

The PCORnet Learning Cycle

**Keith Marsolo, PhD¹, Bradley Hammill, DrPH², Laura Qualls, MS², Jeffrey Brown, PhD³,
Lesley Curtis, PhD²**

¹Cincinnati Children's Hospital Medical Center, Cincinnati, OH; ²Duke Clinical Research Institute, Durham, NC; ³Harvard Pilgrim Health Care Institute, Boston, MA;

Abstract

The National Patient-Centered Clinical Research Network (PCORnet) is a distributed network-of-networks, where partners transform data from electronic health record (EHR) and claims data sources into a Common Data Model. In order to ensure that the underlying data are suitable for use in network queries, PCORnet employs a robust data characterization process. The tools and processes that support PCORnet are not static, however. They are constantly being refined in response to stakeholder needs as well as the findings from analytic queries and characterization results. This panel will describe the Learning Cycle that informs the activities of the PCORnet Distributed Research Network Operations Center (DRN OC) and will highlight the relationship between the PCORnet CDM, the Data Characterization process, the findings of the PCORnet Demonstration Projects, and downstream analytical tool development. In addition, the panel will discuss how all of these findings are communicated out to network participants and other stakeholders. Attendees will gain an understanding of why these components are all necessary to the successful operation of a distributed research network, and recognize how similar Learning Cycles can be used to improve the quality and reproducibility of any network's research.

General description of panel

Distributed research networks (DRNs) are critical components of the strategic roadmaps for agencies like the National Institutes of Health and Food and Drug Administration as they seek to move towards large-scale systems of evidence generation (1). The promise and sustainability of these systems hinges on the ability to extract usable, high-quality data from sources like the electronic health record (EHR), claims and others that are fit to support translational, interventional, and observational research initiatives. There are several active DRNs in the United States, including the Sentinel Initiative (formerly Mini-Sentinel) (2, 3), the Health Care Systems Research Network (HCSRN) (formerly the Health Maintenance Organization Research Network (HMORN))(4-6) and the National Patient-Centered Clinical Research Network (PCORnet) (7, 8). While these DRNs all have different operational characteristics, they share some commonalities. The first is that each network utilizes a common data model (CDM) to ensure that all partners (network participants) represent their data in a standardized manner. The second is that the data stay local, and the queries or analyses are distributed. The analytical tools differ by network, but are generally tailored to leverage the available data and to support the major use cases or study designs of interest. What they do have in common is that typically only the results, in the form of aggregate counts or summary statistics, are sent back to the requestor (9, 10). Therefore, before conducting any analysis, it is important to first understand the quality of the data that are to be queried to ensure that the results returned by each partner are valid and trustworthy (11). To that end, many DRNs have developed data characterization routines which are used to summarize missing values, outliers, and frequency distributions, and they also ask partners to complete surveys about the provenance of their data (3, 11). The characterization results then help inform an investigator's decision about whether a dataset is suited to answering a given research question.

These activities do not occur in a silo, however, but inform one another. Data characterization routines may uncover variability in how network partners are populating certain elements of the network's CDM. Study-specific analyses may uncover quality issues that are masked by a data characterization package that is designed to provide a broad assessment on data quality. Such findings can have a number of downstream impacts. For instance, variability in the population of the CDM may result in specific guidance to partners on how they should structure their extract-transform-load procedures and to the analytic teams on how they should structure their queries. Quality issues uncovered through study-specific characterization routines may result in changes to the base characterization process, or modifications to the network's CDM in order to better represent the data in question. All of this then has an impact on the design and developmental roadmap of the analytical tools that support the network. In this manner, a DRN follows a learning cycle with a goal of continuous improvement, which can be linked to a number of metrics, such as overall data quality or query throughput. To that end, this panel will describe the Learning Cycle that informs the activities of the Distributed Research Network Operations Center (DRN OC) of the PCORnet Coordinating Center.

PCORnet is a distributed network of networks, comprised of 13 Clinical Data Research Networks (CDRNs) and 21 Patient-Powered Research Networks (PPRNs). PCORnet's CDRN network partners include hospitals, integrated health systems, ambulatory care clinics, and health plans, which contribute electronic health record data, billing system data, and other electronic health data such as clinical registries and health insurance claims. Altogether, these CDRNs operate more than 80 individual datamarts, which contain data on tens of millions of patients standardized to the PCORnet CDM. The DRN OC of the PCORnet Coordinating Center is responsible for characterizing the datamart(s) of each network partner and evaluating the data's fitness-for-use across a broad research portfolio through a foundational data characterization process. Data characterization is designed to complement any of the CDRNs' internal data quality efforts while generating meaningful, actionable information for the CDRNs, Coordinating Center, sponsor, and other stakeholders. In particular, these results are used to inform the development and deployment of the analytical query tools that support the network, which includes both rapid, prep-to-research queries, as well as observational and comparative effectiveness studies. While the foundational data characterization process does not necessarily determine whether a given datamart is fit-for-use in a specific research study, the results inform study planning decisions, including what to include in the more comprehensive study-specific data characterization routines. These routines are executed prior to the main study analyses, and help assess the quality of the outcomes and variables of interest for the cohort in question.

Panelist presentations

Laura Qualls is a Project Leader at the Duke Clinical Research Institute (DCRI) and is responsible for the implementation of PCORnet's foundational data characterization process. She will provide an overview of the current process and illustrate how it has evolved over time. In addition, Ms. Qualls will share selected results for the current data characterization cycle and explain how they are used to inform the query activities of PCORnet.

Keith Marsolo is an Associate Professor in the Division of Biomedical Informatics at Cincinnati Children's Hospital Medical Center. He is a co-investigator in the PCORnet Coordinating Center's DRN OC where he provides faculty oversight for the efforts related to data characterization and the PCORnet CDM. He will describe how the results of the initial data characterization cycle led to the creation of a document that provides implementation guidance to network partners as they populate the PCORnet CDM. Dr. Marsolo will also cover the process used to update this "Implementation Guidance" and how the material is disseminated and communicated to stakeholders.

Bradley Hammill is a Faculty Statistician at the DCRI. He is a co-investigator on the PCORnet Demonstration Study, Aspirin Dosing: A Patient-centric Trial Assessing Benefits and Long-Term Effectiveness (ADAPTABLE), a pragmatic clinical trial that is comparing the effectiveness of two different daily doses of aspirin to prevent heart attacks and stroke in patients with heart disease (12) as well as a study that transformed Medicare fee-for-service claims data into the PCORnet CDM and tested linkage rates between PCORnet patients and Medicare beneficiaries. Dr. Hammill will provide a high-level overview of these activities and describe how his findings led to changes in the PCORnet CDM specification and implementation guidance, as well as the foundational data characterization process.

Jeff Brown is an Associate Professor in the Department of Population Medicine at the Harvard Pilgrim Health Care Institute. He is a co-lead of the DRN OC and oversees the team responsible for the fulfillment of PCORnet queries and for the development and support of the PCORnet analytical query tools. He will provide an overview of the current suite of analytic tools and describe how the findings of data characterization and the initial PCORnet studies have influenced the developmental roadmap of future tools, as well as plans to increase the throughput of query activities, such as allowing query fulfillment activities to occur outside of the PCORnet Coordinating Center.

Lesley Curtis is a Professor in the Department of Medicine and Director of the Center for Population Health Sciences in the Duke University School of Medicine. She is a co-lead of the DRN OC along with Dr. Brown. Dr. Curtis will serve as moderator for the panel, setting the stage for the presentations of the other panelists by providing a high-level overview PCORnet, DRN OC activities and a summary of queries issued to-date. She will also discuss the learning cycle that underpins the work of the DRN OC and how it ties together the different workstreams.

Importance of topic

DRNs provide an important platform for the conduct of large-scale observational and comparative effectiveness research. Understanding the relationship between the CDM, data characterization and analytic tool utilization/development is important, as it can have an impact on the reproducibility of the research results and the ability of the network to meet the needs of its stakeholders, which is crucial for long-term sustainability. In addition, the general principles articulated in this panel have applicability beyond distributed research networks. Any project that seeks to use EHR data that have been transformed into a CDM will face similar issues. Raising awareness is the first step towards minimizing error and variability, ultimately leading to more high-quality research.

Target audience

Informatics professionals who generate EHR or CDM datasets; researchers who analyze EHR or CDM datasets.

Discussion questions

- Given the inter-relatedness between the CDM, data characterization and the use/development of analytical tools, how can findings be best communicated to all stakeholders?
- As the quality of the underlying data and sophistication of the query tools increases, should efforts be made to repeat earlier studies / queries?
- Can a 'scorecard' be created that describes the state of network/data at the time of analysis?
- How can these processes be translated to run outside of a DRN, so that they can be employed at a single institution?

Statement of participation

All panelists have agreed to take part in the panel

References

1. Califf RM, Robb MA, Bindman AB, Briggs JP, Collins FS, Conway PH, Coster TS, Cunningham FE, De Lew N, DeSalvo KB, Dymek C, Dzau VJ, Fleurence RL, Frank RG, Gaziano JM, Kaufmann P, Lauer M, Marks PW, McGinnis JM, Richards C, Selby JV, Shulkin DJ, Shuren J, Slavitt AM, Smith SR, Washington BV, White PJ, Woodcock J, Woodson J, Sherman RE. Transforming Evidence Generation to Support Health and Health Care Decisions. *N Engl J Med*. 2016;375(24):2395-400. doi: 10.1056/NEJMs1610128. PubMed PMID: 27974039.
2. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. *Pharmacoepidemiol Drug Saf*. 2012;21 Suppl 1:82-9. doi: 10.1002/pds.2321. PubMed PMID: 22262596.
3. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, Raebel MA, Beaulieu NU, Rosofsky R, Woodworth TS, Brown JS. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidem Dr S*. 2012;21 Suppl 1:23-31. PubMed PMID: Medline:22262590.
4. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, Steiner JF. The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration. *Egems*. 2014;2(1):1049. doi: 10.13063/2327-9214.1049. PubMed PMID: 25848584; PMCID: PMC4371424.
5. Steiner JF, Paolino AR, Thompson EE, Larson EB. Sustaining Research Networks: the Twenty-Year Experience of the HMO Research Network. *Egems*. 2014;2(2):1067. PubMed PMID: 25848605; PMCID: PMC4371441.
6. Vogt TM, Lafata JE, Tolsma DD, Greene SM. The Role of Research in Integrated Health Care Systems: The HMO Research Network. *Perm J*. 2004;8(4):10-7. PubMed PMID: 26705313; PMCID: PMC4690686.
7. Califf RM. The Patient-Centered Outcomes Research Network: a national infrastructure for comparative effectiveness research. *N C Med J*. 2014;75(3):204-10. PubMed PMID: Medline:24830497.
8. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014. doi: 10.1136/amiajnl-2014-002747. PubMed PMID: 24821743.
9. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care*. 2010;48(6 Suppl):S45-51. Epub 2010/05/18. doi: 10.1097/MLR.0b013e3181d9919f. PubMed PMID: 20473204.
10. Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, Brown JS. Design of a national distributed health data network. *Annals of internal medicine*. 2009;151(5):341-4. Epub 2009/07/30. PubMed PMID: 19638403.
11. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care*. 2013;51(8 Suppl 3):S22-9. doi: 10.1097/MLR.0b013e31829b1e2c. PubMed PMID: 23793049; PMCID: 4306391.
12. Johnston A, Jones WS, Hernandez AF. The ADAPTABLE Trial and Aspirin Dosing in Secondary Prevention for Patients with Coronary Artery Disease. *Curr Cardiol Rep*. 2016;18(8):81. doi: 10.1007/s11886-016-0749-2. PubMed PMID: 27423939.