NCATS

COLLABORATE. INNOVATE. ACCELERATE.

National COVID Cohort Collaborative (N3C)

8/2020

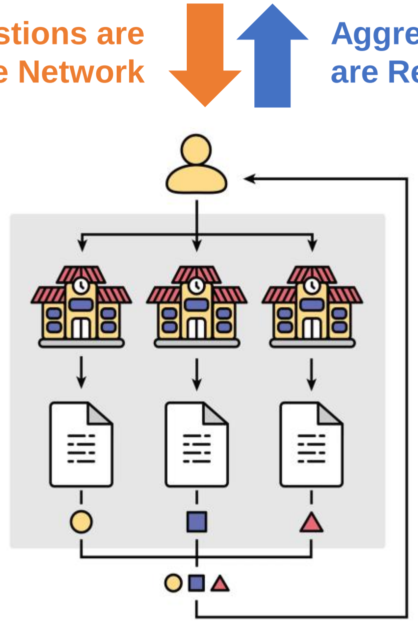
National
COVID
Cohort
Collaborative

National COVID Cohort Collaborative (N3C)

Federated Query

Questions are Sent to the Network

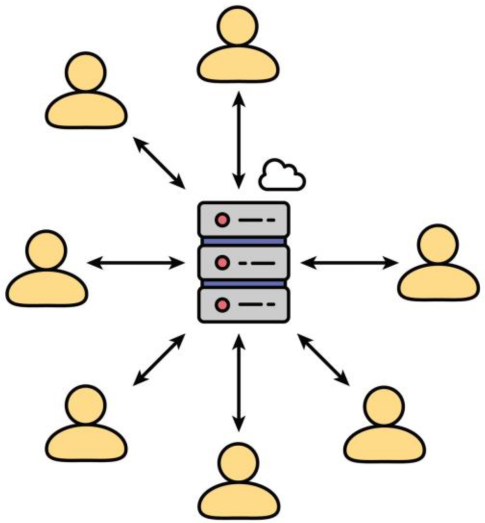
Aggregated Results are Returned



Is **drug X** beneficial to COVID-19 patients?
Does **disease Y** impair course?

Centralized Analytics

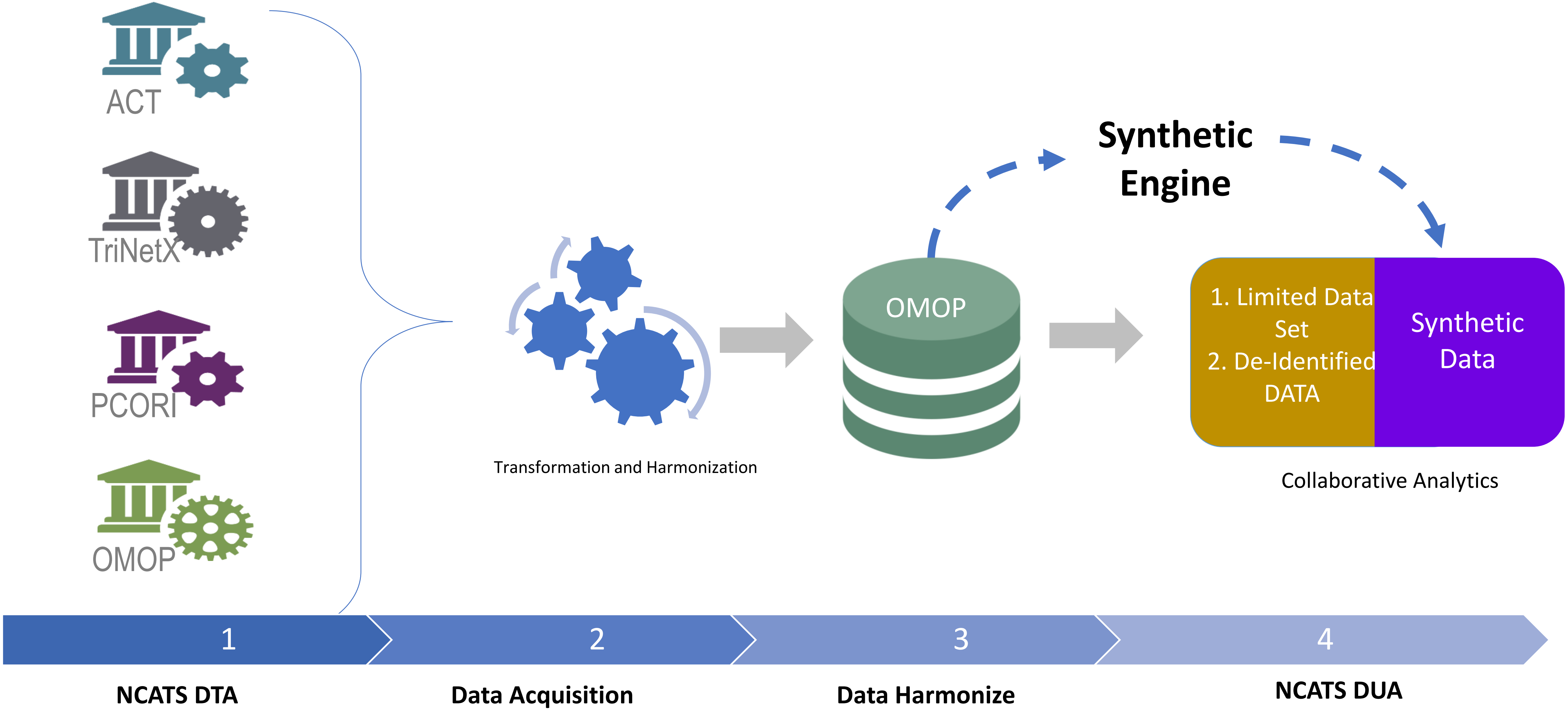
Data Resides Centrally in a Secure enclave



What **drugs** help or hinder COVID patients?
What **factors** predict being placed on a ventilator?
What comorbid **diagnosis** are risk factors

National
COVID
Cohort
Collaborative

N3C Data Pipeline



ACT

TriNetX

PCORI

OMOP

Transformation and Harmonization

OMOP

Synthetic Engine

1. Limited Data Set
2. De-Identified DATA

Synthetic Data

Collaborative Analytics

1 **NCATS DTA**

2 **Data Acquisition**

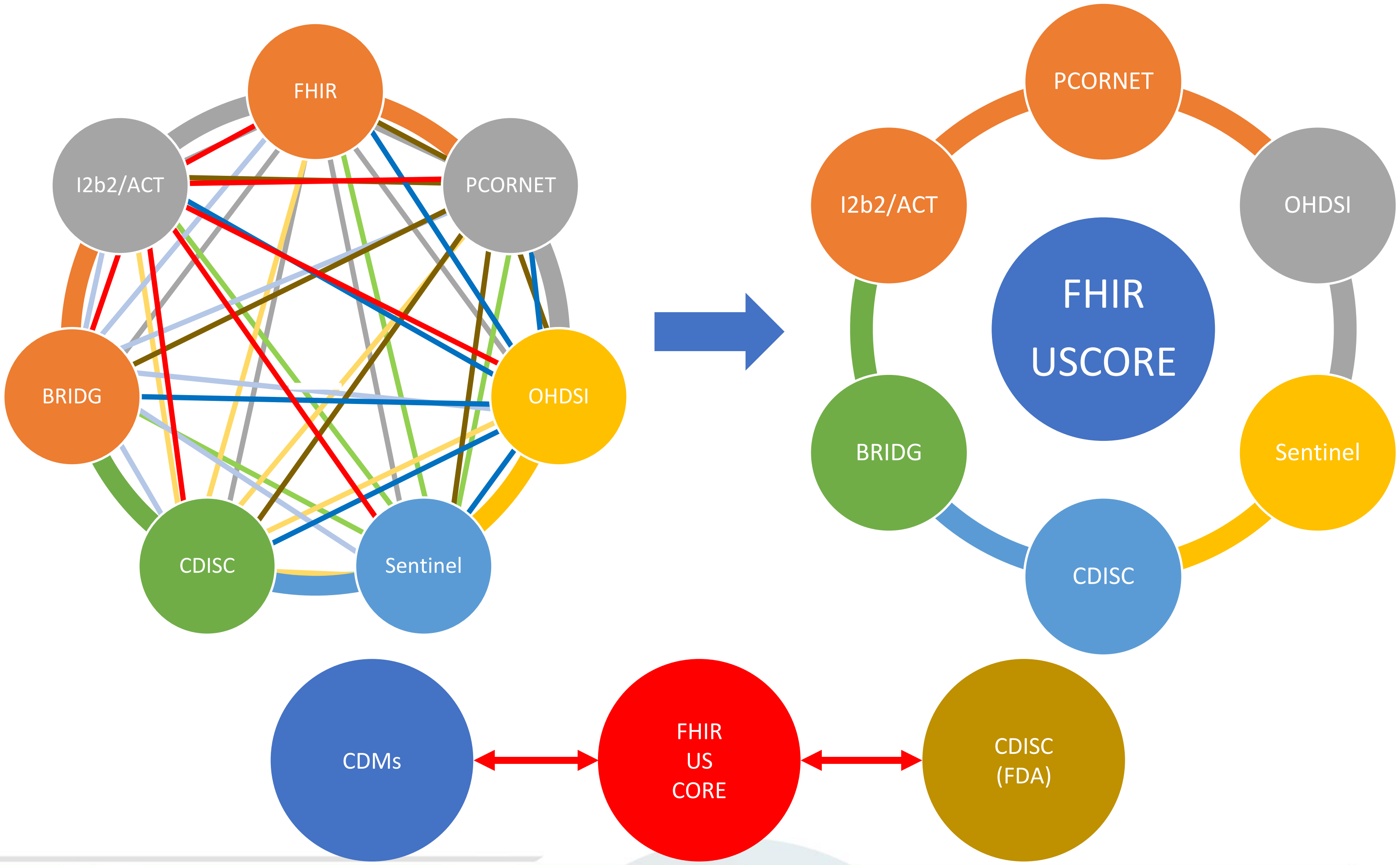
3 **Data Harmonize**

4 **NCATS DUA**

National COVID Cohort Collaborative

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**



National Center for Advancing Translational Sciences

Data Partnership and Governance

Goal of the Data Use Agreement 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**
- **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

Data Tiers

Access Level	Level 1 - Synthetic	Level 2 –Deidentified		Level 3 - LDS
Data Type	Synthetic Data	**Aggregate Data** (i.e., summary statistics)	**HIPAA Safe Harbor**	**HIPAA Limited Data Set**
Description	Computational data derivative statistically resembles original data	Counts and summary statistics representing 10 or more individuals	Data stripped of 18 direct identifiers called out in the HIPAA Privacy Rule	Data that is stripped of all PHI under HIPAA except dates and zip code

Data Sharing Initiative: Synthetic Data

*Computer Derived Synthetic Data: Validation of Sepsis Prediction

Public / Private Partnership

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

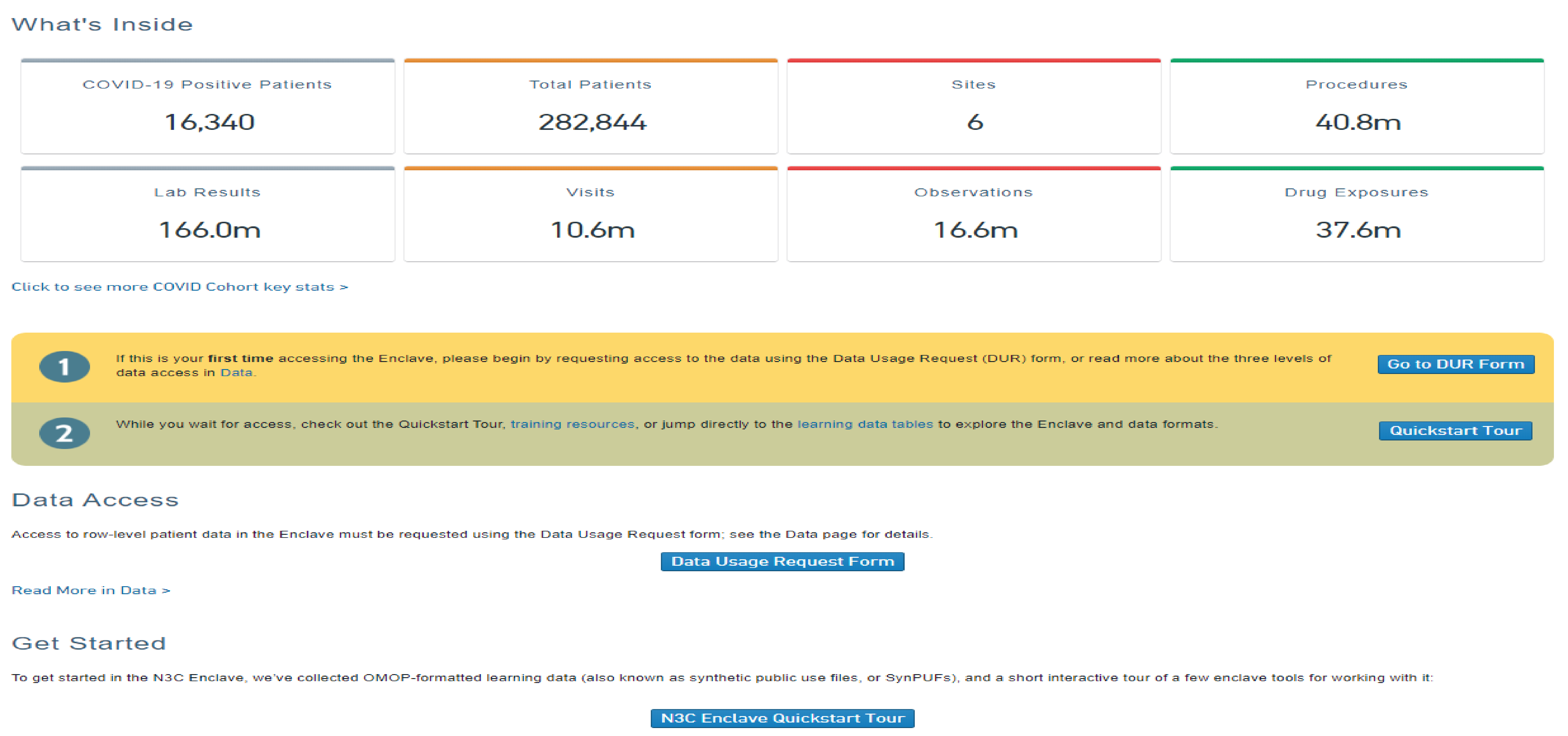
		Trained on real data Tested on real data	Trained on synthetic data Tested on real data
Train	Accuracy	0.925	0.911
	Precision	0.95	0.925
	Recall	0.817	0.799
	F-Score	0.879	0.858
10-fold cross-validation	Accuracy	0.839	0.816
	Precision	0.802	0.754
	Recall	0.704	0.666
	F-Score	0.745	0.704
Test	Accuracy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

N3C Statistics



What's Inside

COVID-19 Positive Patients	Total Patients	Sites	Procedures
16,340	282,844	6	40.8m
Lab Results	**Visits**	**Observations**	**Drug Exposures**
166.0m	10.6m	16.6m	37.6m

Click to see more COVID Cohort key stats >

1 If this is your **first time** accessing the Enclave, please begin by requesting access to the data using the Data Usage Request (DUR) form, or read more about the three levels of data access in Data.

2 While you wait for access, check out the Quickstart Tour, training resources, or jump directly to the learning data tables to explore the Enclave and data formats.

Data Access

Access to row-level patient data in the Enclave must be requested using the Data Usage Request form; see the Data page for details.

Read More in Data >

Get Started

To get started in the N3C Enclave, we've collected OMOP-formatted learning data (also known as synthetic public use files, or SynPUFs), and a short interactive tour of a few enclave tools for working with it:

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

National
COVID
Cohort
Collaborative

Thank you!