Building a Nationwide COVID-19 Cohort Through Informatics: A new initiative being coordinated by CD2H & NCATS
Agenda

• Brief introduction to AMIA’s Webinar Series and the role of CD2H/CTSA NCATS

• Introduction to the new NCATS COVID-19 Cohort Collaborative N3C
  • Melissa Haendel, PhD, Director, Center for Data to Health (CD2H), Oregon Health & Science University
  • Christopher Chute, MD, DrPH; CD2H Co-Program Director, Bloomberg Distinguished Professor, Chief Research Information Officer, Deputy Director, Hopkins CTSA, Johns Hopkins University
  • Mitra Rocca, Dipl. Inform. Med., FAMIA; Senior Medical informatician, Office of Translational Sciences; Center for Drug Evaluation & Research, Food & Drug Administration
  • Ken Gersing, MD, Director of Informatics NCATS DCI, NCATS/NIH

• Audience Q&A
Health Informatics is the science of how to use data, information, and knowledge to improve human health, including the execution of scientific research, the delivery of health care services, and the promotion of public health. AMIA is the multi-disciplinary, inter-professional home for 5,400+ health informatics experts.
Working Groups of AMIA

- Biomedical Imaging Informatics
- Clinical Decision Support
- Clinical Information Systems
- Clinical Research Informatics
- Consumer and Pervasive Health Informatics
- Dental Informatics
- Education
- Evaluation
- Bioinformatics
- Ethical, Legal and Social Issues
- Genomics and Translational Global Health Informatics
- People and Organizational Issues
- Intensive Care Informatics
- Knowledge Discovery and Data Mining
- Knowledge Representation and Semantics
- Nursing Informatics
- Open Source Student
- Pharmacoinformatics
- Primary Care Informatics
- Public Health Informatics
- Regional Informatics Action
- Visual Analytics
- Natural Language Processing
The Globe of Health Informatics & COVID-19

- Analysis of Coronavirus
- Development of Therapeutics and symptom identification
- Treatment of patients via EHRs & Information Exchange
- Tools for contact tracing and for study of transmission

- DNA
- Small Molecules
- Disease
- Patient
- Practice
- Population
- Global

- 10^-9
- 10^-6
- 10^-3
- 10^0
- 10^3
- 10^6
- 10^9

- TBI
- CRI

- Clinical
- Consumer Health

- Public Health
To highlight how our members and the broader informatics community is addressing this global pandemic we are launching the AMIA COVID-19 Webinar Series.

We will look at the pandemic through a health informatics lens and is designed to share informatics responses to the COVID-19 pandemic. Panelists will share their specific domain expertise, including clinical informatics, public health informatics, translational bioinformatics, clinical research informatics, and consumer health informatics. We will also have special emphasis webinars covering topics related to global health, telemedicine, and public policy during the COVID-19 pandemic. These webinars are open to all at no cost.
Several additional webinars are being planned to highlight members of AMIA and the wider informatics community

- Nursing Informatics highlighted 4/14 @ 12pm ET
- Visit AMIA.org/COVID19
AMIA COVID-19 Webinar series
Building a Nationwide COVID-19 Cohort Through Informatics: A New Initiative being coordinated by CD2H & NCATS
April 13, 2020

These slides: bit.ly/n3c-amia
@data2health
@ncats.nih.gov
https://covid.cd2h.org/
Panelists

Melissa Haendel, PhD
Director, Center for Data to Health (CD2H)
Oregon Health & Science University

Christopher Chute, MD, DrPH
CD2H Co-Program Director
Bloomberg Distinguished Professor
Chief Research Information Officer
Deputy Director, Hopkins CTSA
Johns Hopkins University

Ken Gersing, MD
Director of Informatics
NCATS DCI
NCATS/NIH

Mitra Rocca, Dipl.-Inform. Med. FAMIA
Senior Medical Informatician
Office of Translational Sciences Center for Drug Evaluation & Research, FDA
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
Introducing the National COVID Cohort Collaborative (N3C)

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

- A **partnership** among several HHS agencies, the CTSA network, distributed clinical data networks (e.g. PCORnet, OHDSI, ACT/i2b2, and TriNetX), and other clinical partners

It is being (rapidly) organized:

**Four community workstreams:**
- Data Partnership & Governance
- Phenotype & Data Acquisition
- Data Ingestion & Harmonization
- Collaborative Analytics
Distributed clinical data network advantages

- Maximizes #records
- Flexibility in diversity of querying
- More complete, longitudinal data

Federated Data Model
The results are aggregated

Data resides locally

Questions are sent to network Data Partners

Aggregate answers are sent back
Centralized, harmonized COVID-19 dataset advantages

Centralized model advantages
- Large dataset
- Consistency
- Improved ML applications & analytics over patient-level data
- Shared compute infrastructure and application deployment
- Purpose-driven curation/data modeling for covid-19
Data Partnership & Governance Workstream

Clinical institutional partners

Central IRB
DUA

Data Ingest

Harmonized covid data

Approve access

Data Access Committee:
Stakeholder representation

Member of

Request access

Open covid data

Synthetic derivation

Register & access

Everyone

Qualified researchers, clinicians & data contributors
Since the data could be identifiable to the patient and institution, these analyses are only for:

- Analysis of COVID (community spread, risk, treatment)
- No re-identification of patients or contacting of patients
- Only used for Research, Public Health, and Development for Covid-19

Limited data set

- Data de-identified as much as possible when used for research
- Secure platforms, DAC approval

Requirements

- Those using will have to abide by the terms of the agreement
- Time period for use of agreement
- Valid IRB that includes these limits (COVID research and COVID response planning)
- Any findings shared back to the consortium
- No secondary redistribution
N3C Phenotype & Data Acquisition Workstream

Christopher Chute, MD, DrPH

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
Defining a COVID-19 Phenotype: A consensus process (draw from many networks)

Inclusion criteria:
- All ages
- 14 days prior to first case in state
- At least two clinical encounters

Lab Confirmed Positive
- **LOINC** codes Positive result

Lab Confirmed Negative
- **LOINC** codes Negative result
- [may sample if number is large]

Likely Positive
- COVID Dx Code (other strong positive)

Possible Positive
- Two or more suggestive ICD codes

Data to pull:
- [One year record]
  - Observations
  - Specimens
  - Visit
  - Procedures
  - Drugs
  - Devices
  - Conditions
  - Measurements
  - Location
  - Provider

Phenotype and data ingestion effort led by Emily Pfaff at UNC
Example single-site workflow

Local EHR data warehouse → ETL → Local CDM

- Define COVID datamart extract (local CDM model)
- Define COVID cohort

SELECT *
FROM foo
WHERE...

Agreed-upon covid phenotype

Expert A
Expert B
Expert C
Expert D

ETL ~OR~ ETL

NCATS Cloud

Staging Database (multi-CDM)
- TriNetX covid data
- PCORnet covid data
- OMOP covid data
- ACT covid data

ETL

Production Database (unified CDM)

Data QA/Curation/Aggregation

NCATS Cloud

Researcher or clinician, querying secure analytical enclave
## Metadata Registration

- Makes the meaning of data publicly available and reusable, in human and machine-readable format
  - data interpretation, data validation, data transformation
- Persistent, unique identifier including version number
- Normalizes the meaning of the fields and the data values using standard NCIt terminology
- Enables interoperability for data that is not born interoperable

### Model

<table>
<thead>
<tr>
<th>Field Label</th>
<th>CDMH v1.0</th>
<th>PCORnet v 4.0</th>
<th>Sentinel v 6.0.2</th>
<th>i2B2ACT v 1.4</th>
<th>OMOP v 5.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public ID</td>
<td>6153917v1.0</td>
<td>6153919v1.0</td>
<td>6153920v1.0</td>
<td>6153918v1.0</td>
<td>6153921v1.0</td>
</tr>
<tr>
<td>BRIDG Name</td>
<td>Person Biological Entity Ethnic Group</td>
<td>Person Biological Entity Ethnic Group</td>
<td>Person Biological Entity Ethnic Group</td>
<td>Person Biological Entity Ethnic Group</td>
<td>Person Biological Entity Ethnic Group</td>
</tr>
<tr>
<td>CDM Value Domain</td>
<td>CDMH HL7 FHIR v3 Ethnicity Category Code</td>
<td>PCORnet CDM Hispanic Code</td>
<td>Sentinel CDM Hispanic Indicator</td>
<td>ACT I2B2 CDM Hispanic Indicator</td>
<td>OMOP CDM Ethnicity Category Code</td>
</tr>
<tr>
<td>Permissible Value(s)</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Value</th>
<th>Concept</th>
<th>Data Value</th>
<th>Concept</th>
<th>Data Value</th>
<th>Concept</th>
<th>Data Value</th>
<th>Concept</th>
<th>Data Value</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNK</td>
<td>C17998</td>
<td>UN</td>
<td>C17998</td>
<td>U</td>
<td>C17998</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NI</td>
<td>C53269</td>
<td>NI</td>
<td>C53269</td>
<td>NI</td>
<td>C53269</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2135-2</td>
<td>C17459</td>
<td>Y</td>
<td>C17459</td>
<td>Y</td>
<td>C17459</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2186-5</td>
<td>C41222</td>
<td>N</td>
<td>C41222</td>
<td>N</td>
<td>C41222</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTH</td>
<td>C17649</td>
<td>OT</td>
<td>C17649</td>
<td>N</td>
<td>C41222</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASKU</td>
<td>C79729</td>
<td>R</td>
<td>C79729</td>
<td>N</td>
<td>C41222</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Update and verify CDM model transforms**

**Target Data Model: OMOP 5.3**
Data Extraction from Sites

Steps for local data extraction

- Choose which CDM to use
- Execute pre-written query code
- Create local folder of output tables
- Transfer via sFTP to NCATS server

Support Resources for participating CTSA hub sites

- Helpdesk (white-glove service)
- Subject matter expert from corresponding CDM community
- CDM specific query “code”
- COVID data augmentations (optional)
- Transfer assistance to sFTP
- NCATS N3C support supplement
First Stage Data Quality Checks

NCATS Secure Cloud, Staging Area

Reincarnate CDM instance

CDM Data Quality Tooling

First Stage Ingestion

- Reconstitute CDM data into native database structures
- Run CDM specific Data Quality tooling and dashboards
- Check currency of value sets
- Iterate with contributing site to reconcile data (emphasis on first time submission)
N3C Data Ingestion & Harmonization Workstream

**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

**Founded upon ongoing work coordinated by CD2H**

- Interagency Clinical Data Model Harmonization project
- Terminology services and mapping tools
- FHIR as an interchange mechanism across CDM
NCATS Secure Cloud, Staging Area

Reincarnate CDM instance

ADEPTIA
Commercial ETL tool purchased by NCATS

Contributed Hub data in OMOP 5.3 instance

Second Stage Ingestion
- Transform Native CDM into OMOP 5.3
- Leverage library of maps maintained and updated in caDSR
- Identify variations from local CDM instance (second data quality check)
Third Data Quality Check

- Invoke OHDSI data quality tooling to create data quality checks and dashboards
- Return dashboard data to contributing hub sites
- Invoke results in data refresh cycles (no immediate iteration)
Data Integration from contributing sites into master OMOP dataset

Final Merge

- OMOP versioned data from all sources will be combined into analytic database
- Analytic database will migrate to Palantir Analytic Platform
Future Work

- FHIR as pluripotent data model
- Derive all CDMs and protocol specific schema as needed from common source
- Simplify ETL at hub sites using bulk FHIR APIs when available
- Facilitate transform into FDA ready formats to simplify clinical trial data management

Future phase work in partnership with Federal Clinical Data Model Harmonization (CDMH) project
Common Data Model Harmonization Project

Mitra Rocca

April 13, 2020
Agenda

Overview of the Patient-Centered Outcomes Research Trust Fund (PCORTF) Common Data Model Harmonization (CDMH)
Phase I Accomplishments
Phase II Deliverables
Overview: PCORTF CDM Harmonization Project

Goal:
Build a data infrastructure for conducting research using Real World Data (RWD) derived from the delivery of health care in routine clinical settings.

Objective:
Develop the method to harmonize the Common Data Models of various networks, allowing researchers to simply ask research questions on much larger amounts of RWD than currently possible, leveraging open standards and controlled terminologies to advance PCOR.
Different countries use different “outlets”.

There is a need for travel adapters.

**The Solution:**

Use a converter between various adapters.

Allow researchers to ask a question once and receive results from many different sources using a common, agreed-upon standard structure, or a Common Data Model.
Proposed Solution
Additional Goals of CDM Harmonization Project (1)

1. Develop a general framework (i.e., tools, processes, governance and standards) for transformation of various CDMs, curation, maintenance and sustainability.

2. Assess the value of the developed CDM harmonization mechanisms by demonstrating research utility for safety evaluation of cancer drugs that use the body’s immune system [programmed cell death (PD1) and programmed cell death ligand (PDL1) inhibitors] with a focus on patients with autoimmune disorders.

3. Reuse infrastructure developed by currently-funded OS PCORTF projects (NIH Common Data Elements (CDE) Repository, ....)
4. Leverage open standards and controlled terminologies to advance Patient-Centered Outcomes Research.

5. Test methods and tools developed by the collaborative on the universal CDM mapping and transformation approach.
Phase I Accomplishments

1. Harmonized 5 Common Data Models (i.e., Sentinel, PCORnet versions 3.1 and 4.0, OMOP and i2b2/ACT) with an intermediary model (BRIDG).

2. Developed the infrastructure (in collaboration with NIH/NCATS) to build a query, view, and store the results leveraging open, consensus-based standards.

3. Collaborated with Yale/Mayo Clinic as well as Elligo Health Research on the execution of the query focusing on the oncology use case.
Phase II Deliverables

1. Collaborate with new data partners leveraging the CDMH architecture as well as direct query from Electronic Health Records and Clinical Data Repositories.

2. Enhance the existing infrastructure to leverage Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) standard as the exchange data standard.

3. Submit Real World Data (RWD) leveraging clinical trial study data, leveraging Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) via the FDA Gateway.
N3C Collaborative Analytics Workstream

Ken Gersing, MD

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)
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?
ML model performance (random forest)

<table>
<thead>
<tr>
<th></th>
<th>Trained on real data</th>
<th>Trained on synthetic data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.817</td>
</tr>
<tr>
<td></td>
<td>F-Score</td>
<td>0.879</td>
</tr>
<tr>
<td>10-fold cross-</td>
<td>Accuracy</td>
<td>0.839</td>
</tr>
<tr>
<td>validation</td>
<td>Precision</td>
<td>0.802</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.704</td>
</tr>
<tr>
<td></td>
<td>F-Score</td>
<td>0.745</td>
</tr>
<tr>
<td>Test</td>
<td>Accuracy</td>
<td>0.846</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>0.836</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.671</td>
</tr>
<tr>
<td></td>
<td>F-Score</td>
<td>0.745</td>
</tr>
</tbody>
</table>

*Washington U Philip Payne
Computer Derived Synthetic Data: Validation of Sepsis Prediction
CDMH II: Standards and Architecture

**Data Partner with CDMs**

1. Unified Query Builder: (Write once using CQL/FHIR SQL which will work with all CDMs and FHIR Repository Server)

2. BULK FHIR Repository: Cloud-based virtual FHIR repository with security and management controls

3. Identified Warehouse

4. CDM Adapter: Allows CDMs (OMOP, PCORI, l2b2, Sentinel) to work with FHIR Repository and BRIDG model

**HHS Agencies**

6. Manage Results: Combine results and monitor task completion

7. Terminologies to FHIR Harmonization
   - SNOMED CT
   - ICD-10-CM
   - RxNorm
   - LOINC

8. Export Results: Exports to multiple formats for any agency i.e. CDC, FDA, NIH

**Terminology mapping**
- Local proprietary codes → FHIR

**Data Transformation**

**EHR**

**CDM to FHIR**

**HL7 FHIR Repository**

**Data Aggregator**

**Data Repository**

**Reuslts Viewer**

**HLS7 FHIR**

**CDC**

**NIH**

**FDA**

**BRIDG 1 (CDMH I)**

**BRIDG 1 to CDISC SDTM (CDMH I)**

**NCI, NCATS Terminology Server**
- CDM ↔ FHIR ↔ CDISC

**FDA Gateway**
NIDAP: Collaborative Analytics Platform: Palantir

Security and Auditability
- FedRamp Certified
- Can handle PHI
- Granular configuration and access controls - row, column, cell level configuration
- Logging auditability, security review, 2/7 monitoring with security audits
- Single sign-on
- Encryption in transit and at rest

Collaborative Ecosystems
- Common platform shared by many HHS agencies (CDC, FDA, NIH), multiple ICs (NCATS, NCI)
- Accommodate multiple data types: Clinical, diagnostic, genomic, imaging
- Work with time services data

Integration with other tools
- Easy to get data in and out, OpenAPI
- Analytics and Machine Learning and NLP support
- Complete version history, assist with reproducibility

Features
- Interpretability: support open source tools & languages such as SQL, Python, JAVA, Scala
- Complete lineage of dataset provenance
- Supports third party tools such as Tableau, R Studio, SAS, Jupyter, AWS, Azure
Architecting Attribution in the N3C

The N3C Collaborative analytics platform will support robust tracking of provenance and attribution; the DUA will require attribution of all scientific outcomes to everyone who contributed.

cd2h.org/attribution
Agency Partners

NCATS & CD2H

Other NIH ICs:
NIAID, NLM, NCI, NHLBI

Distributed networks:
PCORnet, ACT, OHDSI, TriNetX

Agencies:
FDA, HHS, VA
CDC & DoD (in discussions)
Join the conversation

Onboarding to N3C: [bit.ly/cd2h-onboarding-form](bit.ly/cd2h-onboarding-form)

Joining Workstreams:
- N3C Data Ingestion & Harmonization Workstream
  - Slack Channel Harmonization
  - Google Group Harmonization

- N3C Phenotype & Data Acquisition Workstream
  - Slack Channel Phenotype
  - Google Group Phenotype

- N3C Collaborative Analytics Workstream
  - Slack Channel Analytics
  - Google Group Analytics

- N3C Data Partnership & Governance Workstream
  - Slack Channel Governance
  - Google Group Governance

Additional Information:
- Onboarding N3C, Slack, Google
- Finding and Joining a Google Group
CD2H COVID Website

National COVID Cohort Collaborative (N3C)
In collaboration with several HHS agencies, the CTSA program, and distributed clinical data networks, CD2H has created the National COVID Cohort Collaborative (N3C), a centralized, secure, limited access portal to access COVID-19 clinical data.
Learn More

COVID Resource Search
COVIDsearch allows researchers and clinicians to search global publications, preprints, and clinical trials related to COVID-19. Resources are aggregated multiple times per day from various federal and medical sources.
Search COVID Resources

COVID Outbreak Information
Outbreaks of emerging diseases like COVID-19 increase the need for efficient collection, dissemination, and integration of data. The Outbreak.info online resource aggregates this information in an interactive dashboard.
Learn More

COVID Clinical Trials
Trials Today provides a mechanism to connect patients, their families, researchers, healthcare professionals, and the public with easy access to relevant information on publicly and privately supported clinical trials for COVID-19.
Search COVID Trials

covid.cd2h.org
Thank you!