Research Data Network Ontologies for Precision Cancer Medicine supporting i2b2 and OMOP

W. Scott Campbell, PhD, MBA; James R. Campbell, MD; Christian Reich, MD, PhD; Rimma Belenkaya, MA, MS

1University of Nebraska Medical Center, Department of Pathology and Microbiology, Omaha, Nebraska; 2University of Nebraska Medical Center, Department of Internal Medicine, Omaha, Nebraska; 3Real World Insights, IQVIA, Cambridge, MA; 4Memorial Sloan Kettering Cancer Center, New York, NY

Abstract

Research data networks created over the past decade including OHDSI and PCORnet offer the promise of access to large data sets to investigators for hypothesis testing. The data sets based on information collected and stored in the EHR are largely managed by common data models to support query interoperability. These data models, however, are insufficient to address the precision cancer medicine use case. The introduction and extension of network ontologies based on ONC standards greatly extends the utility and functionality of these research networks to serve the cancer medicine use case.

The development, deployment and use of ONC standards into a shared ontology for cancer precision medicine research will be discussed by this panel. In particular, the integration of SNOMED CT, LOINC and RxNorm will be described along with research use cases in PCORnet and OHDSI networks and across the OMOP and i2b2 data structures. Information capture including discrete pathology and genomics data and its incorporation into these data warehouses will be reviewed.

Participants will gain insights into management and deployment of ONC ontologies and reference terminologies to support research objectives. Database requirements and operational considerations for successful use of ONC-based ontologies will be provided, as well as, example research queries.

Introduction

Research data networks developed over the past decade including PCORnet and OHDSI use EHR-based information as a primary source of data. Through the use of common data architectures and common data models, these networks provide the promise of supporting both local and network-wide research efforts. While these data networks serve laudable research uses, they are not prepared to support precision cancer medicine use cases in their current form. Precision Cancer Medicine (PCM) research requires access to data not typically recorded nor represented well in ONC data standards including anatomic and molecular pathology data and newly developed molecular therapies. Furthermore, the data networks’ use of ONC standards and their underlying concept models is limited. SNOMED CT, LOINC and RxNorm concepts are associated with data elements within these data sets and are critical to PCM research, but the nature of the relationships between the concepts represented within each standard is variable and left to each datamart manager’s discretion. This variation and the lack of domain ontologies to provide conceptual metadata ultimately limits the types of PCM questions that can be readily asked of the datamart.

To address the shortcomings of these data networks to adequately serve the Precision Cancer Medicine use case, consolidated efforts in terminology development, management and deployment are necessary. The panelists will describe the components required for successful precision cancer research in the context of specific use cases that are representative of the domain. Terminology development for use in cancer pathology data capture using SNOMED CT for surgical and molecular pathology will be presented. Management of clinical enterprise workflows to capture clinical information in discrete and encoded form is discussed. How standards are incorporated into OHDSI (OMOP-based) and PCORnet (i2b2-based) data models will be explained, and the development and
deployment of a unified reference terminology and ontology using LOINC, SNOMED CT and RxNorm will be reviewed. These three distinct building blocks of data representation and management will be used to demonstrate how Precision Cancer Medicine research can be supported by international research networks in an era of evolving clinical science. The panel will address the following topics:

**Use of standards in research data networks: OHSDI (OMOP)/PCORnet (i2b2) – C Reich; JR Campbell**

PCORnet has been struggling to support ONC terminology standards in their Common Data Model (CDM) since CDMv3. Deployment of standards in the proposal for CDMv5 is more mature and complete but PCORnet has not yet coped with the deployment of ONC metadata in their information model as is true of OMOP installations. This places the researcher in the awkward position of having to construct their own code value sets for every research question and to distribute the code sets with their queries, a process subject to error and non-reproducibility.

The OHSDI network is prescriptive in its use and application of reference terminology standards, and removes ambiguity surrounding data and its intended meaning. The OHSDI approach reduces the burden of researchers and data mart managers with regards to data management and shared queries of data from disparate sources.

**Terminology development for pathology genomics, lab medicine and therapies – JR Campbell**

A primary impediment to structured reporting of clinical and genomic data for precision medicine is limitations in the structure and content of ONC terminologies. Although LOINC has proposed new terms for some genomic observations, under specification of the LOINC metadata model, poorly defined metadata for RXNORM and inadequate integration with the concept models for SNOMED CT, coupled with incomplete content across the spectrum of Precision Cancer Medicine is insurmountable for the research informatics datamart manager. University of Nebraska has been working with colleagues in PCORnet and OHDSI to extend LOINC, SNOMED CT and RXNORM to the rigor of full ontologies with extensions to content to include precise, complete and logically consistent metadata supporting queries and analytics for the PCM researcher.

**Common concept model for unified ontology of ONC – JR Campbell**

ONC terminology standards were specified as part of a 2000 analysis by NCVHS of those reference terminologies in the US domain which were compliant with ontologic standards proposed by the NLM. Although the individual terminologies were vetted for structure and content, they were of inconsistent semantic rigor. Currently they are a patchwork designed to ‘cover’ the major semantic domains of the EHR but do so with no consistent agreement as to binding of the terminology to the EHR information model nor to the metadata definitions of the concepts themselves. In order to better support analytics for research, Nebraska has been working with NLM and LOINC to develop and publish a unified ontology for the ONC terminologies of SNOMED CT, LOINC and RXNORM.

**Managing the clinical enterprise to create structured data capture model – WS Campbell**

OHSDI and PCORNet research networks both rely heavily upon information collected and stored in the EHR. This information is extracted from the EHR, transformed to conform to a CDM and/or characterized with metadata, and loaded into these datamarts. Non-discrete data is extremely difficult to extract reliably while discrete EHR data is relatively easy to extract although many discrete EHR data elements are not represented or characterized using ONC terminology standards. Encoding these data introduces additional, complex steps in the transformation process which can be eliminated when data is captured and represented by data standards at the point of data collection. Pathology and genomics data collection and storage in the EHR is neither discrete nor encoded which creates a barrier to importing PCM data into research datamarts. Nebraska has deployed methods to capture pathology and genomics data in discrete fashion and encoded using ONC reference terminologies to facilitate PCM data into research data networks.

**Example use case – Colorectal Cancer Precision Medicine – WS Campbell; R Belenkaya**

As part of a PCORNet funded effort, Nebraska investigated the readiness of EHR data as developed and deployed in clinical research warehouses to support PCM. Specifically, investigators sought to identify all patients diagnosed with stage 4 colorectal cancer between 2013 and 2016, or who developed metastatic CRC in the same time period. Further refinement involved identifying those patients who received some form of genetic evaluation and received a
targeted therapy. Conformance of targeted therapy with FDA guidelines was then assessed. Analysis was done by data extraction from CDW’s and compared with manual chart review. Results will be reviewed including data readiness to support PCM queries and data domains in need of evaluation.

Discussion Questions
1. Why do reference terminologies need to interoperate?
2. Why have reference terminologies not worked together before?
3. How can a unified reference terminology be distributed and maintained?
4. Can industry be engaged to facilitate adoption of unified standards?
5. How do the terminologies and common data models interact?

Statement of participation
All participants agree to participate in this panel presentation and discussion.

References


