N3C Timeline

A program of NIH’s National Center for Advancing Translational Sciences
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).

29 Domain teams!
Span manual curation of mapping resources to industrial scale production transformation
Many clinical data research networks are *federated*; N3C is *centralized*. Centralized datasets have some advantages where data quality assessment is concerned.

**Federated Network**

Questions are sent to network Data Partners. Results are aggregated & sent back.

**Centralized Data**

Questions asked directly against all sites’ data combined.
Each of the 50+ sites has a pipeline with 100+ transformations

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

- 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.
~2x increase in usable data from N3C harmonization procedures

Centralized Data QC can rescue a lot of data!
Mapping is all over the place, and lossy
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)
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 $\Rightarrow$ 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