Ensuring Quality: A Core Competency of Federated EHR Data Networks

Jeffrey G. Klann, PhD[1,2,3]; Darren W. Henderson[4]; Shyam Visweswaran, MD, PhD[5]; Hossein Estiri, PhD[1,2]; Shawn N. Murphy, MD, PhD[1,2,3]

[1]Lab of Computer Science, Massachusetts General Hospital, Boston, MA; [2]Harvard Medical School, Boston, MA; [3]Research Information Science and Computing, Massachusetts General Hospital, Boston, MA; [4]University of Kentucky, Lexington, KY; [5]University of Pittsburgh, Pittsburgh, PA

Abstract

The growing prevalence of Clinical Data Research Networks supports nation- and even world-wide research on electronic health record data. Although national programs like PCORnet and ACT have developed network resources over years, COVID-19 accelerated development of innovative new approaches, including the Consortium for Clinical Characterization of COVID-19 by EHR (4CE), an innovative decentralized network with multiple source data formats. The panelists have been involved in methods and tool development for both ACT and 4CE and will identify core network components that ensure data quality and interoperability. These include: consistent organization and term naming (i.e., a biomedical ontology); multi-site evaluations of data completeness; and, deep harmonization across sites through chart review and validation. Together, these enable large-scale data analytics including machine learning in these networks. This panel will discuss these components (naming, completeness, deep harmonization, and impact on analyses) with practical examples that allow these networks to transcend previous accomplishments.

Introduction and Background

The promise of Clinical Data Research Networks is the ability to harness EHR data to make generalizable discoveries using data on large swaths of the population. Some networks, like the Patient Centered Outcomes Research Network (PCORnet) [1,2] and the Accrual to Clinical Trials Network (ACT) [3], have been in development for years and have yielded many methodological improvements in data quality and interoperability. Others, such as the Consortium for Clinical Characterization of COVID-19 by EHR (4CE) have been built in response to COVID-19 and respond with great agility to the continuing pandemic. [4] With the establishment of data research networks as an emerging commodity, we believe it is possible to identify core components needed to ensure network data quality and the generalizability of findings.

To ensure quality, a network must include the following three components. First, sites need to adopt consistent naming in an organized manner across the network. Network sites must agree on the biomedical ontology used for data in the network which must be adapted to each site through mappings. [5] Second, data must be evaluated for completeness. Many existing data quality checks focus on completeness [6] (e.g., "what percent of diabetic patients with a prescription have an A1C test?"). A newer approach is to identify subsets of patients who are 'loyal' - i.e., have complete data because they receive most of their care at an institution that is a member of the network. [7] Third, data must be harmonized across sites. We call this "deep harmonization" because it goes deeper than just mapping codes. We advocate comparing data distributions across sites to ensure consistent mapping, and validation through both chart review and comparison of the underlying data and codes when applying phenotyping algorithms. Code mappings can dramatically vary from site-to-site. [8]

Panel Description

Dr. Shawn Murphy will lead the panel discussion on the core components of quality in data research networks, and in doing so will help the audience better understand the methodologies that can enable high-quality clinical data research within the U.S. and internationally. Dr. Murphy is a Professor of Neurology and Biomedical Informatics at Harvard Medical School, Chief Research Information Officer at Mass General Brigham, and the Associate Director of the Lab of Computer Science at MGH. He is one of the Principal Investigators for the Data Resource Core of the large NIH RECOVER program that focuses on the study of post-acute sequelae of COVID-19 infection.

The panelists will discuss these core quality-enabling components with perspectives and tools from the 4CE Consortium and the ACT network. The ACT network is a well-established national research network embodying established best practices [3]. In contrast, 4CE is an agile decentralized consortium developed at the beginning of the COVID-19 outbreak that has pushed the envelope of new methodological approaches and tools. [4] ACT has

demonstrated the ability to interchange data at scale through mappings to standardized ontologies. [5] The 4CE philosophy is that researchers must stay as close to the data as possible at each site. This way, studies can respond to the actual complexities of the data in implementation. In 4CE, research questions and methodologies are refined by going back to those who know the data best, directly involving the researchers and analysts at each site in network-wide analytics.

This panel is timely and needed because of the *growing urgency* to utilize the burgeoning wealth of EHR data, especially in response to the public health needs of the ongoing pandemic. This group of panelists have multi-institutional experience with two prominent flagship data networks. Given the national scope of ACT and international scope of 4CE, the proposed panel is well suited for the AMIA members and audience.

Presenters

Naming the Data (Visweswaran)

Dr. Shyam Visweswaran is an Associate Professor of Biomedical Informatics and is the Director of the CTSI's Biomedical Informatics Core at the University of Pittsburgh in Pennsylvania. He leads the Data Harmonization Workgroup that develops the ontologies for the ACT network. He will discuss biomedical ontologies in i2b2/SHRINE data research networks, with a focus on the ACT network and the enhanced ontology functionality developed for COVID-19 research. Ontologies offer not only a way to query the data but also an information model to organize the data. The i2b2/SHRINE ontologies have a tree-like hierarchical structure in which concepts closer to the root are more general than concepts located near the leaves. The tree-like structure enables the investigator to navigate the concepts and construct succinct queries by using the most general concepts that are applicable. The ACT network rapidly developed and validated a specialized COVID-19 ontology and deployed it on the network, and ACT sites augmented their data to support the ontology. The latest version of the ACT COVID-19 ontology, Version 4, consists of 52,476 codes in the domains of diagnosis, procedures, medications, and laboratory tests. The COVID-19 ontology has several unique features. It has categorized *emerging terms* from ICD-10, CPT-4, HCPCS, and LOINC terminologies that were introduced in response to SARS-CoV-2. It includes *computable phenotypes* to characterize the course of illness and outcomes in COVID-19 that include illness severity, respiratory therapy management, and level of care. And it contains *harmonized value sets* for the growing number of SARS-CoV-2 nucleic acid antigen and antibody tests. [9]

Assessing Completeness by Finding Loyal Patients (Henderson)

Darren W Henderson is a Database Administrator at the University of Kentucky focused on data warehouse development and performance tuning. He collaborated on the development of a tool to identify whether patients are "loyal" - meaning they get most of their care within a healthcare system and are thus likely to have longitudinally complete data. This tool implements a previously-validated algorithm which relies on the presence of proxies for loyalty such as prostate-specific antigen (PSA) tests, pap smears, and recent visits. [7] This algorithm has been shown to enhance the performance of machine learning methods by choosing this "enriched" cohort. Mr. Henderson will discuss the loyalty methodology and related approaches presently being used to enhance data network completeness.

Deep Harmonization Across Sites (Klann)

Dr. Klann is an Assistant Professor of Medicine in the MGH Laboratory of Computer Science and Harvard Medical School. Data harmonization must include data mapping to common terminologies. But even when this is accomplished, coding differences across sites can still abound. There are many LOINC codes for the same test, and if institutions choose different (but equally valid) codes, it could cause phenotyping algorithms to fail. For data to be truly harmonized, the underlying codes in the data must be examined. To ensure network harmonization, it is essential to explore the network-wide distribution of codes used to represent a given concept. Additionally, chart review on a subset of patients must be performed to challenge assumptions and discover intricacies in data encoding practices. Dr. Klann will discuss his work in two related projects: comparing count distributions in the ACT network, and enhancing an international phenotype of COVID-19 severity in the 4CE network. [8]

Impact of Data Completeness on Downstream Machine Learning (Estiri)

Dr. Hossein Estiri is an Assistant Professor of Medicine at MGH Lab of Computer Science and Harvard Medical School. Dr. Estiri develops novel computational phenotyping and predictive algorithms using clinical data. In his

presentation, Dr. Estiri will discuss ways in which longitudinal completeness of patient records in EHRs may impact the performance of machine learning (ML) algorithms measured by discrimination power, reliability, and bias. In binary classification tasks, which are widely exercised in healthcare research, standard ML performance evaluation is often focused on descrimination metrics, such as the confusion matrix, area under the receiver operating and/or precision-recall curves. Reliability of predictions are also evaluated from time to time, especially in the context of predictive models. Dr. Estiri has developed the MLHO pipeline [10], an end-to-end ML pipeline, that outputs a comprehensive set of ML performance metrics, including discrimination and reliability, as well as algorithm-level and patient-level measurements of bias. Dr. Estiri will showcase how different levels of longitudinal data completeness (quantified through the loyalty methodology) can prospectively alter discrimination, reliability, and bias in MLHO's predictions of mortality, ventilation, and ICU and hospital admission due to COVID-19.

Discussion Questions

1. How can biomedical ontologies aid in data quality? What are the unique features of the COVID-19 ontology that enable phenotyping?
2. Even when sites are working in the same data model, differences in set size, patient population disparities, and hardware/database differences can stymie progress in developing reproducible algorithms. What techniques can mitigate these issues?
3. How can chart review be used to augment the structured data coming from EHRs?
4. How can longitudinal data completeness impact machine learning perfromance in terms of discrimination, reliability, and bias?

Panel Organizer Statement: All participants have agreed to take part in the panel and discuss the topics as outlined above.

References

1 Collins FS, Hudson KL, Briggs JP, *et al.* PCORnet: turning a dream into reality. *J Am Med Inform Assoc* 2014;**21**:576–7.

2 Bian J, Lyu T, Loiacono A, *et al.* Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J Am Med Inform Assoc* 2020;**27**:1999–2010.

3 Visweswaran S, Becich MJ, D'Itri VS, *et al.* Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. *JAMIA Open* 10/2018;**1**:147–52.

4 Brat GA, Weber GM, Gehlenborg N, *et al.* International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med* 2020;**3**:109.

5 Klann JG, Abend A, Raghavan VA, *et al.* Data interchange using i2b2. *J Am Med Inform Assoc* 2016;**23**:909–15.

6 Qualls LG, Phillips TA, Hammill BG, *et al.* Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *EGEMS (Wash DC)* 2018;**6**:3.

7 Lin KJ, Rosenthal GE, Murphy SN, *et al.* External Validation of an Algorithm to Identify Patients with High Data-Completeness in Electronic Health Records for Comparative Effectiveness Research. *Clin Epidemiol* 2020;**12**:133–41.

8 Klann JG, Estiri H, Weber GM, *et al.* Validation of an internationally derived patient severity phenotype to support COVID-19 analytics from electronic health record data. *J Am Med Inform Assoc* 2021;**28**:1411–20.

9 Visweswaran S, Samayamuthu MJ, Morris M, *et al.* Development of a Coronavirus Disease 2019 (COVID-19) Application Ontology for the Accrual to Clinical Trials (ACT) network. *JAMIA Open* 2021;**4**:ooab036.

10 Estiri H, Strasser ZH, Murphy SN. Individualized prediction of COVID-19 adverse outcomes with MLHO. *Sci Rep* 2021;**11**:5322.