Data Harmonization and Synergies

Contrasting the Approaches of UC-ReX, OMOP, PCORnet and NCATS ACT

Douglas S. Bell, MD, PhD
Professor of Medicine; Director, UCLA CTSI Biomedical Informatics Program
Data Harmonization

• **Goal:** Re-use clinical data across organizations

• **Objective:** Choose a representation that
  – Validly represents meaning of the source data
  – Enables accurate aggregation and comparison
  – For specific purposes

• **Approaches**
  – i2b2/SHRINE – Cohort discovery
  – OMOP – Drug Adverse Event Surveillance
  – PCORnet – Comparative Effectiveness
I2b2 Data Model

290-319 Mental, Behavioral and Neurodevelopmental disorders
- 290-299 Psychoses
- 290 Pervasive developmental disorders
  - 290.0 Autistic disorder

F01-F99 Mental, Behavioral and Neurodevelopmental disorders
- F80-F89 Pervasive and specific developmental disorders
  - F84 Pervasive developmental disorders
    - F84.0 Autistic disorder

290.0 F84.0
Network Data Alignment Lifecycle

Data source
- UC-REX
- PCORI grants
- Investigators
- Repeatable site assessment plan

Requirements

Analyses
- Data Harmonization
- Project requirements alignment

Executive Committee
- Approves analysis findings & recommendations

Policy alignment

Data Quality
- Technical Implementation

Analysis

Deployment
- Technical Implementation

Curation
- Data Extract, Transform & Load
- Mapping to standards
Observational Medical Outcomes Partnership
Common Data Model

Michael E. Matheny, MD, MS, MPH
VA Tennessee Valley Healthcare System
Vanderbilt University
Key concepts within OMOP CDM v4

Person-centric (not encounter based)

Delineate between verbatim data and inferred information (ex. standardize process to define episode of care and length of drug exposure)

Hybrid between Entity-relationship (ER) model and Entity-attribute-value (EAV) model

Preserve source values (ex. ICD-9-CM) and established standard concepts (ex. SNOMED)

Multiple domains, each with specific attributes: CONDITION, DRUG, PROCEDURE, VISIT, DEATH

One table can maintain multiple sources of data by distinguishing ‘type’ of condition (ex. diagnosis from medical claim vs. problem list from EHR)

Source: Patrick Ryan OMOP Presentation 10/26/2013 “Harmonizing Methods and Data across Multiple Databases”
• Brings multiple sources of data for a particular domain into one content-centric table
  – Bar coded medication administration
  – Computerized provider order entry for drug
  – Outpatient pharmacy fill records
  – Administrative codes (HCPCS, CPT, ICD-9) for infusion and drug administration
• Uses UMLS and custom mappings to transform large set of standardized vocabularies into core set of standard terminologies
  – Examples:
    • Condition Standard: SNOMED
    • Lab Standard: LOINC
    • Drug Standard: RxNorm
pSCANNER Data Harmonization

Daniella Meeker, PhD

University of Southern California,
RAND Corporation
1. Stage raw data elements and terminologies into target data model ("raw" data fields)

2. Apply NLM standard-to-standard terminology maps to populate a ("standard" data fields)

3. Fill in gaps in the terminology mapping
   – Prioritize by use-case
   – Retain and resolve new maps

4. Quality Assessment and monitoring
   – Monitor new terminologies
   – Update new maps
## 1. The v1.0 Data Domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Relational Framework</th>
<th>Analytic Use and Importance</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEMOGRAPHIC</td>
<td>Demographic</td>
<td>Foundational</td>
<td>Central table for patient identity.</td>
</tr>
<tr>
<td>ENROLLMENT</td>
<td>Demographic</td>
<td>Foundational</td>
<td>Contains records for continuous enrollment periods. The enrollment basis is expected to be determined by 4 potential methods: Insurance, Geography, Algorithmic, or Encounter-based.</td>
</tr>
<tr>
<td>ENCOUNTER</td>
<td>Encounter</td>
<td>Foundational</td>
<td>The ENCOUNTER table should include information on interactions between patients and providers.</td>
</tr>
<tr>
<td>DIAGNOSIS</td>
<td>Encounter</td>
<td>Foundational</td>
<td>ENCOUNTER-based diagnosis codes. Data in this table are expected to be from healthcare-mediated processes and reimbursement drivers (can include both technical/facility billing and professional billing)</td>
</tr>
<tr>
<td>PROCEDURE</td>
<td>Encounter</td>
<td>Foundational</td>
<td>ENCOUNTER-based procedure codes. Data in this table are expected to be from healthcare-mediated processes and reimbursement drivers (can include both technical/facility billing and professional billing)</td>
</tr>
<tr>
<td>VITAL</td>
<td>Encounter basis for healthcare-based measures / patient basis for patient self-report</td>
<td>Foundational</td>
<td>Vital sign measurements from both healthcare and non-healthcare settings. Does not include direct feeds from devices.</td>
</tr>
<tr>
<td>Domain</td>
<td>Relational Framework</td>
<td>Analytic Use and Importance</td>
<td>Comments</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------</td>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Dispensin...</td>
<td>Demographic</td>
<td>Identifying exposures</td>
<td>Straightforward context, but although this domain is common in claims data, outpatient dispensing may not be available in many EHR data sources. An example of this domain is prescription fills through a neighborhood pharmacy. This is not the same as inpatient medication administration (eg, injection performed by a nurse).</td>
</tr>
<tr>
<td>Death</td>
<td>Demographic</td>
<td>The ultimate endpoint</td>
<td>Dependent upon larger PCORnet comprehensive death efforts (eg, NDI, study workflows)</td>
</tr>
<tr>
<td>Cause of Death</td>
<td>Demographic</td>
<td>The ultimate endpoint</td>
<td>Dependent upon larger PCORnet comprehensive death efforts (eg, NDI, study workflows)</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------</td>
<td>-----------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Laborator y result</td>
<td>Encounter result</td>
<td>Patient identification and outcomes analyses (e.g., identify patients with uncontrolled diabetes based on HbA1c)</td>
<td>A very challenging area; approach might involve a multi-stage approach with full standardization and harmonization across sites undertaken for a limited number of labs pertinent to use cases. Daniella Meeker’s summary: There is value in standardizing the way the raw data is organized in a staging table if analysts wish to explore that information for feasibility interests. Although MSCDM does not have ENCOUNTERID in the CDM documentation, Jeff confirms that this is an oversight.</td>
</tr>
<tr>
<td>State Vaccine</td>
<td>Demographic</td>
<td>Immunization exposures</td>
<td>For Mini-Sentinel, some states have electronic immunization registries maintained by state public health, which supports better coverage for immunization exposures. This could potentially be supplemented by EHR data domains.</td>
</tr>
</tbody>
</table>

| | | | HIGH EFFORT (dependent upon other workflows) | HIGH UTILITY (centralized data sources, such as NDI, would improve generalizability) | LOW DIFFICULTY |
| | | | HIGH EFFORT | MOD EFFORT | HIGH UTILITY | HIGH DIFFICULTY |

Recommended v3.0 due to dependencies

Recommended v2.0, but focused on a targeted set of labs of particular value

Later stage
<table>
<thead>
<tr>
<th>Domain</th>
<th>Relational Framework</th>
<th>Analytic Use and Importance</th>
<th>Comments</th>
<th>Effort Needed to Acquire Data</th>
<th>Analytic Utility/Value</th>
<th>Ability to Standardize Across Sites</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient-reported outcomes (PROs) reported via a PRO survey instrument</td>
<td>Patient or encounter</td>
<td>Emphasis for PCORI focus on patient-centric experiences</td>
<td>Current effort led by Reesa Laws, Matthew Wyatt, Jason Doctor, and the PRO task force working group. The analytic use is not yet well-characterized.</td>
<td>MOD EFFORT (more likely to implemented by PPRNs, rather than CDRNs)</td>
<td>MOD UTILITY (use cases have not been highly detailed)</td>
<td>HIGH DIFFICULTY (however, if the PRO TF chooses to adopt PROMIS/LOINC, this would be a different situation)</td>
<td>Proposed v2.0, but dependent upon PRO TF work</td>
</tr>
<tr>
<td>Problem list</td>
<td>Patient basis</td>
<td>Potential use for cohort identification</td>
<td>Problem lists are different from the DIAGNOSIS table because they are not explicitly associated with a per-encounter basis. Felt to be particularly valuable for chronic conditions that might not always be reflected in encounter-level diagnosis associations. Dates associated with problem lists (such as onset date) are generally felt to be non-rigorous.</td>
<td>LOW EFFORT (domain is expected to be codified)</td>
<td>MOD UTILITY (value appears to exist with cohort identification, but not clear for analytic purposes)</td>
<td>MOD DIFFICULTY (concepts appear to be consistent across sites, but date accuracy may be dependent upon site-specific workflows)</td>
<td>Undetermined recommendation for v2.0 vs. v3.0</td>
</tr>
<tr>
<td>Medications ordered</td>
<td>Encounter</td>
<td>Identifying exposures</td>
<td>A moderately challenging area; we should assess whether the scope should expand to all orders, not just medications</td>
<td>MOD EFFORT</td>
<td>HIGH UTILITY</td>
<td>MOD DIFFICULTY</td>
<td>Because of difficulty, a cautious recommendation for v2.0</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------</td>
<td>-----------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-----------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------------------------------------------------------</td>
</tr>
<tr>
<td>Medications administered</td>
<td>Encounter</td>
<td>Identifying exposures</td>
<td>Generally an inpatient basis with nursing workflow</td>
<td>MOD EFFORT</td>
<td>LOW UTILITY (limited coverage because of inpatient basis)</td>
<td>MOD DIFFICULTY</td>
<td>Later stage</td>
</tr>
<tr>
<td>Patient medication reconciliation</td>
<td>Encounter</td>
<td>Analytic use has not been strongly articulated</td>
<td>Sometimes referenced as the “active med list”</td>
<td>MOD EFFORT</td>
<td>MOD UTILITY (because this is reported, the veracity of data is more uncertain)</td>
<td>MOD DIFFICULT (becoming more common because of MU requirement)</td>
<td>Later stage</td>
</tr>
<tr>
<td>Provider Orders</td>
<td>Encounter</td>
<td>Patient identification and outcomes</td>
<td>To be assessed: What are the logical subdivisions in this area? Possibly medication orders, test orders. Are referrals considered to be orders? Is this a strictly EHR domain, or is there an equivalent from claims data? Feedback from Aaron Sorensen: If we decide to do labs, then it would probably make sense to allow the collection of results for certain diagnostic procedures (e.g., pulmonary function test [CPT code 94010], six-minute walk test [CPT code 94620], etc.). In most Epic implementations, procedure and lab results are stored in the same reporting table.</td>
<td>MOD EFFORT</td>
<td>MOD UTILITY</td>
<td>MOD DIFFICULTY</td>
<td>v3.0, unless working group determines a high synergy with other areas</td>
</tr>
</tbody>
</table>
| Contraindications and/or Allergies | Demographics | Quality of care; (eg, with aspirin use case, exclude patients with a contraindication to aspirin) | Contraindications are assumed to be medication/treatment focused, while allergies might be more unrelated to treatment (eg, trees, cats).

Also, “contraindications” would cover cases where two medications should not be taken simultaneously, whether or not the patient has allergies to those meds.

Jon Puro’s comments: generally means any case where a medication, procedure, or other medical process should not be used for a given patient, not just where medications might interact; most common in the area of medications and adverse drug interactions.

Daniella Meeker’s comments: unrelated to treatment (eg, trees, cats).

This concept may also involve alerts. | HIGH EFFORT (likely to be many different methods of data capture) | MOD UTILITY (has not been incorporated in Mini-Sentinel use, but may be consideration for PCORnet observational use cases) | HIGH DIFFICULTY (due to expected heterogeneity of data capture) | Later stage |
Cross-CDRN Activities

• DSSNI CDM working group
• DSSNI Patient Generated Data CDM work group
• Cross-site ETL collaboration infrastructure
  – Vendor-to-CDM (e.g. Clarity-to-PCORNet)
  – CDM-to-CDM (e.g. OMOP to PCORNet)
Data Harmonization Highlights

Davera Gabriel, RN
“The Intrepid Semanticist”
Research Informatics
University of California, Davis
UCReX Data Harmonization
Guiding Principles

1) Collaborative Governance
2) Iterative Process
3) Understand Requirements & Source Data
   - Data profiling of source instances always a part of the analysis
   - Solution to fit the need
   - Only code what needs to be
4) Manage Scope & Expectations
   - Start simple
   - Eliminate low volume transactions, “noise”
   - Scope “down” to meet resource challenges
5) Conflate with Care
   - Document assumptions and results
6) Don’t Create Data
   - Example: null handling
Global Considerations

- Flavors of null
- Multiple values per pt, per encounter ...
- Inclusion / exclusion of source data
- Date / time data types
- Implementation variation
  - Clarity tables
  - Encounter type
- Multiple source systems
Sex

- HL7: Administrative Gender
- Transgender folks
- True hermaphrodites
Age

- Date / Time format
- Rounding to whole year
- Pediatric cases
- Obfuscating the elderly
- Single years vs ranges
- Age at encounter
Race & Ethnicity

• Combined vs 2 question format: not equal
• Multiple Race designations
• Flavors of null
  – Target format for “non-Hispanic”
Lab Results

• Most transaction volumes are in a small # of tests
• Units of measure must be normalized
• LOINC is very specific
  – Event sans method axis
• Textual results
  – Misspellings
  – Requires development of normal value sets
• Microbiology is a different animal: requires special focus
Diagnoses

• Divergent Dx types
  – Discharge
  – Admitting (sometimes Reason for Visit)
  – Problem List
  – Billing: Primary vs Secondary...

• ICD-9-CM
  – Big buckets
  – No concept persistence
  – IMO implementations: NOT really I9

• ICD-10
  – Implementation solution: maps (per S. Murphy)
Medications

• Proprietary content: FDB, Multum, MediSpan...
  – Combination drugs
  – Dose forms
  – Administration strength, overall dose

• Utilize NDC 11 digit codes

• Maps to RxNorm
  – SCD, SCDG, SBD, GPCK, BPCK term types