Seminal OHDSI Vocabulary Paper

OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization
OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization

Christian Reich, MD, Anna Ostropolets, PhD, Patrick Ryan, PhD, Peter Rijnbeek, PhD, Martijn Schuemie, PhD, Alexander Davydov, MD, Dmitry Dymshyts, MD, and George Hripcsak, MD
Observational research needs large scale for sample size and diverse populations.
That needs standardization, all major data networks (Sentinel, PCORNet, OHDSI) adopted a standard.
Standardization of format and representation (vocabularies)
Standardization of medical vocabularies can be done ad-hoc or through a central reference.
OHDSI chose the central system.
Supports
- Cohort definition
- Covariate construction
- Large-scale analytics
- Result reporting
UMLS is such a repository of vocabularies, but not focused on our use cases.
## Requirements

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard concepts</td>
<td>Unique concepts of fully pre-coordinated medical entities, to be stated as fact, no negations of facts, no reference to the past, and no flavors of null (unknown, not reported, etc.)</td>
</tr>
<tr>
<td>Concept domains</td>
<td>Assignment of concepts to domain categories (condition, drug, visit, etc.)</td>
</tr>
<tr>
<td>Comprehensive coverage</td>
<td>In each domain, standard concepts must cover all possible entities and mappings from terms and codes used in databases around the world</td>
</tr>
<tr>
<td>Polyhierarchies</td>
<td>Precalculated hierarchies organizing concepts</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Computationally efficient data model</td>
</tr>
<tr>
<td>Use case focus</td>
<td>Storing and analyzing patient-level data for evidence generation</td>
</tr>
</tbody>
</table>
Methods

• Vocabularies and concepts
• Domains
• Standardization of concepts
• Mapping, hierarchical, and other relationships between concepts
• Life cycle and distribution
• Quality assurance
Structure
Overall content

• 136 vocabularies
  – 101 from external sources

• 10,574,359 concepts
  – 8,761,976 valid ones
    • 40.5% standard ones
    • 50.1% non-standard ones
    • 9.4% classification (mostly Drug and Measurement)

• 28 million valid relationships
  – 38.3% Is a
  – 14.1% Maps to (covering 66.8% of non-standard concepts)
### Equivalence relationships per type and domain

<table>
<thead>
<tr>
<th>Type of “Maps to” relationship, % (n)</th>
<th>Domain</th>
<th>One-to-one</th>
<th>Many-to-one</th>
<th>One-to-many</th>
<th>Many-to-many</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>1.9% (54 671)</td>
<td>10.3% (292 507)</td>
<td>&lt;0.1% (38)</td>
<td>3.2% (90 034)</td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td>6.1% (172 216)</td>
<td>3.6% (101 774)</td>
<td>&lt;0.1% (4)</td>
<td>&lt;0.1% (10)</td>
<td></td>
</tr>
<tr>
<td>Drug</td>
<td>18.1% (515 360)</td>
<td>42.6% (1 208 579)</td>
<td>&lt;0.1% (181)</td>
<td>1.6% (44 780)</td>
<td></td>
</tr>
<tr>
<td>Measurement</td>
<td>0.5% (14 502)</td>
<td>0.4% (11 580)</td>
<td>1.2% (33 655)</td>
<td>1.2% (33 489)</td>
<td></td>
</tr>
<tr>
<td>Observation</td>
<td>3% (86 014)</td>
<td>2.4% (69 458)</td>
<td>&lt;0.1% (3)</td>
<td>0.2% (45 79)</td>
<td></td>
</tr>
<tr>
<td>Procedure</td>
<td>1.1% (29 973)</td>
<td>2.5% (70 974)</td>
<td>&lt;0.1% (16)</td>
<td>0.2% (55 81)</td>
<td></td>
</tr>
</tbody>
</table>
Discussion

• Work in progress, will never end
  – Errors need constantly be addressed
  – Workgroups are taking care of the system
  – Standard concepts particularly challenging – UMLS helps but not enough
  – Quality system being built

• Not a Knowledge Base
  – Lateral relationships only adopted “lazily”