The National COVID Cohort Collaborative: A Social Experiment in Team Science

CADRE Research-in-Progress
July 28, 2020

@data2health @ncats_nih_gov
https://covid.cd2h.org/
https://ncats.nih.gov/n3c
Introducing the National COVID Cohort Collaborative (N3C)

- **A centralized**, secure portal for hosting patient-level COVID 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
Major workstreams of the National COVID Cohort Collaborative

1) Data Partnership & Governance
   - NIH National Center for Advancing Translational Sciences
   - ACT
   - TriNetX
   - PCORNet
   - OMOP
   - Other

2) Phenotype & Data Acquisition

3) Data Ingest & Harmonization

4) Collaborative Analytics
   - NCATS cloud

5) Synthetic Clinical Data
**Key Stats**

<table>
<thead>
<tr>
<th>7/28/2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>53 DTAs executed</td>
</tr>
<tr>
<td>31 IRB protocols approved (26 reliance, 5 local)</td>
</tr>
<tr>
<td>29 Regulatory complete (both DTA and IRB)</td>
</tr>
<tr>
<td>38 Met with Data Acquisition Group</td>
</tr>
<tr>
<td>16 Deposited data:</td>
</tr>
<tr>
<td>5 - PCORI</td>
</tr>
<tr>
<td>3 - ACT</td>
</tr>
<tr>
<td>4 - TriNetX</td>
</tr>
<tr>
<td>4 - OMOP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7/28/2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onboarded Individuals: 782</td>
</tr>
<tr>
<td>Onboarded Unique Institutions: 234</td>
</tr>
<tr>
<td>Onboarded from Clinical Organization Hubs: 64</td>
</tr>
</tbody>
</table>

Distribution of cases from 4 sources
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
Governance & sIRB Overview

Single/Central IRB (sIRB)

- Johns Hopkins serving as central IRB
- Smart IRB makes it easy - all CTSAs are already members, so if you’re willing to rely on sIRB, the paperwork is basically complete
- Not required - if you want to do the work locally, you can do so

Who to contact about reliance or local filing

Tricia Francis pfranci4@jhu.edu
What data is in the N3C?

Community maintained computable phenotype for COVID-19

**DATA FOR 1 YEAR**
- Observations
- Specimens
- Visit
- Procedures
- Drugs
- Devices
- Conditions
- Measurements
- Location
- Provider

**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

**STATS FOR RESULTING COHORT**

<table>
<thead>
<tr>
<th></th>
<th>Sites</th>
<th>COVID+ cases</th>
<th>Deaths</th>
<th>Visits</th>
<th>Clinical measures</th>
<th>Medication records</th>
<th>Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sites</td>
<td>6</td>
<td>21,972</td>
<td>4,559</td>
<td>10.6 mil</td>
<td>166 mil</td>
<td>93.4 mil</td>
<td>282,844</td>
</tr>
</tbody>
</table>

Emily Pfaff
UNC

Community maintained computable phenotype for COVID-19 as of: 7/28/20
<table>
<thead>
<tr>
<th>Access Level</th>
<th>Registered</th>
<th>Controlled</th>
<th>Controlled-Plus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Type</td>
<td>Synthetic Data (pending pilot)</td>
<td>Aggregate Data (i.e., counts)</td>
<td>HIPAA Limited Dataset</td>
</tr>
<tr>
<td>Description</td>
<td>Computational data derivative that statistically resembles the original data</td>
<td>Counts and summary statistics representing 10 or more individuals</td>
<td>Data stripped of 18 direct identifiers per HIPAA rules (dates, full zip code, and any age)</td>
</tr>
<tr>
<td>Downloadable data</td>
<td>Planned: pending validation &amp; organizational agreement</td>
<td>Downloadable query results</td>
<td>No</td>
</tr>
<tr>
<td>Custom software</td>
<td>Yes</td>
<td>Yes - on downloaded query results</td>
<td>Yes with DAC approval</td>
</tr>
</tbody>
</table>

**Increasing regulatory control**
When a new site joins the project, they...
1. Work with the Data Partnership & Governance group to get their regulatory ducks in a row.
2. Attend a Data Ingest onboarding meeting to walk through the process and get questions answered.
3. Assign appropriate technical personnel to review documentation and scripts on our GitHub site.
4. Run scripts locally and transmit data extracts on a regular basis, with the support of the N3C Phenotyping and Data Acquisition team.
First Stage Ingestion

- Unpack Zipped CSV Files. Check data manifests.
- Reconstitute into native CDM formats.
- Hybrid Data Quality checks adapting OHDSI Data Quality Dashboard.

<table>
<thead>
<tr>
<th></th>
<th>Verification</th>
<th>Validation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pass</td>
<td>Fail</td>
<td>Total</td>
</tr>
<tr>
<td>Plausibility</td>
<td>159</td>
<td>21</td>
<td>180</td>
</tr>
<tr>
<td>Conformance</td>
<td>637</td>
<td>34</td>
<td>671</td>
</tr>
<tr>
<td>Completeness</td>
<td>369</td>
<td>17</td>
<td>386</td>
</tr>
<tr>
<td>Total</td>
<td>1165</td>
<td>72</td>
<td>1237</td>
</tr>
</tbody>
</table>
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
NCATS Secure Cloud, Staging Area

Merge

Contributed Hub data as OMOP databases

Combined Hub Data as OMOP 5.3 instance

NCATS Secure Cloud, Staging Area

Final Merge

- OMOP versioned data from all sources will be combined into analytic database
- Analytic database will migrate to Palantir Analytic Platform
Secure, reproducible, transparent, versioned, provenanced, attributed, and shareable analytics on patient-level EHR data
Artifacts are associated with ORCiDs using the Contributor Attribution Model (CAM) cd2h.org/attribution

Provenance graph showing linkages between results, code, and source data allowing for full end-to-end reproducibility
Druggable proteins that interact indirectly with SARS-CoV-2

<table>
<thead>
<tr>
<th>SARS-CoV-2 protein</th>
<th>human protein 1</th>
<th>human protein 2</th>
<th>drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsp8</td>
<td>HLA-A</td>
<td>C5</td>
<td>eculizumab</td>
</tr>
<tr>
<td>S protein</td>
<td>CCNB1</td>
<td>BCL2</td>
<td>ribavirin</td>
</tr>
<tr>
<td>S protein</td>
<td>CCNB1</td>
<td>BCL2</td>
<td>vincristine</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Analyze drugs for positive/negative correlations using NBC cohort

Browse data in N3C enclave

Social Deprivation Index (SDI): composite measure of seven demographic characteristics

Distribution of SDI scores; we selected 70 for high vs. low

Categorize counties as urban or rural based on density.

<70 SDI substantially higher incidence/death rate from public data
## COVID-19 Collaborative Analytical Task Teams

<table>
<thead>
<tr>
<th>Clinical topic</th>
<th>Analytical questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKI/ARB/ACE</td>
<td>How to predict which patients will develop AKI?  How to predict when AKI will progress to CKD?  How do outcomes correlate with dialysis timing?  Oxygenation?  ACEI vs. ARBs vs. ARNI differentiation?</td>
</tr>
<tr>
<td>Critical Care</td>
<td>How to best prioritize limited resources?  What predictors help define which patients will fare best with any given intervention?</td>
</tr>
<tr>
<td>Diabetes</td>
<td>What is the association between HbA1c at baseline and COVID outcomes for patients with diabetes?  Are outcomes equivalent among patients with type 2 diabetes and COVID-19 using different anti-hyperglycemic medications?  Relationship between COVID correlated diabetes development/exacerbation and outcome and treatment response.</td>
</tr>
<tr>
<td>Imaging</td>
<td>Integrative analysis of image and clinical data to predict outcome and treatment response.</td>
</tr>
<tr>
<td>Immuno-suppressed/compromised</td>
<td>How effective is convalescent plasma?  What are the predictors of effectiveness?</td>
</tr>
<tr>
<td>Oncology</td>
<td>What germ line mutations predispose cancer patients to severe COVID outcomes?</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>What endophenotypes exist for MIS-C patients?  What are the consequences of childhood COVID infection?  Can we build a classifier to predict MIS-C?</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>Determine birth outcomes across COVID-19 severity, intervention, and vaginal versus c-section deliveries; postpartum morbidity and complications in positive cases.</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Is there a racial disparity to access in testing?  What is the transmission intensity among populations by race/ethnicity, rural/urban, income, etc?  Are there differences in therapy response?</td>
</tr>
<tr>
<td>Short/long term Complications</td>
<td>Assess longer term conditions, complications, and health care utilization; do these patients have readmissions?  What are their outcomes?</td>
</tr>
</tbody>
</table>
What does this mean for Hopkins investigators?

- N3C Analytics open to any Hopkins community member
- This is the ground floor
- Social experiment is to foster collaboration (180 authors on methods)
- Clinical questions are being addressed by self-organizing group
- Hopkins faculty and student can and should assume leadership roles
- No cost for data access, analytic suite, participation
- But then, no direct funding for analytics (though one could apply)
- This will happen without us if we do not “show up”
Joining the N3C Community Workstreams

1) Data Partnership & Governance

2) Phenotype & Data Acquisition

3) Data Ingest & Harmonization

4) Collaborative Analytics

5) Synthetic Clinical Data

N3C “soft launch” TODAY!

ENGAGE:
Onboarding to N3C

cd2h.org/onboard

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

Get data access:

- Institutions execute their DUAs (these are going out now)
Thank you!