EHR Data Methodologies in Clinical Research: Perspectives from the Field

Session 1: Semantic Harmonization; Definition; Content; Ontologies
Common Data Models for Sharing EHR data across Settings

Michael G. Kahn MD, PhD
Professor, Pediatric Epidemiology
University of Colorado
Michael.Kahn@ucdenver.edu

11 December 2014

Funding by: PCORI Contract ME-1303-5581 (Kahn), NCATS UL1TR001082 (Sokol), AHRQ R01HS022956 (Schilling)
Disclosures

Presentation based on EDM Forum commissioned paper:

**Clinical Informatics**

Data Model Considerations for Clinical Effectiveness Researchers

Michael G. Kahn, MD, PhD,* † ‡ Deborah Batson, BS, † and Lisa M. Schilling, MD, MSPH§

**Introduction:** Growing adoption of electronic health records and increased emphasis on the reuse and integration of clinical care and administration data require a robust informatics infrastructure to inform health care effectiveness in real-world settings. The Scalable

**Key Words:** data models, databases, Comparative Effectiveness Research

*(Med Care 2012;50: S60–S67)*

Medical Care 2012:50: S60-67 (open access)
A common data model is critical!

- EHR-1
  - Local Data Warehouse
    - Limited Data Set
    - Common Data Model
    - Common Terminology

- Other
  - EHR-2
    - Local Data Warehouse
      - Limited Data Set
      - Common Data Model
      - Common Terminology

- Other
  - EHR-3
    - Clinical Registries
      - Limited Data Set
      - Common Data Model
      - Common Terminology

Common Query
What is a data model & why should I care?

• A data model determines:
  – What data elements can be stored
  – What relationships between data can be represented
  – Technical stuff: data type, allowed ranges, required versus optional (missingness)

• You should care because it determines:
  – How easy can data be recorded, extracted & queried
  – Contributes to data quality
Visit-centric versus Patient-centric
Query: “For various age groups, how many medications were filled?”
Query: “For various age groups, what is the average number of prescriptions per visit?”

Two-table join

Three-table join + Date comparisons
SAFTINet Asthma Cohort Definition

• Adults (ages 18 and over) as of Jan 1, 2009 receiving care in selected sites who:
  – Have had at least 2 visits separated by at least 30 days coded as 493.xx in the 18 months prior to July 1, 2011, OR
  – A single diagnosis of 493.xx AND two filled prescriptions for an asthma maintenance medication separated by at least 30 days in the past 12 months.

• Exclusion criteria: Patients with other concomitant chronic lung disease
  – Cystic fibrosis
  – COPD, emphysema, chronic bronchitis
  – Alpha-1-antitrypsin deficiency
  – Pulmonary fibrosis
  – Active TB
Key questions for a data model

- From Jeff Brown regarding FDA Sentinel Initiative*:
  1. What does the system need to do?
  2. What data are needed to meet system needs?
  3. Where will the data be stored?
  4. Where will the data be analyzed?
  5. Is a common data model needed, and if so, what will the model look like?

Eight dimensions of data models

Modified from Moody and Shanks*

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Original definition</th>
<th>Recasted definition for CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Completeness</td>
<td>Does the data model contain all user requirements?</td>
<td>Can the data model store and retrieve data to meet investigator CER needs?</td>
</tr>
<tr>
<td>2. Integrity</td>
<td>Does the data model conform to the business rules and processes to guarantee data integrity and enforce policies?</td>
<td>Does the data model enforce meaningful data relationships and constraints that uphold the intent of the data’s original purpose, i.e., clinical care, billing?</td>
</tr>
<tr>
<td>3. Flexibility</td>
<td>Does the data model deal with changes in business and/or regulatory change?</td>
<td>Can new data elements and relationships be added if project scope or if regulatory rules (e.g., patient identification) changes?</td>
</tr>
<tr>
<td>4. Understandability</td>
<td>Are the concepts and structures in the data model easily understood?</td>
<td>Do the concepts, structures and relationships make sense to investigators, data managers, and statisticians?</td>
</tr>
</tbody>
</table>

## Eight dimensions of data models

Modified from Moody and Shanks*

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Original definition</th>
<th>Recasted definition for CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Correctness</td>
<td>Does the data model conform to the rules of the data modeling technique?</td>
<td>Does the model conform to good data modeling practices such as limited data storage redundancy?</td>
</tr>
<tr>
<td>6. Simplicity</td>
<td>Does the data model contain the minimum possible entities and relationships?</td>
<td>Are concepts represented as straightforwardly as possible? Are all data element necessary?</td>
</tr>
<tr>
<td>7. Integration</td>
<td>Is the data model consistent with the rest of the organization’s data?</td>
<td>Do all of the various data domains, such as demographics, observations, labs and medications “hang together” in a consistent and logical fashion?</td>
</tr>
<tr>
<td>8. Implementability</td>
<td>Can the data model be implemented within existing time, budget, and technology constraints?</td>
<td>Can the data model be implemented and maintained by current and future partners given anticipated budgets, time, and technical constraints?</td>
</tr>
</tbody>
</table>

## Major common data models

<table>
<thead>
<tr>
<th>Name</th>
<th>Developing entity</th>
<th>Initial Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observational Medical Outcomes Project</strong> (OMOP)</td>
<td>Foundation of the NIH, now Reagan-Udall Foundation</td>
<td>Comparative Drug Outcomes Studies</td>
</tr>
<tr>
<td>i2b2</td>
<td>Partners Healthcare</td>
<td>Informatics framework for clinical and biological data integration. Widely used across NCATS CTSA</td>
</tr>
<tr>
<td>HMORN Virtual Data Warehouse (VDW)</td>
<td>HMO Research Network</td>
<td>Distributed data warehouse to allow comparative studies across collaborating sites: HMORN, CRN, Oregon CTRI</td>
</tr>
<tr>
<td>Mini-Sentinel</td>
<td>FDA</td>
<td>Derivative of VDW focused on large-scale drug surveillance</td>
</tr>
<tr>
<td><strong>PCORnet</strong></td>
<td>PCORI</td>
<td>Derivative of Mini-Sentinel focused on PCOR research</td>
</tr>
</tbody>
</table>
A common data model is critical!
Public domain tools to help “Cross the CER Chasm”

- Data profiling with OHDSI White Rabbit

Public domain tools to help “Cross the CER Chasm”

- Data profiling with OHDSI White Rabbit
- Data transformation documentation with OHDSI Rabbit in a Hat

Ensuring Data Consistency/Comparability

ETL Conventions for use with PEDSnet CDM V1.0

Revision Date: October 29, 2014

Version Tracking

<table>
<thead>
<tr>
<th>Date</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/20/2014</td>
<td>Version 1: Includes observation period, uses CHOP's codes for height/weight</td>
</tr>
<tr>
<td>10/20/2014</td>
<td>Version 1a: Readjusted logic for SBP/DBP and position. incorporated changes from TO</td>
</tr>
<tr>
<td>10/22/2014</td>
<td>Incorporated (most of) changes from CB, GS, and KM</td>
</tr>
<tr>
<td>10/24/2014</td>
<td>Post ETL call changes; added CF's future elements list</td>
</tr>
<tr>
<td>10/27/2014</td>
<td>Incorporated edits from JB</td>
</tr>
<tr>
<td>10/28/2014</td>
<td>Removed BMI from observations -- all sites can provide source height and weight. DCC will calculate and insert BMI observations using one common algorithm</td>
</tr>
<tr>
<td></td>
<td>Replaced vocabulary 99 with identical and additional concept_ids from new vocabulary 60, which includes new concept_id codes for &quot;unknown&quot; and &quot;no information.&quot;</td>
</tr>
<tr>
<td></td>
<td>Untangled confusion regarding which fields can be/should be obfuscated (should be obfuscate key source values rather than ID fields)</td>
</tr>
<tr>
<td>10/29/2014</td>
<td>Added suggestion that *_id field could be used as obfuscated value of source value as long as mapping between ID field and real source value is maintained by site.</td>
</tr>
<tr>
<td></td>
<td>Added text that biobank flag and chart availability can be entered into observations table more than once if patient changes consent status over time.</td>
</tr>
<tr>
<td></td>
<td>Clarified that discharge disposition and discharge status observations occur once per visit_occurrence.</td>
</tr>
</tbody>
</table>

The PEDSnet Common Data Model is an evolving specification, based in structure on the OMOP Common Data Model, but expanded to accommodate requirements of both the PCORnet Common Data Model and the primary research cohorts established in PEDSnet.

Version 1 of the PEDSnet CDM reflects the ETL processes developed during the first six months of network development. As such, it closely follows version 1 of the PCORnet CDM. We anticipate that the PEDSnet CDM will expand to include additional data domains such as medication usage, laboratory testing, and other types of clinical observations. However, in order to minimize discordance with the PCORnet CDM, specification of these domains has been deferred until the cognate portion of the PCORnet CDM is developed, or active research in one of the PEDSnet cohorts requires those data types.

This document provides the ETL processing assumptions and conventions developed by the PEDSnet data partners that should be used by a data partner for ensuring common ETL business rules. This document will be modified as new situations are identified, incorrect business rules are identified and replaced, as new analytic use cases impose new/different ETL rules, and as the PEDSnet CDM continues to evolve.
EHR Data Methodologies in Clinical Research: Perspectives from the Field

Session 1: Semantic Harmonization; Definition; Content; Ontologies
Common Data Models for Sharing EHR data across Settings

Michael G. Kahn MD, PhD
Professor, Pediatric Epidemiology
University of Colorado
Michael.Kahn@ucdenver.edu

11 December 2014

Funding by: PCORI Contract ME-1303-5581 (Kahn), NCATS UL1TR001082 (Sokol), AHRQ R01HS022956 (Schilling)