Stephanie Hong, BS
Biomedical Informatics and Data Science
Johns Hopkins University

Davera Gabriel, RN
Biomedical Informatics and Data Science
Johns Hopkins University

Session 5: Importance of Interoperability

Hello and welcome to the Center for Leading Innovation and Collaboration's Insights to Inspire 2021 Informatics, A Journey to Interoperability. My name is Davera Gabriel. Stephanie Hong and I are members of the Biomedical Informatics and Data Science group at Johns Hopkins University. Today we will demonstrate/provide examples of the role interoperability placed in the COVID epidemic (N3C).

Objectives -Today aims to address the importance of interoperability. And we will provide information about our work with interoperable data during the COVID 19 pandemic. In today's presentation, we will provide some background on N3C's implementation status principles and goals. We will address how N3C supports data contributing sites, the mechanics of achieving interoperability within the N3C infrastructure, as well as approaches to ensure data quality and completeness will be discussed. And finally, we will touch upon future data and infrastructure enhancements that are planned for N3C.

National COVID Cohort Collective N3C - The National COVID Cohort Collective N3C is a partnership that aims to aggregate and use COVID-19 clinical data to answer critical research questions and to address the short and long-term effects of COVID-19 in the U.S. population. The partnership is led by the NCAT's CD2H and is supported through participation of data contributing sites that includes CTSAs, Idea Centers and health information exchanges, as well as data, technology partners, supporting the infrastructure and investigators pursuing discovery on the N3C technical platform. The primary data asset offered is the N3C data enclave. This is the result of an enormous data extraction, harmonization and aggregation efforts stored on the OMOP common data model. In N3C, the data being gathered from data partners has changed over time as the pandemic has unfolded and testing and cases have risen and now dropped dramatically. A dedicated team of specialists works with each site to facilitate all of this data collection as a service. Currently, data are extracted from confirmed COVID positive patients with records spanning back to January of 2018. To support comparative analysis, the phenotype and data extraction team is also collecting data for demographically matched controlled patients in a two to one ratio and phenotype descriptions and scripts for sites to extract data customized for each contributing data model are kept up to date and available on the N3C public GitHub repository.

N3C Timeline- The N3C data enclave was built very quickly as a result of the intensive effort that harmonized established implementations of common data models. This effort resulted in the world's largest single repository of COVID-19 case data and continues to aggregate over time, moving forward as contributing site data are refreshed on a continual basis.

N3C Dashboard - New releases of the data are available almost every week and up-to-date statistics and characterizations of data in the enclave are available via the N3C publicly available dashboard.
N3C Goals - From a programmatic perspective and N3C represents more than the considerable data asset. It offers to researchers as mentioned previously, N3C goals include utilization of the considerable investments. Contributing sites have made to extract and transform their EHR data into existing common data models. Harmonizing and aggregating these data into a centralized data national repository. In turn, the centralized data repository provides the opportunity for researchers to perform detailed analytics, not possible on established federated research network implementations. Utilization of N3C requires research transparency and generation of data and analytic artifacts that support reproducibility of research results, and thus N3C is a demonstration of real-world data utilization at scale. Further, the recognition of the contributions of all the multidisciplinary team members whose work contributes to aspects of the research work product.
N3C Principles - This expands on traditional evidence, dissemination models, which only record and preserve authors of scientific manuscripts. In alignment with the objectives of open science, N3C adheres to a set of operational principles, including: **Partnership**. N3C members must adhere to the N3C community guiding principles and user code of conduct. **Inclusivity**. N3C is open to any U.S. organization that wishes to contribute data. N3C also allows access to the data by registered researchers from any country who follow our governance processes. N3C also includes access for citizen and community scientists. **Transparency**. As mentioned previously, N3C data artifacts are preserved to support research. Reproducibility descriptions of projects are available to NTC partners and searchable to promote collaborations. **Reciprocity**. Contributions are acknowledged and results from analysis, including providence and attribution are expected to be shared within the N3C community. **Accountability**. N3C members take responsibility for their activity and hold each other accountable for achieving N3C objectives. **Security**. Activities are conducted in a secure, controlled access cloud-based environment and are recorded for auditing and attribution purposes.
Hub Participation - N3C has employed a number of processes, which aim to lower the burden to sites contributing data to the enclave. The NCATS smart IRB facilitated through Johns Hopkins as the single IRB of record provide sites the opportunity to significantly lower administrative cycles in order to contribute data. The phenotype and data acquisition work stream has developed multiple parallel data extraction scripts based on the common data model options open to N3C data partners. The combination of extraction scripts, which leverages common data model implementations, significantly lowers site expertise and labor requirements to successful data submission. Additionally, the data acquisition team provides personalized consultation and troubleshooting with site data managers responsible for N3C data payload contributions. Weekly office hour meetings provide drop-in opportunities to supplement the individual site white glove consultation sessions and facilitate collaboration with peer data managers among the data partner community. And finally N3C provides feedback about the quality of the data submissions to data partners.
This information can be used by the data partners to improve their source common data model implementations, or explore further upstream changes to their EHR data gathering and extraction processes.
N3C Interoperability Approach - Within the core open science principles, N3C's approach to implementation and interoperable data includes developing and supporting data sets that are transparent, easily shared, versioned, fully auditable, provide data and analytic provenance, and support reproducible analytics. The data ingestion and harmonization pipeline that Stephanie will review in detail next, performs syntactic and semantic alignment, preserves original source data that retains clinical intent and specificity and aligns, but does not impute data. Next. I will turn the presentation over to Stephanie Hong who will present some of the technical specifics of the data ingestion and harmonization pipeline and data quality processes.
Leveraging Common Data Models - The reality of COVID 19 urgency demanded collaboration across all medical centers in the U.S. CTSA hubs willingly comply by utilizing the CDM format that was already readily available and in use. Leveraging common data model OMOP, ACT, PCORnet, and Tri-Net X, N3C aims to support consistency in the data acquisition process across four CDMs. The goal of the data ingestion and harmonization workstream is to translate and harmonize this impacting semantic of data from all contributing sites into a single data model. Retaining as much specific city, original clinical intent as possible, as well as data quality and transparency. These steps support, N3C's ultimate goal

of producing comparable and consistent data to enable effective and efficient analytics. Common data model also allows us to write a consistent translation maps that can be run with few local changes at site with one or more data models.

N3C Interoperability Implementation: Data Element & Value Set Mapping - We have approached Translation in multiple levels. First of all, at the data ingestion level, data is submitted and ingested from a secured FedRAMP SFTP site where only the submitting data partners can see their data. Each CDM data elements are mapped to OMOP domain elements, and each of the value sets that are found in the data elements are also translated at two levels. Static value sets across a crosswalk table is used to map to OMOP concept. By these using OMOP vocabulary and dynamic semantic value sets, a mapping table is generated using the OMOP concept relationship ontology and is used to map to form up concept by these or ICD10CM, ICD 10 PCs, RX NORM, or LOINC codes to SNOMEDs. During the harmonization process, laboratory results are harmonized to canonical units and missing units are based on the numerical data distribution. A concept such a defined and shared for their data variables and also concepts that are defined and utilized to generate cohorts. Formulas and codes are shared from the knowledge store to generate common concepts like macro visits, BMI, Glasgow Coma Score, Convalescent Plasma therapy for data consistency.

N3C Data Ingestion & Harmonization Pipeline - We have built a CDM specific data pipeline to manage the ingestion and harmonization at scale. We have built data source specific ingestion pipeline, one for each CDM format: OMOP, ACT, PCORnet, and Tri-Net X. We ingest data by downloading and parsing the zip file from the secured SFTP location. And we have created value set mapping crosswalk table to translate all study codes like discharge status, race, gender, ethnicity. We also dynamically generate value set mapping crosswalk table for each site in order to translate ICD 10 codes or PCs codes, LOINC codes, RX Norm codes to SNOWMEDs. Laboratory results are harmonized to canonical units for data comparison, missing units based on the numerical data distribution. CDM specific pipeline code that make use of conformance checks, domain mapping, reference of crosswalk mapping tables for both static and dynamic value sets, and terminology. Domain, key ID, key generation and ID collision checks are all bundled into a template.

N3C Data Ingestion & Harmonization Pipeline - Shown is one instance of data ingestion and harmonization pipeline. We make use of the templatized code in order to ingest all sites data, and one OMOP vocabulary tables are used to ingest all data. Meaning whenever the OMOP vocabulary tables are updated, all of the sites data is re-ingested to update to one version of the OMOP vocabulary tables. This allows us to use the one version for vocabulary translation and the maintenance of code reusable using codes and ingestion and harmonization of all sites data can be performed at scale.

Each of 50+ sites has a pipeline with 100+ Transformations - Each of the 55 sites has a pipeline with over 5,000 transformation. The Providence between 5,000 plus transformation across all 55 plus sites is automatically checked. This enables pipeline developers to quickly identify the root cause of data quality issues and data pipeline can be refreshed less than 20 minutes. Whatever news sources submitted, all data sources ingested using one version of the OMOP vocabulary resulting in unified for vocabulary translation for all data sets. When new vocabulary files downloaded all the data states again are re-ingested.

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 automated data health checks. These run every time data updates so that if new data doesn't meet the expectation pipeline, administrators are immediately alerted to take action.

N3C's Data Quality Process - N3C's data quality processes are performed for every data submission, data ingestion and harmonization. Team checks our quality metrics for each site multiple times per week, utilizing the data quality portal within the enclave. Some of the checks that we perform weekly are a number of positive and negative COVID test results and it's ratio per site, average rows of data per patient per domain, references to patients found in the person table across all domains, demographic, distribution, and visits with negative lengths of stay.

Unit Harmonization example - We are able to rescue unreal usable data by harmonizing to canonical units and converting data by using data conversion formulas. We also look at the sites data distribution to generate units.

Harmonizing numeric data - Sometimes different sites provide laboratory results in different units. So we harmonize the standard units so that data can be correctly compared for analysis. Harmonizing numerical data helps us to contact the source data partners to let them be informed of areas where they may be able to correct. We also use the inference engine to calculate different values in different units to the same units.

Harmonization Progress - We check harmonized measurement units, not just by data partners, but across all data partners for accuracy. These kinds of checks are only possible with N3C data sets.
Measurement Data Set is Enriched with harmonized values as number - We have enriched the measurement data set with harmonized value_as_number and harmonized value as units, so that it's available to all domain research groups.
N3C Data Element & Infrastructure Enhancement Plans - So in the future N3C will continue to work with data partner sites to improve data quality. Additionally, N3C will support efforts to include additional data not available due to site configurations or common data model specifications. N3C has provided an opportunity to compare in detail the differences in value set utilization for common data classes among the common data models, and will seek opportunities to promote uniformity of these sets in the future. N3C has a cross- cutting domain team that has developed a natural language processing module that is in a pilot implementation phase, and will seek to leverage the results of this work and to add data from text sources into the enclave. N3C has also engaged leadership in the Odyssey community, focusing on the OMOP vocabulary and is developing N3C quality improvement approaches as a result of this close collaboration. Lastly, N3C in partnership has developed a privacy preserving record linking or PPRL governance and solution, which leverages an intermediary honest broker that will allow researchers to access data beyond the enclave and the ingestion pipeline to include in their analysis. Pilot implementation of this record linking approach has worked well with imaging databases and will expand in the future to include a wide array of additional data opportunities.
N3C Summary Points - I hope our presentation today has provided some insight regarding our example of the power and utility that interoperable data bring to the research enterprise.
In summary

- N3C is the largest available repository of COVID-19 case data.
- Is based on open science principles and an open technical platform
- Represents a comprehensive approach to data harmonization
- Has rapidly developed a complex and robust infrastructure giving new life to prior cite data investments
- As achieved deployment of a powerful data resource that supports of advanced analytics, despite known persistent issues of EHR source data heterogeneity.

If you are interested in more information about N3C, especially how you can become involved, please explore the information on the pages accessible via the links provided here.

Register with N3C: https://lahs.cd2h.org/registration/
Join a Domain Team: https://covid.cd2h.org/domain-teams
Submit a manuscript or a publication: https://covid.cd2h.org/publication-review

We encourage you to view the following webcasts as they provide foundational information for the rest of the series.

Session 1 – Introduction to Insights to Inspire
Session 2 – Language of Informatics
Session 3 - Introduction to Informatics
Session 4 - Introduction to Maturity Models
Session 5 - Importance of Interoperability

After that, please view the remaining webcasts in the order of your choice. The webcasts are available on the CLIC website: https://clic-ctsa.org/ or you can access them by searching CLIC_CTSA on Vimeo or YouTube.

Thank you for viewing Session 5 – ***The Importance of Interoperability***.
Please join us for the rest of the ***Informatics, The Journey to Interoperability series***.