eMERGE Data Dictionary
Harmonization and Best Practices for Standardized Phenotype Data Representation

Jyoti Pathak

April 13th, 2010
Acknowledgment

• Chris Chute
• Dan Masys
• Janey Wang
• Sudha Kashyap
• Melissa Basford
Overall Objective

• Without terminology standards:
  • Health data is non-comparable
  • Health systems cannot meaningfully interoperate
  • Secondary uses of data for research and applications (e.g., clinical decision support) is not possible

• **Our goal**: Standardized and consistent representation of eMERGE network-wide phenotype data submitted to dbGaP
Standardized Resources

• NCI caDSR (Cancer Data Standards Repository)
• CDISC SDTM (Study Data Tabulation Model)
• NCI Thesaurus
• SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms)
Background: Clinical Terminology Standards and Resources

• NCI Cancer Data Standards Repository
  • Metadata registry based on ISO/IEC 11179 standard for storing common data elements (CDEs)
  • Allows creating, editing, deploying, and finding of CDEs
  • Provides the backbone for NCI’s semantic-computing environment, including caBIG (Cancer Biomedical Informatics Grid)
  • Approx. 40,000 CDEs
Background: Clinical Terminology Standards and Resources

• CDISC Terminology
  • To define and support terminology needs of the CDISC models across the clinical trial continuum
  • Used as part of the Study Data Tabulation Model: an international standard for clinical research data, approved by the FDA as a standard electronic submission format
  • Comprises approx. 2300 terms covering demographics, interventions, findings, events, trial design, units, frequency, and ECG terminology
Background: Clinical Terminology Standards and Resources

• NCI Thesaurus
  • Reference terminology for clinical care, translational and basic cancer research
  • Comprises approx. 70,000 concepts representing information for nearly 10,000 cancers and related diseases
  • NCI Enterprise Vocabulary Services (LexEVS) provides the terminology infrastructure for caBIG, NCBO etc.
Background: Clinical Terminology Standards and Resources

• SNOMED-CT
  • Systematized Nomenclature of Medicine Clinical Terms is a comprehensive terminology covering most areas of clinical information including diseases, findings, procedures, microorganisms, pharmaceuticals etc.
  • Comprises approx. 370,000 concepts
  • Acquired by International Health Terminology Standards Development Organization (IHTSDO) in 2007
Methods: DD Harmonization

- Collected the latest data dictionaries from all the eMERGE sites
- Preliminary cleaning
  - Uniform variable names (e.g., BMI vs. Body_Mass_Index)
  - Added new variables as required (e.g., Observation_Age)
- Manual mapping of DD variables and permissible values
  - String search
  - Pre-existing mappings
Results: Preliminary Mapping (11/2009)

Group Health: 32 Total CDEs, 17 CDEs Harmonized
Marshfield: 42 Total CDEs, 19 CDEs Harmonized
Mayo Clinic: 27 Total CDEs, 23 CDEs Harmonized
Northwestern: 26 Total CDEs, 22 CDEs Harmonized
Vanderbilt: 28 Total CDEs, 23 CDEs Harmonized
Observations

• One size does not fit all; coverage is not uniform across all the standards

• High degree of overlap for commonly used enumerated variables (e.g., Race)

• Additional curation is required to improve coverage for eMERGE data elements
  • Communicated with the caDSR and NCI Thesaurus teams
  • 54 new eMERGE-specific data elements in caDSR were created during 11/2009 and 03/2010
    • Released under “draft” status
  • 9 new NCI Thesaurus concepts
54 New eMERGE CDEs
9 new NCI Thesaurus Concepts

- Ankle-Brachial Index (ABI) (C87304)
- Decade (C87556)
- Current Procedural Terminology (C87308)
- Cognitive Abilities Screening Instrument (C87307)
- Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (C86966)
- Diagnostic and Statistical Manual of Mental Disorders, 3rd Edition (C86967)
- Fulfill (87531); Synonym “meets”
- NINCDS-ADRDA Criteria for Alzheimer's Disease (86983)
- Quartile (C87306)
Results: Revised Mapping (04/2010)

<table>
<thead>
<tr>
<th>eMERGE Network Sites</th>
<th>Total CDEs</th>
<th>CDEs Harmonized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Health</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Marshfield</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>Northwestern</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>28</td>
<td>28</td>
</tr>
</tbody>
</table>
eleMAP Data Harmonization Tool

- Demo. http://www.gwas.net/eleMAP
Discussion: How to Proceed/Next Steps

• Is caDSR our “default” mapping?
  • Almost 100% mapping to caDSR CDEs

• Implication to dbGAP submissions
  • Individual sites going to “re-map” their phenotype data based on harmonized data dictionaries?

• Feedback on eleMAP implementation
  • Thanks to Luke from Marshfield!

• Publications
  • AMIA poster on eleMAP submitted
  • JAMIA manuscripts: (1) harmonization [draft] (2) SNOMED post-coordination [under preparation]

• Collaboration with PhenX
Thank You!

Q & A
Results: NCI Thesaurus

[Bar chart showing the total number of eMERGE CDEs for different sites, with categories for Total CDEs, Proper Match, Partial Match, and No Match.]
Results: SDTM Terminology

The chart illustrates the total number of eMERGE CDEs across various eMERGE Network Sites, categorized by Total CDEs, Proper Match, Partial Match, and No Match.

- **Group Health**: Total CDEs: 32, Proper Match: 6, Partial Match: 0, No Match: 0
- **Marshfield Clinic**: Total CDEs: 44, Proper Match: 5, Partial Match: 0, No Match: 0
- **Mayo Clinic**: Total CDEs: 39, Proper Match: 12, Partial Match: 15, No Match: 0
- **Northwestern**: Total CDEs: 27, Proper Match: 9, Partial Match: 17, No Match: 0
- **Vanderbilt**: Total CDEs: 26, Proper Match: 14, Partial Match: 14, No Match: 0
Results: SNOMED (pre-coordination)