Relationships among Biomedical Ontologies and Classifications: The ICD11 Use-case

Christopher G. Chute, MD DrPH
Professor Biomedical Informatics
Vice-Chair, Data Governance
Mayo Clinic College of Medicine

Chair, International Classification of Disease Revision
Chair, ISO Technical Committee 215 on Health Informatics

OntologySummit2011: panel session-9 "Grand Challenges"
Webcast, 24 March, 2011
From Practice-based Evidence to Evidence-based Practice

Data

Clinical Databases

Registries et al.

Inference

Shared Semantics

Medical Knowledge

Standards

Vocabularies & Terminologies

Expert Systems

Decision support

Clinical Guidelines

Knowledge Management
Whither Phenotype?
Spans spectrum from enzymes to disease

• Pharmacogenomics – enzyme functionality
• Physiologist – cellular function
• Systems biologist – pathway circuit flow
• Sub-specialist – organ functioning
• Patient/Clinician – disease manifestation
• Public Health – population characteristics

Highly specific to use-case context
ICD11 Use Cases

• Scientific consensus of *clinical phenotype*

• Public Health Surveillance
  • Mortality
  • Public Health Morbidity

• Clinical data aggregation
  • Metrics of clinical activity
  • Quality management
    • Patient Safety
  • Financial administration
    • Case mix
    • Resource allocation
Traditional Hierarchical System
ICD-10 and family
Addition of structured attributes to concepts

**Concept name**
- Definition
  - Language translations
- Preferred string
  - Language translations
- Synonyms
  - Language translations
- Index Terms
THE CONTENT MODEL

Any Category in ICD is represented by:

1. ICD Concept Title
   1.1. Fully Specified Name
   1.2. Preferred Name
   1.3. Synonyms

2. Classification Properties
   2.1. Parents
   2.2. Type
   2.3. Use and Linearization(s)

3. Textual Definition(s)

4. Terms
   4.1. Base Index Terms
   4.2. Inclusion Terms
   4.3. Exclusions

5. Body System/Structure
   5.1. Body System(s)
   5.2. Body Part(s) [Anatomical Site(s)]
   5.3. Morphological Properties

6. Manifestation Properties
   6.1. Signs & Symptoms
   6.2. Investigation findings

7. Causal Properties
   7.1. Etiology Type
   7.2. Causal Properties - Agents
   7.3. Causal Properties - Causal Mechanisms
   7.4. Genomic Linkages
   7.5. Risk Factors

8. Temporal Properties
   8.1. Age of Occurrence & Occurrence Frequency
   8.2. Development Course/Stage

9. Severity of Subtypes Properties

10. Functioning Properties
    10.1. Impact on Activities and Participation
    10.2. Contextual factors
    10.3. Body functions

11. Specific Condition Properties
    11.1. Biological Sex
    11.2. Life-Cycle Properties

12. Treatment Properties

13. Diagnostic Criteria
Lung Cancer

- **has_anatomy**: Lung (all lobes), Bronchus
- **has_histology**: Adenocarcinoma, squamous, large cell, small cell, bronchioloalveolar
- **has_exclusion**: mesothelioma
- **has_symptom**: cough, hemoptysis, wheeze and stridor, dyspnea, and obstructive pneumonitis
- **has_finding**: pleural effusion; radiologic nodules, cavitation destruction of tissue or bone
Addition of semantic arcs - Ontology

Relationships
- Logical Definitions
- Etiology
- Genomic
- Location
  - Laterality
- Histology
- Severity
- Acuity
Serialization of “the cloud”

Algorithmic Derivation
High Level Structure – Core Model

Terms
Definitions
Comments

ICD11 Entity

Reference Ontologies

Category Layer

Linearization

World Health Organization; ICD-11
Linear views may serve multiple use-cases

Morbidity, Mortality, Quality, ...
ICD Collaborative Authoring Tool

Details for LB4 Immunobullous diseases of the skin

<table>
<thead>
<tr>
<th>Title &amp; Definition</th>
<th>Classification Properties</th>
<th>Terms</th>
<th>Clinical Description</th>
<th>Manifestation Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causal Properties</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporal Properties</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severity Properties</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functioning Properties</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specific Condition Properties</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnostic Criteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICD 10 Notes and Hints</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ICD 10 Code

Sorting label: LB4

ICD Title
Immunobullous diseases of the skin

Fully Specified Name

Short Definition
A group of disorders characterized by the presence of circulating auto-antibodies directed against specific skin or mucous membrane antigens and resulting in blisters or erosions.

Detailed Definition

Add new value
Wise Crowd-Sourcing
Wiki science?

• Acknowledge that the “full” ICD11 exceeds our resources
• Engage a hierarchy of expert talent to “wiki” proposals
  • Fill in details of models, extend depth
• Pre-process to manageable input
• Use TAGs as editorial boards
• Anticipate prototypes by fall 2011
Relationship with IHTSDO

SNOMED content

• IHT (SNOMED) will require high-level nodes that aggregate more granular data
  • Use-cases include mutually exclusive, exhaustive,…
  • Sounds a lot like ICD

• ICD-11 will require lower level terminology for value sets which populate content model
  • Detailed terminological underpinning
  • Sounds a lot like SNOMED

• Memorandum of Agreement – July 2010!
  • WHO right to use for authoring and interpretation
Potential Future States

ICD-11

Ghost SNOMED

SNOMED

Ghost ICD
How should a Disease Taxonomy be Constructed?

- Phenomenological – symptomatic, syndromic?
- Logical assertions – pathophysiological facts?
- Etiological – causes and predispositions
  - At what level of granularity?
- Anatomically structured – what goes wrong where?
  - At what level of anatomy?
- Descriptive – random observations?
- Clinical – disease course and management?
- Genomic – disease polymorphisms?
Classification Structure

- Random list
- Structured Hierarchy
- Frame-structures
- Description Logics
Description Logics

• Basis for Modern Terminology Authoring
• A Computational Subset of Predicate Logic
  • Subset of First Order Predicate Logic (FOPR)
  • FOPR is Not Computationally Tractable
• Derived from “Frame Language” Heritage
• Always Balance Computational Tractability with Expressiveness
• Notation for DL ≠ FOPR
• OWL (Web Ontology Language) most popular dialect
  • Semantic evolution: XML → RDF → OWL

Logic assertions cannot handle probability
NCBO Browser View MI in SNOMED

- Myocardial disease
- Injury of anatomical site
- Structural disorder of heart
- Acute heart disease
- Acute myocardial infarction

Myocardial infarction
# Standardized diagnostic algorithm

## WHO-ARIC-Olmsted Co

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Cardiac pain present</th>
<th>Cardiac pain absent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abnormal</strong></td>
<td>ECG</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>Evolving Diag</td>
<td>Abnormal</td>
<td>Definite MI</td>
</tr>
<tr>
<td>Diagnostic</td>
<td>Definite MI</td>
<td>Diagnostic</td>
</tr>
<tr>
<td>Evolving ST-T</td>
<td>Definite MI</td>
<td>Evolving ST-T</td>
</tr>
<tr>
<td>Equivocal</td>
<td>Definite MI</td>
<td>Equivocal</td>
</tr>
<tr>
<td>Absent</td>
<td>Probable MI</td>
<td>Absent</td>
</tr>
<tr>
<td><strong>Equivocal</strong></td>
<td>ECG</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>Evolving Diag</td>
<td>Equivocal</td>
<td>Definite MI</td>
</tr>
<tr>
<td>Diagnostic</td>
<td>Probable MI</td>
<td>Diagnostic</td>
</tr>
<tr>
<td>Evolving ST-T</td>
<td>Probable MI</td>
<td>Evolving ST-T</td>
</tr>
<tr>
<td>Equivocal</td>
<td>Suspect MI</td>
<td>Equivocal</td>
</tr>
<tr>
<td>Absent</td>
<td>Suspect MI</td>
<td>Absent</td>
</tr>
<tr>
<td><strong>Incomplete</strong></td>
<td>ECG</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>Evolving Diag</td>
<td>Incomplete</td>
<td>Definite MI</td>
</tr>
<tr>
<td>Diagnostic</td>
<td>Incomplete</td>
<td>Diagnostic</td>
</tr>
<tr>
<td>Evolving ST-T</td>
<td>Incomplete</td>
<td>Evolving ST-T</td>
</tr>
<tr>
<td>Equivocal</td>
<td>Incomplete</td>
<td>Equivocal</td>
</tr>
<tr>
<td>Absent</td>
<td>No MI</td>
<td>Absent</td>
</tr>
<tr>
<td><strong>Normal</strong></td>
<td>ECG</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>Evolving Diag</td>
<td>Normal</td>
<td>Definite MI</td>
</tr>
<tr>
<td>Diagnostic</td>
<td>No MI</td>
<td>Diagnostic</td>
</tr>
<tr>
<td>Evolving ST-T</td>
<td>No MI</td>
<td>Evolving ST-T</td>
</tr>
<tr>
<td>Equivocal</td>
<td>No MI</td>
<td>Equivocal</td>
</tr>
<tr>
<td>Absent</td>
<td>No MI</td>
<td>Absent</td>
</tr>
</tbody>
</table>

NHGRI eMERGE (U01) Goals

- GWAS – Genome Wide Association Study
  - 600k Affy chip

- High-throughput Phenotyping
  - Disease algorithm scans across EMRs
  - “catch up” with high throughput genomics

- Generalize Phenotypes across the Consortium
  - Measure reproducibility of algorithms among members
  - Vanderbilt, Northwestern, Marshfield, Group Health Seattle, Mayo
### SHARP: Area 4: Secondary Use of EHR Data

**A $15M National Consortium**

<table>
<thead>
<tr>
<th>Themes</th>
<th>Projects</th>
<th>Players</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Normalization</td>
<td>Clinical Data Normalization</td>
<td>IBM, Mayo, Utah, Agilex</td>
</tr>
<tr>
<td>Phenotype Recognition</td>
<td>Natural Language Processing (NLP)</td>
<td>Harvard, Group Health, IBM, Utah, Mayo, MIT, SUNY, i2b2, Pittsburgh, Colorado</td>
</tr>
<tr>
<td>Data Quality and Evaluation Frameworks</td>
<td>High-Throughput Phenotyping</td>
<td>CDISC, Centerphase, Mayo, Utah</td>
</tr>
<tr>
<td>Data Quality</td>
<td>UIMA and Scaling Capacity</td>
<td>IBM, Mayo</td>
</tr>
<tr>
<td></td>
<td>Data Quality</td>
<td>Mayo, Utah</td>
</tr>
<tr>
<td></td>
<td>Evaluation Framework</td>
<td>Agilex, MN HIE, Mayo, Utah</td>
</tr>
</tbody>
</table>
ICD11 Use Cases

• Scientific consensus of *clinical phenotype*

• Public Health Surveillance
  • Mortality
  • Public Health Morbidity

• Clinical data aggregation
  • Metrics of clinical activity
  • Quality management
    • Patient Safety
  • Financial administration
    • Case mix
    • Resource allocation
Where is This Going?

• ICD is embracing a rich information model
  • Frame-based, hybrid assertions
• Explicit linkage to related ontologies is core to populating ICD11’s Foundation Component
• Phenomenological description of phenotype should be undertaken with ontological elements
  • Raises questions about aggregation logics
  • May not be ideally suited to description logics
  • Must account for probabilistic information