Clinical Terminologies from the Research Perspective

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
Learning objectives

- Describe the history of biomedical ontologies
- List and describe the main biomedical ontologies used in 21st century healthcare
- Discuss the purpose of biomedical ontologies in knowledge management, clinical decision support and analytics
References

Review articles


Additional references


Outline

◆ (Brief) Historical perspective

◆ “High-Impact” Biomedical Ontologies
  ● Structural perspective

◆ Biomedical Ontologies “in Action”
  ● Functional perspective
Clinical Terminologies
Part 1

Historical perspective
History of Medical Ontologies

Synopsis
Nosologiae
Methodicae

1603
1700
1785
1855
1900
1975

ICD
ICD9

1700
1855
1900
1975

OPCS
OPCS3
OPCS4
OPCS4.3

SNOP
CPT

1975
1985
1995
2005

ICPC
READ
SNOMED
SNOMED-RT
SNOMED-CT

1985
1995
2005

UMLS
FMA
GALEN
DM&D

SNOMED-2
SNOMED International
SNOMED-RT

CTV3

[Based on Bodenreider, BIB 2006]
“High-Impact” Biomedical Ontologies

A Structural Perspective
Overview

◆ Structural perspective
  ● What are they (vs. what are they for)?

◆ “High-impact” biomedical ontologies
  ● International Classification of Diseases (ICD)
  ● Logical Observation Identifiers, Names and Codes (LOINC)
  ● SNOMED Clinical Terms
  ● Foundational Model of Anatomy
  ● Gene Ontology
  ● RxNorm
  ● Medical Subject Headings (MeSH)
  ● NCI Thesaurus
  ● Unified Medical Language System (UMLS)

[J. Cimino, YBMI 2006]
International Classification of Diseases
ICD Characteristics (1)

- Current version: ICD-10 (2016)
  - Annual updates
- Type: Classification
- Domain: Disorders
- Developer: World Health Organization (WHO)
- Funding: WHO
- Publicly available: Yes
- Used for: Mortality and morbidity statistics worldwide
- URL: http://www.who.int/classifications/icd/en/
ICD Characteristics (2)

◆ Number of
  ● Concepts: 12,320 (ICD-10, 2004)
  ● Terms: 1 per concept (tabular)

◆ Major organizing principles:
  ● Tree (single inheritance hierarchy)
  ● No explicit classification criteria
    - Idiosyncratic inclusion/exclusion mechanism
  ● .8 slots for \textit{Not elsewhere classified} (NEC)
  ● .9 slots for \textit{Not otherwise specified} (NOS)

◆ Specific coding rules

◆ Distribution: Proprietary format
ICD Top level

ICD-10 Version: 2016

- I Certain infectious and parasitic diseases
- II Neoplasms
- III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- IV Endocrine, nutritional and metabolic diseases
- V Mental and behavioural disorders
- VI Diseases of the nervous system
- VII Diseases of the eye and adnexa
- VIII Diseases of the ear and mastoid process
- IX Diseases of the circulatory system
- X Diseases of the respiratory system
- XI Diseases of the digestive system
- XII Diseases of the skin and subcutaneous tissue
- XIII Diseases of the musculoskeletal system and connective tissue
- XIV Diseases of the genitourinary system
- XV Pregnancy, childbirth and the puerperium
- XVI Certain conditions originating in the perinatal period
- XVII Congenital malformations, deformations and chromosomal abnormalities
- XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- XIX Injury, poisoning and certain other consequences of external causes
- XX External causes of morbidity and mortality
- XXI Factors influencing health status and contact with health services
- XXII Codes for special purposes
ICD Example

Idiosyncratic inclusion/exclusion criteria

---

Type 1 diabetes mellitus

**Incl.**
- diabetes (mellitus):
  - brittle
  - juvenile-onset
  - ketosis-prone

**Excl.**
- diabetes mellitus (in):
  - malnutrition-related (E12.-)
  - neonatal (P70.2)
  - pregnancy, childbirth and the puerperium (O24.-)

glycosuria:
- NOS (R81)
- renal (E74.8)

impaired glucose tolerance (R73.0)
postsurgical hypoinsulinaemia (E89.1)
**ICD Example**

- *Not elsewhere classified (NEC)*
- *Not otherwise specified (NOS)*

---

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>E84</td>
<td>Cystic fibrosis</td>
</tr>
<tr>
<td></td>
<td><em>Incl.</em>: mucoviscidosis</td>
</tr>
<tr>
<td>E84.0</td>
<td>Cystic fibrosis with pulmonary manifestations</td>
</tr>
<tr>
<td>E84.1</td>
<td>Cystic fibrosis with intestinal manifestations</td>
</tr>
<tr>
<td></td>
<td>Distal intestinal obstruction syndrome</td>
</tr>
<tr>
<td></td>
<td>Meconium ileus in cystic fibrosis† (<em>P75</em>)</td>
</tr>
<tr>
<td></td>
<td><em>Excl.</em>: meconium obstruction (ileus) in cases where cystic fibrosis is known not to be present (<em>P76.0</em>)</td>
</tr>
<tr>
<td>E84.8</td>
<td>Cystic fibrosis with other manifestations</td>
</tr>
<tr>
<td>E84.9</td>
<td>Cystic fibrosis, unspecified</td>
</tr>
</tbody>
</table>
ICD-10-CM

- Derived from: ICD-10
  - Finer-grained (both clinically and administratively)
- Type: Classification
  - 92,042 codes (2015)
  - Terms: 1.2 per concept
- Domain: Disorders
- Developer: National Center for Health Statistics (NCVHS)
- Funding: U.S. Government
- Publicly available: Yes
- Used for: Billing
- URL: http://www.cdc.gov/nchs/icd/icd10cm.htm
ICD-10 vs. ICD-10-CM

E72 Other disorders of amino-acid metabolism

Excl.: abnormal findings without manifest disease (RZ).

Excludes:
- aromatic amino-acid metabolism (E70.-)
- branched-chain amino-acid metabolism (E71.3)
- fatty-acid metabolism (E71.3)
- purine and pyrimidine metabolism (E79.-)
- gout (M10.-)

E72.0 Disorders of amino-acid transport

Cystine storage disease* (N29.8*)
Cystinosis
Cystinuria
Fanconi(-de Toni)(-Debré) syndrome
Hartnup disease
Lowe syndrome

Excl.: disorders of tryptophan metabolism (E70.8)
ICD-10 vs. ICD-10-CM

W58 Bitten or struck by crocodile or alligator

W58 Contact with crocodile or alligator

The appropriate 7th character is to be added to each code from category W58
A - initial encounter
D - subsequent encounter
S - sequela

W58.0 Contact with alligator

W58.01 Bitten by alligator
W58.02 Struck by alligator
W58.03 Crushed by alligator
W58.09 Other contact with alligator

W58.1 Contact with crocodile

W58.11 Bitten by crocodile
W58.12 Struck by crocodile
W58.13 Crushed by crocodile
W58.19 Other contact with crocodile

W58.01A Bitten by alligator, initial encounter
W58.01D Bitten by alligator, subsequent encounter
W58.01S Bitten by alligator, sequela
Logical Observation Identifiers, Names and Codes (LOINC)
LOINC Characteristics (1)

- Current version: 2.59 (Feb. 2017)
  - 2 annual releases
- Type: Controlled terminology*
- Domain: Laboratory and clinical observations
- Developer: Regenstrief Institute
- Funding: NLM and other sources
- Publicly available: Yes
- Used for: information exchange
- URL: https://loinc.org/
LOINC Characteristics (2)

◆ Number of
  ● Concepts: 73,958 active codes (2.52, June 2015)
  ● Terms: 1 per concept (“long name”)

◆ Major organizing principles:
  ● No hierarchical structure among the main codes
  ● 6 axes
    ▪ Component (analyte [+ challenge] [+ adjustments])
    ▪ Property
    ▪ Timing
    ▪ System
    ▪ Scale
    ▪ [Method]

◆ Distribution: proprietary database format
LOINC Example

- **Sodium [Moles/volume] in Serum or Plasma**
  [the molar concentration of sodium is measured in the plasma (or serum), with quantitative result]

<table>
<thead>
<tr>
<th>Axis</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component</td>
<td>Sodium</td>
</tr>
<tr>
<td>Property</td>
<td>SCnc – Substance Concentration (per volume)</td>
</tr>
<tr>
<td>Timing</td>
<td>Pt – Point in time (Random)</td>
</tr>
<tr>
<td>System</td>
<td>Ser/Plas – Serum or Plasma</td>
</tr>
<tr>
<td>Scale</td>
<td>Qn – Quantitative</td>
</tr>
<tr>
<td>Method</td>
<td>--</td>
</tr>
</tbody>
</table>
2951-2 Sodium [Moles/volume] in Serum or Plasma

<table>
<thead>
<tr>
<th>Fully-Specified Name:</th>
<th>Sodium</th>
</tr>
</thead>
</table>

**PART DEFINITION/DESCRIPTION(S)**

Sodium is an essential nutrient that regulates blood volume, blood pressure, osmotic equilibrium and electrolyte balance. Sodium chloride is the principal source of sodium in the diet, and is used for seasoning and as a preservative. Increased levels of sodium intake can cause hypertension and reportedly leads to 7.6 million premature deaths worldwide. Sodium is also important in neuron function and osmoregulation between cells and the extracellular fluid.

**Copyright:** Text is available under the Creative Commons Attribution/Share-Alike License. See http://creativecommons.org/licenses/by-sa/3.0/ for details.

**Source:** Wikipedia, URL: Sodium (Wikipedia)

<table>
<thead>
<tr>
<th>BASIC ATTRIBUTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class/Type:</td>
</tr>
<tr>
<td>CDISC Lab Test:</td>
</tr>
<tr>
<td>Common Lab Results Rank:</td>
</tr>
<tr>
<td>Common SI Lab Results Rank:</td>
</tr>
<tr>
<td>Common Orders Rank:</td>
</tr>
<tr>
<td>Last Updated in Version:</td>
</tr>
<tr>
<td>Order vs. Obs.:</td>
</tr>
<tr>
<td>Status:</td>
</tr>
</tbody>
</table>

**EXAMPLE UNITS**

<table>
<thead>
<tr>
<th>Unit</th>
<th>Source Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>mmol/L</td>
<td>EXAMPLE UCUM UNITS</td>
</tr>
<tr>
<td>mmol/L</td>
<td>REGENSTRIEF</td>
</tr>
<tr>
<td>mmol/L</td>
<td>eCHN</td>
</tr>
</tbody>
</table>

**UNITS AND RANGE**

<table>
<thead>
<tr>
<th>Range</th>
<th>Units Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>mmol/L.[136,145]</td>
<td></td>
</tr>
</tbody>
</table>
SNOMED Clinical Terms
SNOMED CT Characteristics (1)

- Current version: January 31, 2017
  - 2 annual releases
- Type: Reference terminology / ontology
- Domain: Clinical medicine
- Developer: IHTSDO
- Funding: IHTSDO member countries
- Publicly available: Yes*
- Used for: clinical documentation, information exchange, analytics
- URL: http://www.ihtsdo.org/
SNOMED CT Characteristics (2)

- Number of
  - Concepts: 320,912 active concepts (Sept. 2016)
  - Terms: 2.6 per concept (“descriptions”)

- Major organizing principles:
  - Polyhierarchy
  - Rich set of associative relationships
  - Logical definitions (incomplete: many primitives)
  - Built using description logics (EL++)

- Distribution: RF2 (proprietary)
SNOMED CT Top level

- SNOMED CT Concept
  - Body structure (body structure)
  - Clinical finding (finding)
  - Environment or geographical location (environment / location)
  - Event (event)
  - Observable entity (observable entity)
  - Organism (organism)
  - Pharmaceutical / biologic product (product)
  - Physical force (physical force)
  - Physical object (physical object)
  - Procedure (procedure)
  - Qualifier value (qualifier value)
  - Record artifact (record artifact)
  - Situation with explicit context (situation)
  - SNOMED CT Model Component (metadata)
  - Social context (social concept)
  - Special concept (special concept)
  - Specimen (specimen)
  - Staging and scales (staging scale)
  - Substance (substance)
**SNOMED CT Example**

### Parents
- Operation on appendix (procedure)
- Partial excision of large intestine (procedure)

### Appendectomy (procedure)
- SCTID: 80146002
- 80146002 | Appendectomy (procedure) |
  - Appendectomy
  - Excision of appendix
  - Appendicectomy
  - Appendectomy (procedure)

### Children (8)
- Appendectomy with drainage (procedure)
- Emergency appendectomy (procedure)
- Excision of appendiceal stump (procedure)
- Excision of ruptured appendix by open approach (procedure)
- Incidental appendectomy (procedure)
- Interval appendectomy (procedure)
- Laparoscopic appendectomy (procedure)
- Non-emergency appendectomy (procedure)
SNOMED CT Example

80146002
Appendectomy (procedure)

27010001
Partial excision of large intestine (procedure)

8613002
Operation on appendix (procedure)

405813007
Procedure site - Direct (attribute)

66754008
Appendix structure (body structure)

260686004
Method (attribute)

129304002
Excision - action (qualifier value)
RxNorm
RxNorm Characteristics (1)

- Current version: March 2017
  - Monthly releases (+weekly updates)
- Type: Controlled terminology
- Domain: Drug names
- Developer: NLM
- Funding: NLM
- Publicly available: Yes*
- Used for: e-prescribing, information exchange, analytics
- URL: http://www.nlm.nih.gov/research/umls/rxnorm/
RxNorm Characteristics (2)

- **Number of**
  - Concepts: 117,774 (March 2016)
  - Terms: 1.5 per concept

- **Major organizing principles:**
  - Generic vs. brand
  - Ingredient + Strength + Dose form
  - No hierarchical structure; rich graph of associative relations
  - Integrates all major US drug information sources
  - No clinical information

- **Distribution:** similar to UMLS RRF format
**RxNorm** Normalized form

- **Ingredient**: Fluoxetine
- **Strength**: 4mg/ml
- **Dose form**: Oral Solution

Semantic clinical drug component

Semantic clinical drug form

Semantic clinical drug
**RxNorm Example**

- **Ingredient**: *Azithromycin*
- **Brand Name**: *Zithromax*

**C. Drug Comp.**
- *Azithromycin 250 MG*

**C. Drug Form**
- *Azithromycin Oral Tablet*

**B. Drug Comp.**
- *Azithromycin 250 MG*  
  *Azithromycin Oral Tablet*

**B. Drug**
- *Zithromax 250 MG Oral Tablet*

**G. Pack**
- *{6 (Azithromycin 250 MG Oral Tablet)} Pack*

**B. Pack**
- *Z-PAK*
Product Identification: NDCs

- **National Drug Codes**
  - Product identification system
  - Three components
    - Manufacturer
    - Product
    - Packaging

- Introduced in 1972 by FDA
- Only format permitted by NCPDP
- Mandated by HIPAA regulations for drug transactions

Source: Dan Malone
NDC Elements: 3 segments

XXXXX-XXXX-XX

Manufacturer  Product  Packaging
NDC Forms

Warfarin Sodium 1 MG Oral Tablet

XXXX-XXXX-XX (4-4-2) → 0XXXX XXXX XX
  0555-0831-02 (Teva Pharmaceuticals USA, Inc.; 100 in 1 BOTTLE) → 00555083102

XXXXXX-XXX-XX (5-3-2) → XXXXX 0XXX XX
  21695-672-30 (Rebel Distributors Corp; 30 in 1 BOTTLE) → 21695067230

XXXXXX-XXXX-X (5-4-1) → XXXXX XXXX 0X
  50090-1213-0 (A-S Medication Solutions; 30 in 1 BOTTLE) → 50090121300
Clinical Terminologies
Part 3

Biomedical Ontologies “in Action”

A Functional Perspective
Overview

◆ Functional perspective
  ● What are they for (vs. what are they)?

◆ “High-impact” biomedical ontologies

◆ 3 major categories of use
  ● Knowledge management (indexing and retrieval of data and information, access to information, mapping among ontologies)
  ● Data integration, exchange and semantic interoperability
  ● Decision support and analytics (data selection and aggregation, decision support, natural language processing applications, knowledge discovery)

[Bodenreider, YBMI 2008]
Knowledge management
Knowledge management

Annotating data and resources
Terminology in ontology

◆ Ontology as a source of vocabulary
  ● List of names for the entities in the ontology (ontology vs. terminology)

◆ Most ontologies have some sort of terminological component

◆ Not all surface forms represented
  ● Often insufficient for NLP applications
  ● Large variation in number of terms per concept across ontologies
Annotating data

◆ Gene Ontology
  ● Functional annotation of gene products in several dozen model organisms

◆ Various communities use the same controlled vocabularies

◆ Enabling comparisons across model organisms

◆ Annotations
  ● Assigned manually by curators
  ● Inferred automatically (e.g., from sequence similarity)
## GO Annotations across species

### ALDH2  aldehyde dehydrogenase 2 family (mitochondrial) [Homo sapiens (human)]

<table>
<thead>
<tr>
<th>Function</th>
<th>Evidence Code</th>
<th>Pubs</th>
</tr>
</thead>
<tbody>
<tr>
<td>aldehyde dehydrogenase (NAD) activity</td>
<td>EXP</td>
<td></td>
</tr>
<tr>
<td>aldehyde dehydrogenase (NAD) activity</td>
<td>IDA</td>
<td>PubMed</td>
</tr>
<tr>
<td>aldehyde dehydrogenase [NAD(P)+] activity</td>
<td>TAS</td>
<td>PubMed</td>
</tr>
<tr>
<td>electron carrier activity</td>
<td>TAS</td>
<td>PubMed</td>
</tr>
</tbody>
</table>

### Aldh2  aldehyde dehydrogenase 2, mitochondrial [Mus musculus (house mouse)]

<table>
<thead>
<tr>
<th>Function</th>
<th>Evidence Code</th>
<th>Pubs</th>
</tr>
</thead>
<tbody>
<tr>
<td>NADH binding</td>
<td>ISO</td>
<td></td>
</tr>
<tr>
<td>aldehyde dehydrogenase (NAD) activity</td>
<td>IBA</td>
<td></td>
</tr>
<tr>
<td>aldehyde dehydrogenase (NAD) activity</td>
<td>ISO</td>
<td></td>
</tr>
<tr>
<td>identical protein binding</td>
<td>ISO</td>
<td></td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>IAE</td>
<td></td>
</tr>
<tr>
<td>oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor</td>
<td>IEP</td>
<td>PubMed</td>
</tr>
<tr>
<td>protein binding</td>
<td>IP1</td>
<td></td>
</tr>
</tbody>
</table>
Indexing the biomedical literature

◆ **MeSH**
  - Used for indexing and retrieval of the biomedical literature (MEDLINE)

◆ **Indexing**
  - Performed manually by human indexers
    - With help of semi-automatic systems (suggestions)
      e.g., Indexing Initiative at NLM
  - Specific indexing rules
Free cortisol in sepsis and septic shock.
Bendel S¹, Karlsson S, Pettilä V, Loisa P, Varpula M, Ruokonen E; Finnsepsis Study Group.

Abstract
BACKGROUND: Severe sepsis activates the hypothalamopituitary axis, increasing cortisol production. In some studies, hydrocortisone substitution based on an adrenocorticotropic hormone-stimulation test or baseline cortisol measurement has improved outcome. Because only the free fraction of cortisol is active, measurement of free cortisol may be more important than total cortisol in critically ill patients. We measured total and free cortisol in patients with severe sepsis and related the concentrations to outcome.

METHODS: In a prospective study, severe sepsis was defined according the American College of Chest Physicians/Society of Critical Care Medicine criteria. Blood samples were drawn within 24 h of study entry. Serum cortisol was analyzed by electrochemiluminescence immunoassay. The Coolens method was used for calculating serum free cortisol concentrations.

RESULTS: Blood samples were collected from 125 patients, of whom 62 had severe sepsis and 63 septic shock. Hospital mortality was 21%. Calculated free serum cortisol correlated well with serum total cortisol (r = 0.90, P < 0.001). There was no difference in the total cortisol concentrations in patients with sepsis and septic shock (728 +/- 386 nmol/L vs 793 +/- 439 nmol/L, P = 0.44). Nonsurvivors had higher calculated serum free (209 +/- 151 nmol/L) and total (980 +/- 458 nmol/L) cortisol concentrations than survivors (119 +/- 111 nmol/L, P = 0.002, and 704 +/- 383 nmol/L, P = 0.002). Depending on the definition, the incidence of adrenal insufficiency varied from 8% to 54%.

CONCLUSIONS: Clinically, calculation of free cortisol does not provide essential information for identification of patients who would benefit from corticoid treatment in severe sepsis and septic shock.
MeSH MEDLINE indexing

MeSH Terms

- Adrenal Cortex Function Tests
- Adrenal Insufficiency/blood*
- Adrenal Insufficiency/drug therapy
- Adrenal Insufficiency/mortality
- Adult
- Biomarkers/blood
- Female
- Finland/epidemiology
- Hospital Mortality
- Humans
- Hydrocortisone/blood*
- Hydrocortisone/therapeutic use
- Kaplan-Meier Estimate
- Male
- Predictive Value of Tests
- Prospective Studies
- Sepsis/blood*
- Sepsis/drug therapy
- Sepsis/mortality
- Severity of Illness Index
- Shock, Septic/blood*
- Shock, Septic/drug therapy
- Shock, Septic/mortality
- Treatment Outcome
SNOMED CT/ICD  Coding clinical data

◆ SNOMED CT
  ● Used for clinical documentation
  ● E.g., problem lists

◆ ICD-10-CM
  ● Used for coding clinical data for billing purposes
  ● Other uses of ICD
    ▪ Morbidity and mortality reporting worldwide
  ● Specific coding rules
Knowledge management

Accessing biomedical information
Resources for biomedical search engines

- Synonyms
- Hierarchical relations
- High-level categorization
- [Co-occurrence information]
- Translation
MeSH "synonyms" MEDLINE retrieval

- MeSH entry terms
  - Used as equivalent terms for retrieval purposes (query expansion)
  - Not always synonymous

- Increase recall without hurting precision

<table>
<thead>
<tr>
<th>MeSH Heading</th>
<th>Entry Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td>Addison's Disease</td>
</tr>
<tr>
<td></td>
<td>Primary Adrenal Insufficiency</td>
</tr>
<tr>
<td></td>
<td>Primary Adrenocortical Insufficiency</td>
</tr>
<tr>
<td></td>
<td>Primary Hypoadrenalism</td>
</tr>
</tbody>
</table>
MeSH “synonyms” MEDLINE retrieval

"addison disease"[MeSH Terms] OR
("addison"[All Fields] AND
"disease"[All Fields]) OR "addison disease"[All Fields] OR
("primary"[All Fields] AND
"hypoadrenalism"[All Fields]) OR "primary hypoadrenalism"[All Fields]
MeSH hierarchies  MEDLINE retrieval

◆ MeSH “explosion”
  • Search for a given MeSH term and all its descendants
  • A search on Adrenal insufficiency also retrieves articles indexed with its descendant, Addison disease
Knowledge management

Mapping across biomedical ontologies
Terminology integration systems

- Terminology integration systems (UMLS, RxNorm) help bridge across vocabularies

- Uses
  - Information integration
  - Ontology alignment
  - Medication reconciliation
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Other subdomains
- SNOMED CT
- OMIM
- MeSH
- Biomedical literature
- NCBI Taxonomy
- Genomic annotations
- Model organisms
- FMA
- GO
- Anatomy
- UMLS
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Anatomy
- Model organisms
- Other subdomains
Trans-namespace integration

Addison's disease (363732003)

Other subdomains

Clinical repositories

SNOMED CT

NCBI Taxonomy

Model organisms

FMA

Anatomy

Genome annotations

Genetic knowledge bases

Biomedical literature

Addison Disease (D000224)
UMLS Source Vocabularies

- 154 families of source vocabularies
  - Not counting translations
- 25 languages
- Broad coverage of biomedicine
  - 10.3M names (normalized)
  - 3.4M concepts
  - ~13M relations among concepts
- Common presentation
Metathesaurus Basic organization

◆ Concepts
  • Synonymous terms are clustered into a concept
  • Properties are attached to concepts, e.g.,
    - Unique identifier
    - Definition

◆ Relations
  • Concepts are related to other concepts
  • Properties are attached to relations, e.g.,
    - Type of relationship
    - Source
Decision support and analytics
Data selection

◆ The structure of biomedical ontologies helps define groups of values from a high-level value
  ● Vs. enumerating all possible values
◆ Useful for data selection in clinical studies
◆ ICD is used pervasively for this purpose
  ● E.g., Study on supraventricular tachycardia (SVT), based on 2 high-level ICD codes
◆ Similarity with the definition of value sets for use in the information model
Data aggregation

◆ Ontologies help partition/aggregate data in data analysis

- Clinical studies: Study a variable in groups of patients corresponding to the top level categories in ICD
- Biology studies: Functional characterization of gene expression signatures with high-level concepts from the Gene Ontology
  - Recent trend: co-clustering
Decision support

◆ Clinical decision support
  ● Ontologies help normalize the vocabulary and increase the recall of rules
  ● Ontologies provide some domain knowledge and make it possible to create high-level rules (e.g., for a class of drugs rather than for each drug in the class)

◆ Other forms of decision support
  ● Based on automatic reasoning services for OWL ontologies (e.g., grading gliomas with NCIt)
Natural language processing applications

- Ontologies provide background domain knowledge for NLP applications
  - Question answering
  - Document summarization
  - Literature-based discovery

- The UMLS is often used, but other specific resources have been developed
Knowledge discovery

- By standardizing the vocabulary in a given domain, ontologies are enabling resources for knowledge discovery through data mining.
- Less frequently, the structure of the ontology is leveraged by data mining algorithms.
- Example of available datasets:
  - ICD-coded clinical data (in conjunction with non-clinical information, e.g., environmental data)
  - Annotation of gene products to the GO (function prediction)
Controlled Vocabularies

Summary
Summary

◆ History of biomedical ontologies
◆ Structure of the main clinical ontologies used
  ● ICD, SNOMED CT, LOINC, RxNorm
◆ Purpose of biomedical ontologies
  ● Knowledge management, [health information exchange and semantic interoperability], and clinical decision support and analytics
Topics not discussed

◆ Semantic Web, URIs, Linked Data
◆ Ontology creation, Protege
◆ Accessing terminology resources (APIs)
◆ Ontology repositories
  ● [UMLS], NCBO BioPortal, EBI Ontology Lookup Service
◆ NLP, named entity recognition, MetaMap
◆ Mapping local terms to standard terminologies
◆ VSAC, value sets, common data elements
◆ OBO ontologies, OBO Foundry
◆ Coordinated development of ontologies, harmonization
◆ Boundary between terminology and information model
◆ […]
Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: https://mor.nlm.nih.gov

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA

U.S. National Library of Medicine