LHC eSource Symposium – Summary of Presentation by Dr. Christopher Chute

8 September 2021

"Synergy Between Federated and Centralized Clinical Data Repositories and Its Impact on Data Quality"

Dr. Christopher Chute [Bloomberg Distinguished Professor of Health Informatics, Professor of Medicine, Public Health, and Nursing at Johns Hopkins University, and Chief Research Information Officer for Johns Hopkins Medicine] initiated his presentation for this symposium by reminding the audience of the fact that data is always more useful when it is 'consistently classified' as compared to collections of data that are like Legos of different sizes and shapes and colors. He referred to the FAIR principle: Findable, Accessible, Interoperable and Reusable. Interoperable, when referring to data, means *harmonized.* Syntactic interoperability refers to the structure in data exchanges while semantic interoperability refers to ensuring that the meaning is preserved when exchanging data. Harmonization of multiple datasets or models typically requires the selection of a 'target' or a reference standard to which everything else can then be harmonized.

Dr. Chute then introduced the National Covid Cohort Collaborative (N3C), which is a partnership among CTSA program institutions, distributed networks that work with clinical data, including OHDSI, ACT/i2b2 and PCORNet, TriNetX and many others. The N3C has developed a centralized, secure portal that houses Covid data; it provides a shared space or common repository for research data access; it evaluates methods and tools for researchers to apply to these data. The data are harmonized, with the reference standard being the OMOP model. Approximately 2,500 individuals are involved in this project. The N3C workstreams are as follows:

1) Data Partnerships and Governance
2) Phenotype and Data Acquisition
3) Data Ingestion and Harmonization
4) Collaborative Analytics
5) Synthetic Clinical Data

There is a very detailed process that is applied when new sites join the N3C. This is designed to ensure regulatory requirements are met and the site is appropriately onboarded. N3C data are ingested and harmonized across the various data models and eventually to the target model, OMOP. Use of HL7 FHIR is a goal for the future. Manual curation of mapping resources is done to an industrial scale.

Collaborative analytical teams have formed around various areas, including critical care, imaging, immunosuppressed or immunocompromised, oncology, short-term and long-term complications, hypercoagulability, diabetes, pediatrics, pregnancy and even social determinants of health (SDoH). As of early September, the N3C was analyzing data from over 2.5 million Covid-positive patients out of over 7.5 million total patients from 64 sites. Dr. Chute emphasized that "centralizing patient-level data makes it possible to ask qualitatively different and more powerful questions, but is only possible due to each institution having their data in a common data model." Local teams understand the way that their data are collected and stored and the nuances of those data and provide value from this federation. Centralized data benefits from local curation and quality control.

Dr. Chute gave several examples of how the N3C data were informative. He also showed how the data are harmonized, for example, ensuring that each numeric value is associated with the same (standard) units (e.g. pounds, ounces, grams) in the centralized data. When values are missing, the source is contacted or an inference engine can be applied. Harmonization leads to homogeneity. Harmonization significantly increases data usability. In the case of N3C, the usable data was doubled with harmonization procedures.

In summary, Dr. Chute presented the following observations:

- Local data curation is critical to data quality; federated common data models foster additional consistency.
- Centralized data enables unique benchmarking and can occasionally 'repair' or salvage data. They can iteratively enhance submissions with feedback.
- Both federated and centralized models add tremendous benefits to data quality.

The N3C project was funded by NIH/NCATS. Attributions were made to Dr. Ken Gersing (NCATS) and Dr. Melissa Haendel (co-lead with Dr. Chute) along with Emily Pfaff, Katie Bradwell, Richard Moffitt and numerous (hundreds of) others within the N3C consortium.

See https://ncats.nih.gov/n3c for more information.