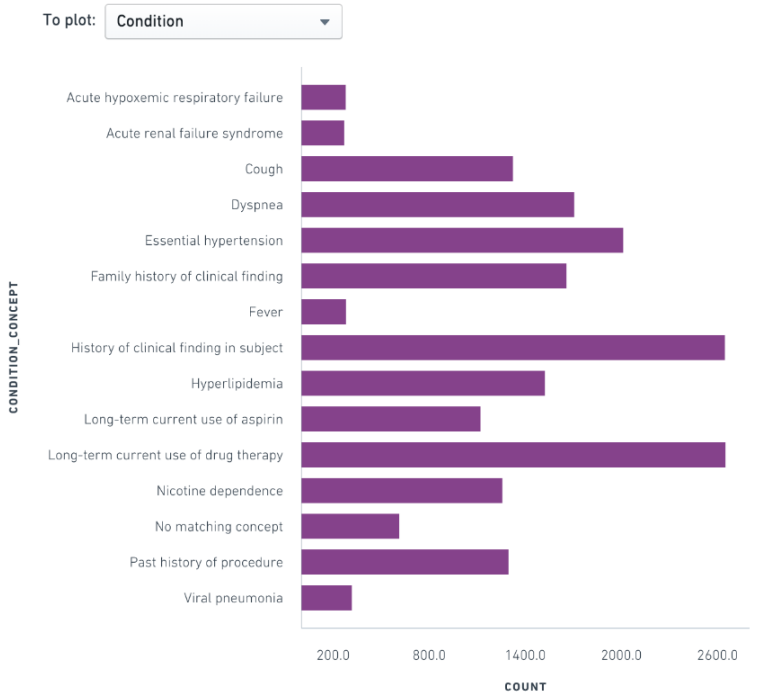
National COVID Cohort Collaborative

Cohort Characterisation

Cohort characteristics

Summary statistics for WUSTL patients

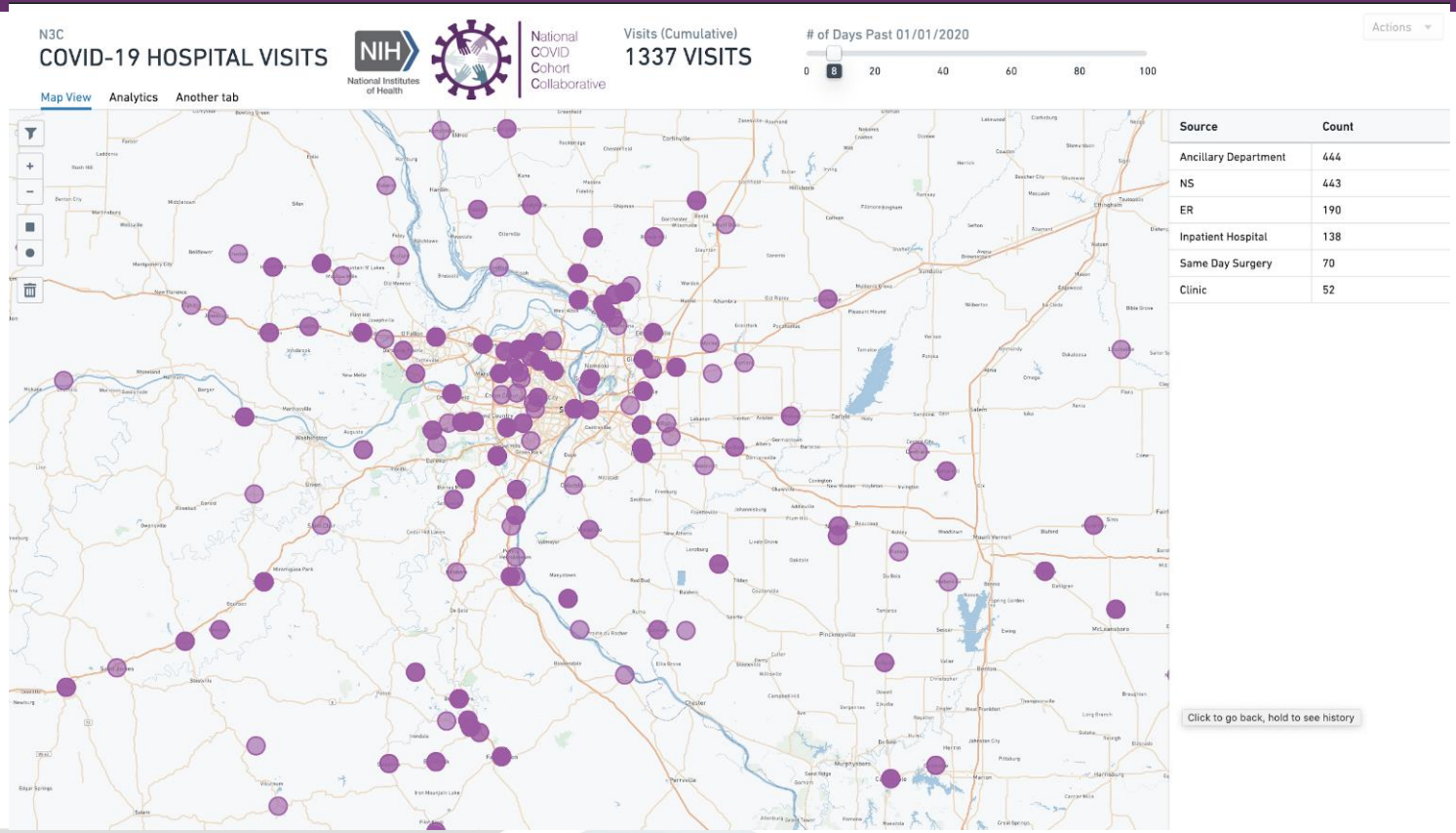
	COVID (N=1161)	Non-COVID (N=5904)	Overall (N=7065)
Gender			
Male	1059	7022	2141
Female	1091	8069	9160
Null		3	3
Age			
0 - 17	46	2095	2141
18 - 29	303	2043	2346
30 - 49	616	3638	4254
50 - 64	584	3488	4072
65+	523	3498	4021
Race			
White	614	8110	8724
Asian	127	1225	1352
American Indian or Alaska Native	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



NIH National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

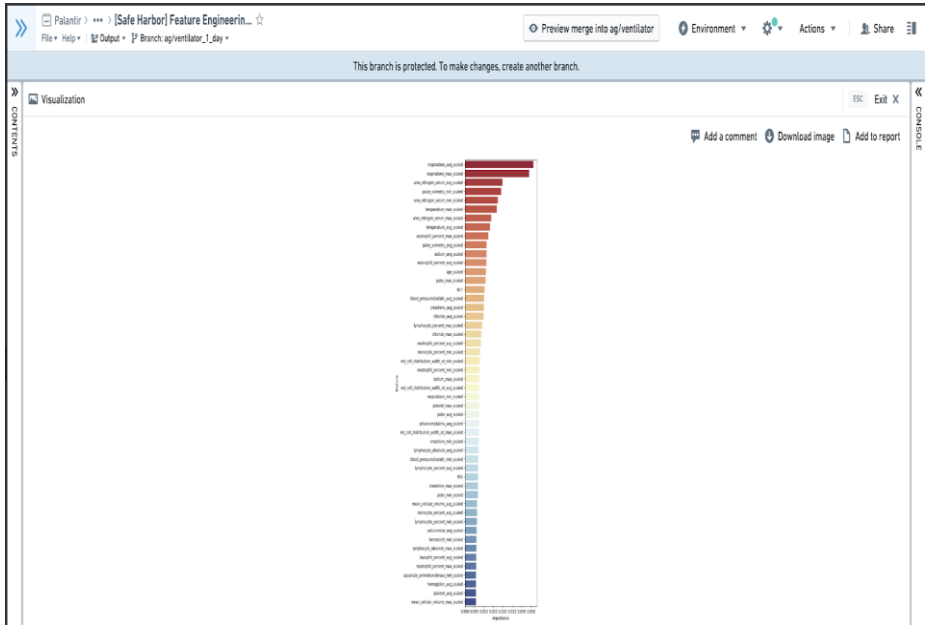
Time/Space Vector - Live Example



National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

Predictive Modeling: Risk of Ventilation and AKI



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.

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 COVID Cohort Collaborative

Synthetic Data

Philip Payne, PhD

Washington University

Atul Butte, MD, PhD

UCSF

Workstream GOAL

- Pilot the generation of synthetic EHR data from the N3C cohort for broad data sharing and community analytics

National COVID Cohort Collaborative

Synthetic Data: Objectives

1) Create a ”pipeline” that can be used to generate computationally derived synthetic clinical data
2) Demonstrate the pipeline by populating it with data from a group of 3-5 pilot sites
3) Provision access to resulting synthetic data sets for evaluation and use by N3C participants and broader research and innovation community
    - **Conduct targeted verification and validation studies informed by “real world” use cases (comparing results of analyses between source and synthetic data)**
4) Plan for future expansion of collaborative, commensurate with overall growth of N3C and community needs

Microsoft Azure

Microsoft Research

MDCLONE

National COVID Cohort Collaborative

N3C Community Workstreams

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

Question and Answers

National COVID Cohort Collaborative

Thank You