The National Covid Cohort Collaborative

Interoperability Standards Priorities Task Force Meeting - April 16, 2021

Christopher Chute & Melissa Haendel

National
COVID
Cohort
Collaborative

A program of NIH's National Center
for Advancing Translational Sciences

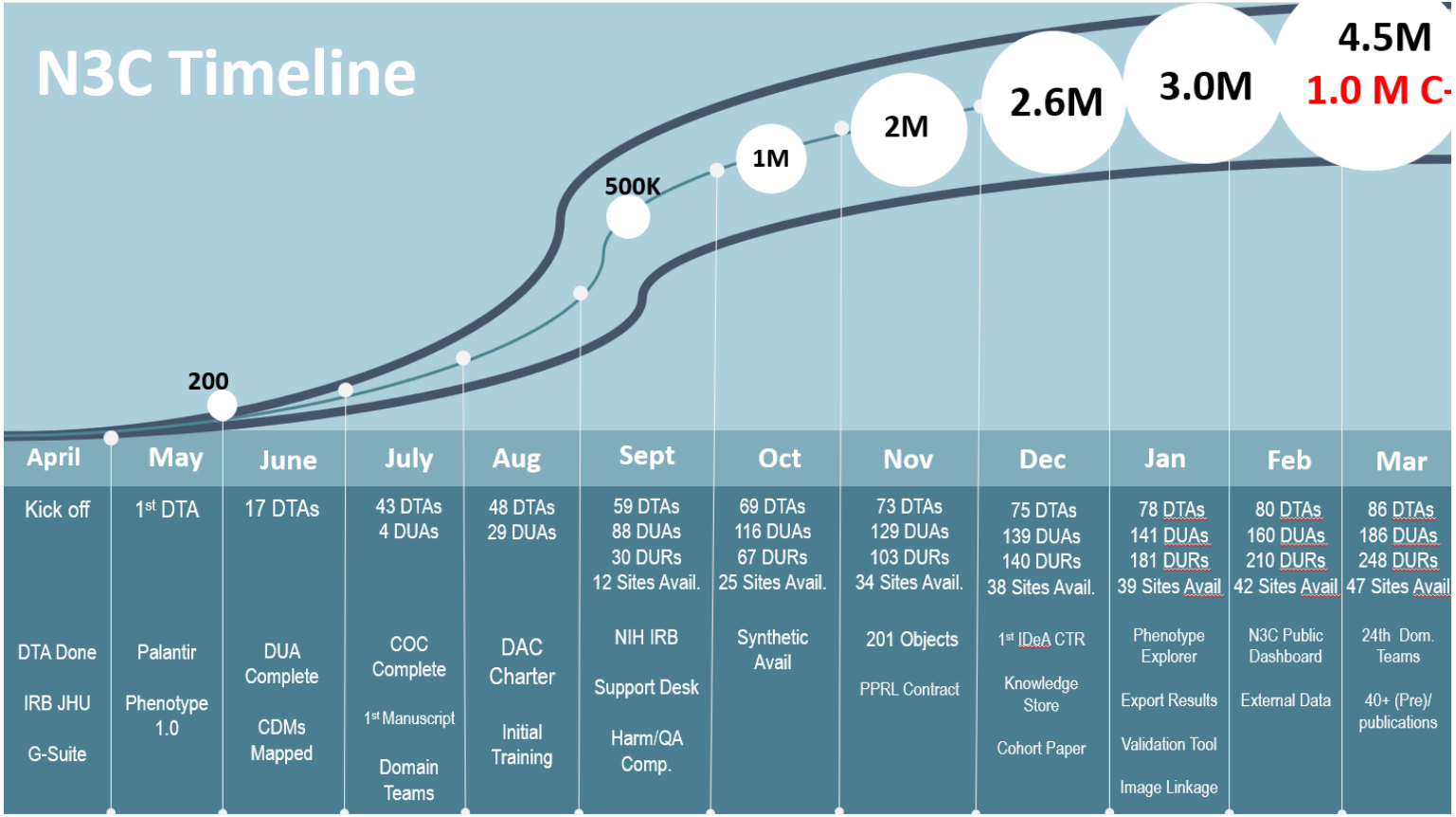
NATIONAL CENTER
FOR DATA TO HEALTH

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

N3C Timeline

NATIONAL CENTER FOR DATA TO HEALTH



National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

N3C Dashboard

covid.cd2h.org/dashboard

NATIONAL CENTER FOR DATA TO HEALTH

Sites: 50

Persons: 5.0 million

COVID+ Cases: 1,222,296

Total Number of Rows: 5.8 billion

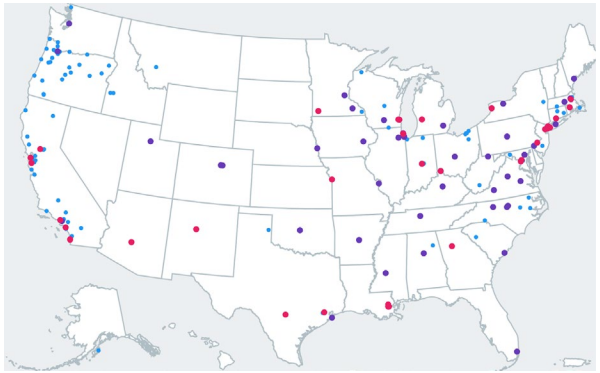
Clinical Observations: 721.4 million

Lab Results: 2.6 billion

Medication Records: 949.0 million

Procedures: 287.3 million

Visits: 257.8 million



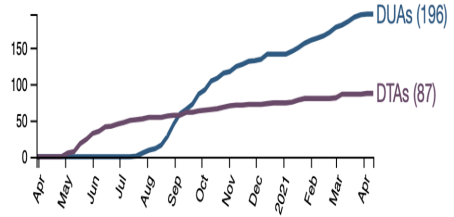
50 sites *with data released (purple) and* ***37 sites*** *with data pending (pink). OCHIN is a national network of* ***131 sites*** *(blue).*

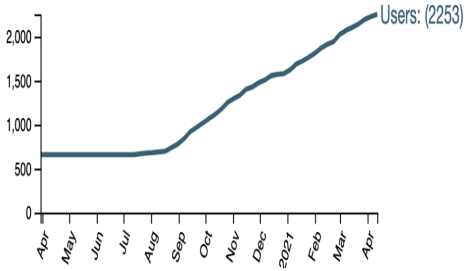
covid.cd2h.org/team

s

29 Domain teams!

Engagement and Registration Statistics

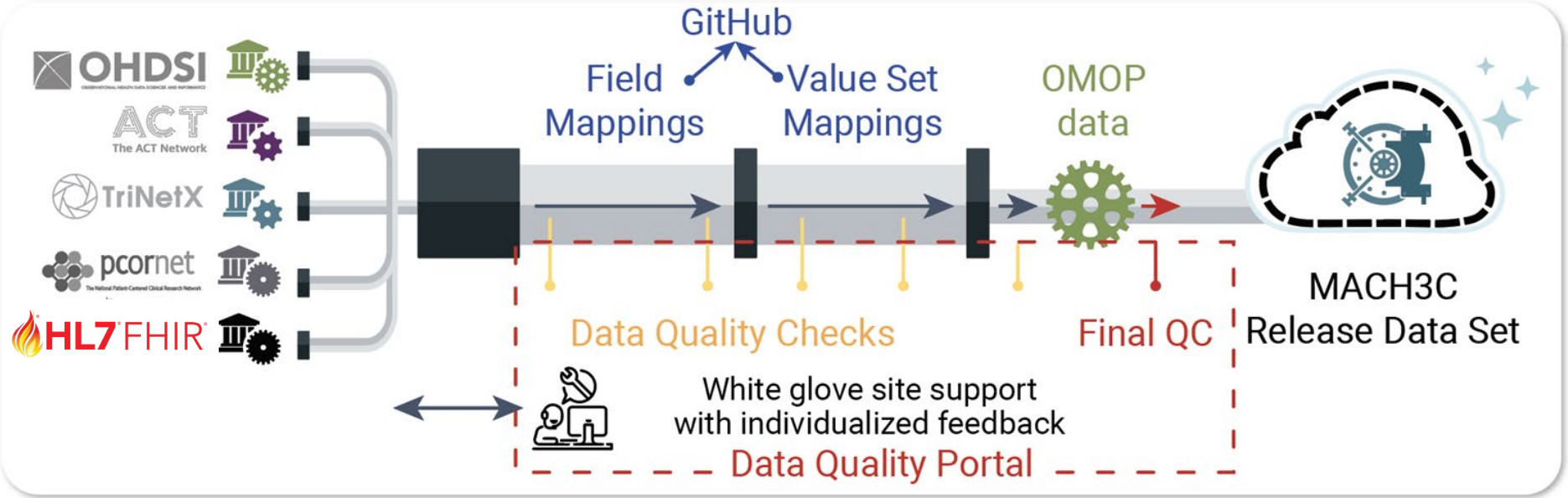




National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

N3C Data Ingestion & Harmonization Pipeline

NATIONAL CENTER FOR DATA TO HEALTH

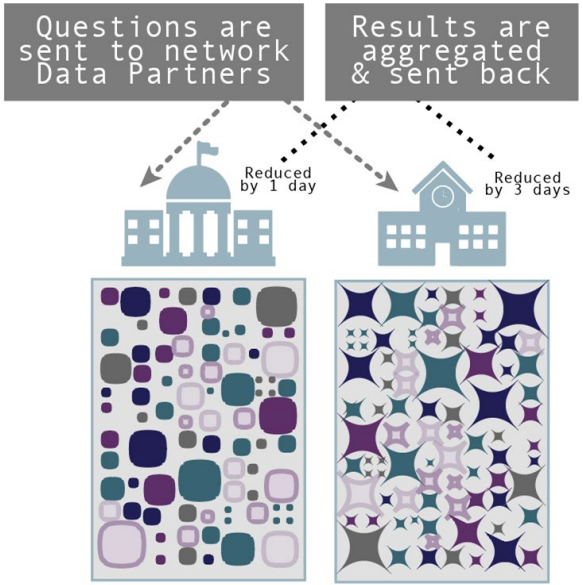


Span manual curation of mapping resources to industrial scale production transformation

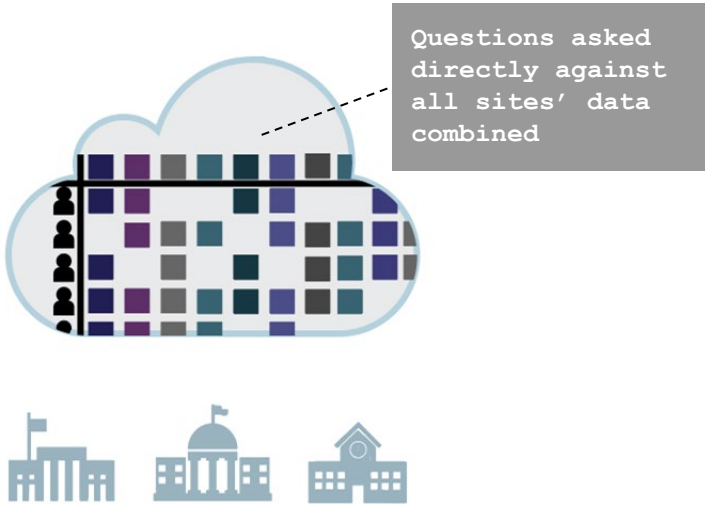
Federated versus Centralized DQ

Many clinical data research networks are *federated*; N3C is *centralized.* Centralized datasets have some advantages where data quality assessment is concerned.

Federated Network



Centralized Data

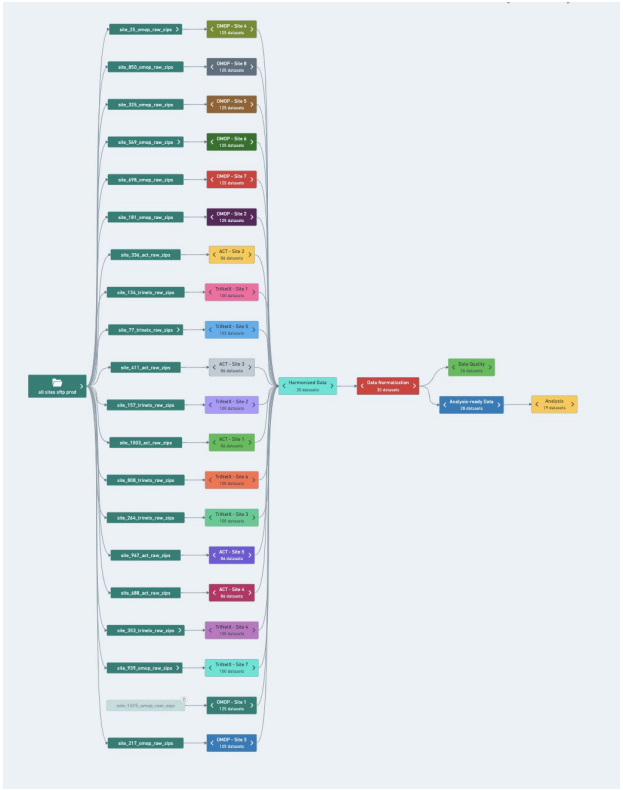


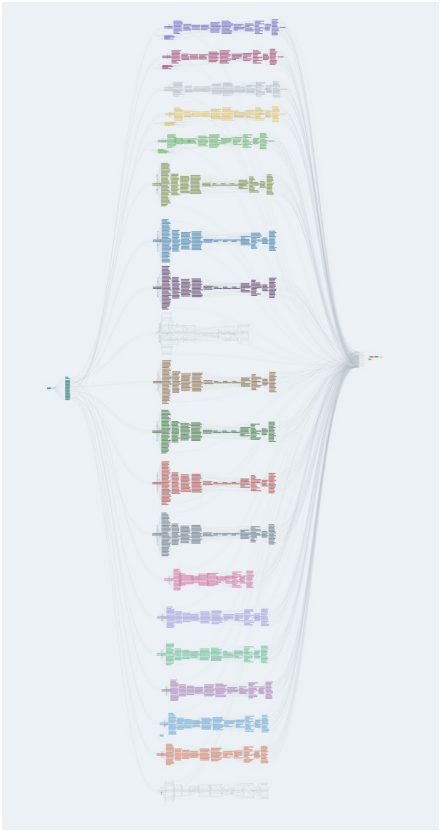
National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

Each of the 50+ sites has a pipeline with 100+ transformations

NATIONAL CENTER FOR DATA TO HEALTH





The provenance between 5000 transformations across the 50 sites is automatically tracked.

This enables:

- pipeline developers to very quickly identify the root cause of data quality issues
- data pipelines can be refreshed in <20 minutes whenever the source data updates

National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Each site has its own set of data health checks that run each time new data is submitted

NATIONAL CENTER FOR DATA TO HEALTH

- When the CDM mapping pipeline is deployed for a new site, it comes with a set of automated data health checks.
- These run every time the data updates - so that if new data doesn't meet expectations, the pipeline administrators are immediately alerted and can take action

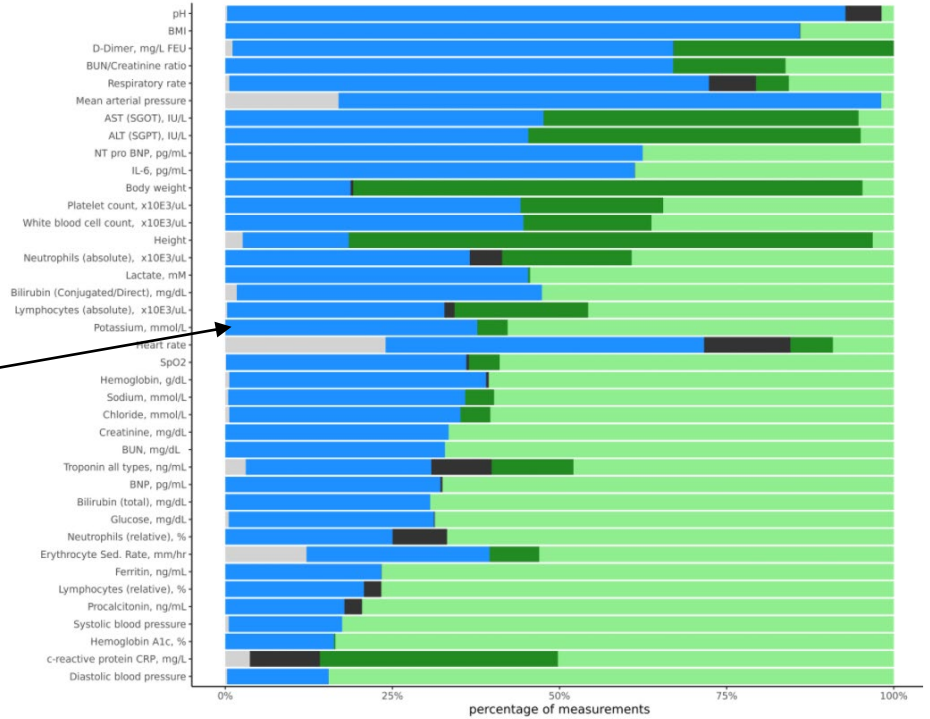
National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Unit harmonization example

NATIONAL CENTER FOR DATA TO HEALTH

~2x increase in usable data from N3C harmonization procedures

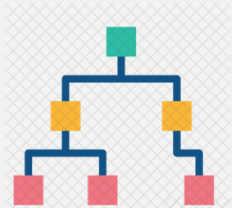
- Canonical unit
- Uses a known conversion
- Unit not plausible
- **Missing unit inferred**
- Unit still missing

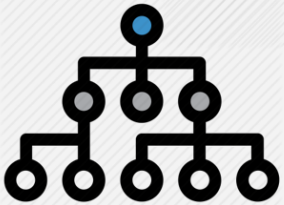


Centralized Data QC can rescue a lot of data!

Mapping is all over the place, and lossy

Source terminologies



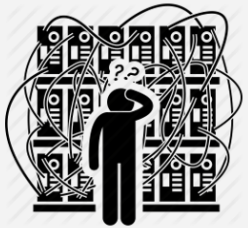


“Mappers”

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

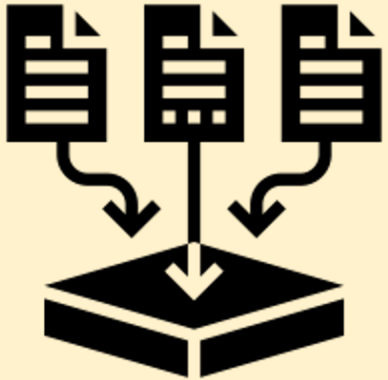
NIH NATIONAL CANCER INSTITUTE
Enterprise Vocabulary Services

Coded Data



Uncoded/locally coded Data

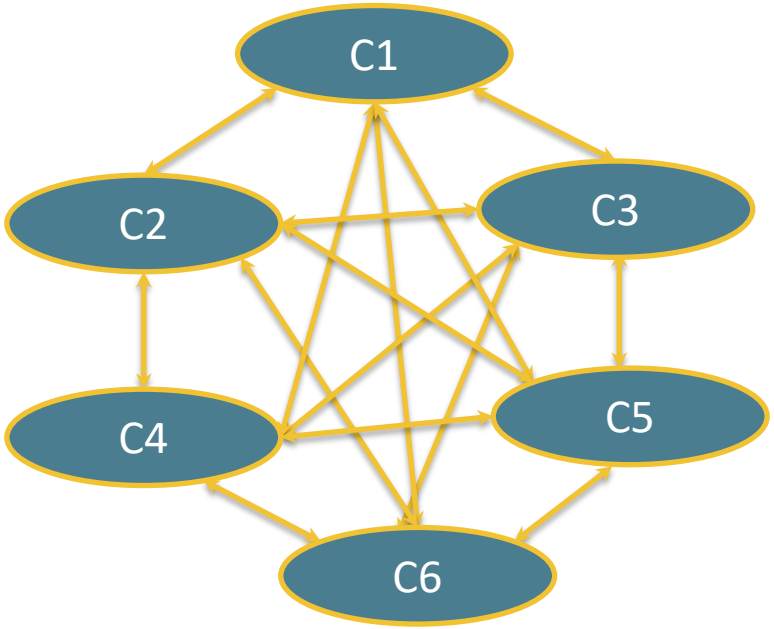
Codesets to unify data



{ }

Mapping is problematic for computational use

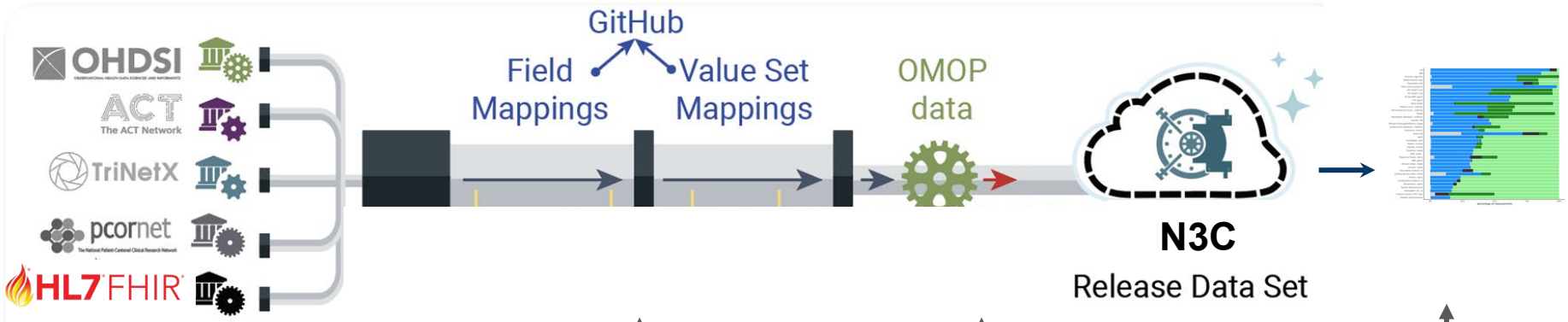
- Proliferation of mappings
  - Too many combinations
  - Frequently conflicting
  - Frequently stale
- Semantics unclear
  - Equivalent?
  - Exact?
  - Broad/Narrow/Related?
  - Without precise equivalence mapping, merging is not possible
  - No curation rules or provenance provided



(N^2)-N sets of mappings
(if each source provides their own mappings to all)

National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Potentially Lossy Mapping along the N3C Pipeline



Local codes to CDM terminology/codeset mapping

Field/Value set mappings to OMOP

Terminology mappings to OMOP concepts

Codeset creation across institutions

N3C Takeaways

What N3C has revealed most in terms of needs:

- Interoperability - we need syntactic and semantic!
  - FHIR ⇒ OMOP (syntactic)
  - Common vocabulary/codeset mapping provenance and management (semantic)
- Approach data harmonization from an end-to-end data life cycle perspective
- Leverage USCDI, but build for interoperable semantic modeling and extensions