



Short course – Summer 2008
Biomedical Ontology in Practice

June 9-11, 2008

Biomedical Ontology in Practice



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

Objectives

- ◆ Learn about biomedical ontologies
 - History
 - Design principles, formalisms and tools
 - What are they?
 - What are they used for?
- ◆ Work with biomedical ontologies
 - Search
 - Analyze
 - Extend
 - Use for data integration



Lister Hill National Center for Biomedical Communications 2

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration



Lister Hill National Center for Biomedical Communications 3

References Bio-ontology courses

- ◆ Barry Smith, U. Buffalo / NCBO
 - http://ontology.buffalo.edu/smith/Ontology_Course.html
- ◆ Stefan Schulz, U. Freiburg, Germany / KR-MED 2008 tutorial
 - <http://www.kr-med.org/2008/index.html>



Lister Hill National Center for Biomedical Communications 4

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration



Lister Hill National Center for Biomedical Communications 5



Short course – Summer 2008
Biomedical Ontology in Practice

June 9, 2008 – Session #1

Introduction to Biomedical Ontologies



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

Outline

- ◆ Historical perspective
- ◆ Introduction to biomedical terminologies through an example
- ◆ Biomedical terms as names for biomedical classes
- ◆ Terminological relations as a surrogate for ontological relations


 Lister Hill National Center for Biomedical Communications 7

Historical perspective

Why biomedical terminologies?

- ◆ To support a theory of diseases
- ◆ To classify diseases
- ◆ To support epidemiology
- ◆ To index and retrieve information
- ◆ To serve as a reference


 Lister Hill National Center for Biomedical Communications 9

To support a theory of diseases

- ◆ Hippocrates
 - Dismisses superstition
 - Four humors
 - Blood
 - Phlegm
 - Yellow bile
 - Black bile
- ◆ Thomas Sydenham (1624-1689)
 - *Medical observations on the history and cure of acute diseases* (1676)





 Lister Hill National Center for Biomedical Communications

To classify diseases (and plants)

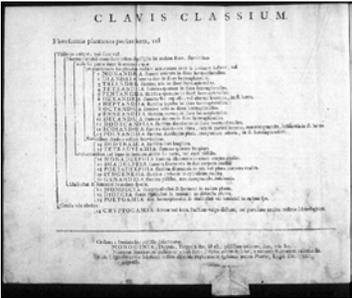
- ◆ Carolus Linnaeus (1707-1778)
 - *Genera Plantarum* (1737)
 - *Genera Morborum* (1763)
- ◆ François Boissier de La Croix a.k.a. F. B. de Sauvages (1706-1767)
 - *Methodus Foliorum* (1751)
 - *Nosologia Methodica* (1763/68)
- ◆ William Cullen (1710-1790)
 - *Synopsis Nosologiae Methodicae* (1785)






 Lister Hill National Center for Biomedical Communications 11

From plants...

... to diseases

- ◆ Four categories (W. Cullen)
 - Fevers
 - Nervous disorders
 - Cachexias
 - Local diseases

“The distinction of the genera of diseases, the distinction of the species of each, and often even that of the varieties, I hold to be a necessary foundation of every plan of physic, whether dogmatical or empirical.”
– William Cullen, Edinburgh, 1785
Synopsis Nosologia Methodicae

(Cited by Chris Chute)

Lister Hill National Center for Biomedical Communications

13

To support epidemiology

- ◆ John Graunt (1620-1674)
 - Analyzes the vital statistics of the citizens of London
- ◆ William Farr (1807-1883)
 - Medical statistician
 - Improves Cullen’s classification
 - Contributes to creating ICD
- ◆ Jacques Berthillon (1851-1922)
 - Chief of the statistical services (Paris)
 - Classification of causes of death (161 rubrics)

Lister Hill National Center for Biomedical Communications

14

London Bills of Mortality

Lister Hill National Center for Biomedical Communications

15

Limitations of existing classifications

“The advantages of a uniform statistical nomenclature, however imperfect, are so obvious, that it is surprising no attention has been paid to its enforcement in Bills of Mortality. Each disease has, in many instances, been denoted by three or four terms, and each term has been applied to as many different diseases; vague, inconvenient names have been employed, or complications have been registered instead of primary diseases. [The nomenclature is of as much importance in this department of inquiry as weights and measures in the physical sciences, and should be settled without delay.]”

– William Farr
First annual report.
London, Registrar General of England and Wales, 1839, p. 99.

Lister Hill National Center for Biomedical Communications

16

To index and retrieve information

- ◆ Biomedical literature
 - MEDLINE (15M citations from 4600 journals)
 - Manually indexed
 - Medical Subject Headings (MeSH)
- ◆ Genome
 - Model organism databases (Fly, Mouse, Yeast, ...)
 - Manually / semi-automatically curated
 - Gene Ontology

Lister Hill National Center for Biomedical Communications

17

MEDLINE and MeSH

At a recent meeting, I saw a slide that said: “Black bile and psychomotor retardation: shades of melancholia in Dante’s Inferno.”

Witness: DA.

Memorial Sloan-Kettering Cancer Center, New York, NY 10021, USA. volkmar@mskcc.org

The history of melancholy depression is rich with images of movement retardation and mental dysfunction. The reconcentration of psychomotor symptoms to the single terminology of affective disorder is not novel to the students of modern psychiatry. The move back to the biology of this psychomotor dysfunction with the neural substrates in brain receptors, as well as neurochemical substrates, is testimony to the centrality of movement changes in the depressive condition. The Inferno, the first canto of Dante Alighieri’s *Commedia*, has a wonderful abundance of allusions to the experience of psychomotor symptoms in describing the depressed individual. *Stanza 4* (line 1, *perché, se non, frusto, non*), there and many other images from the physical manifestations of psychomotor suffering in the forefront of the reader’s mind. Considering life-long and Transatlantic settings on melancholy suffering, it is fitting that Dante chose a bodily illness reflected in the hellish torments visited on the damned. From the roots of the Italian to those of the violent, the panorama of psychomotor symptoms plays a prominent role in the poem as well as in the medical and literary prose of succeeding centuries.

MeSH Terms

- Depressive Disorder/therapy*
- History of Medicine, Medieval
- Human
- Italy
- Literature, Medieval/literature*
- Medicine in Literature*
- Poetry/literature*
- Psychomotor Disorder/therapy*

Mouse Genome Database and GO

Entrez Gene

1: NF2 neurofibromin 2 [Mus musculus]

GeneID: 18016 Locus tag: MGL27307

General gene information

Gene Ontology

Provided by MGI

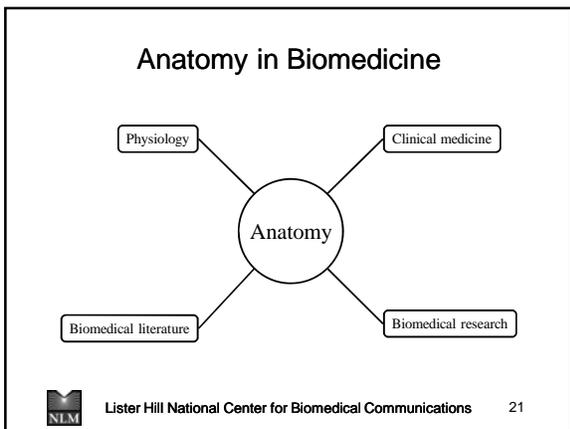
Category	Term	Evidence
Function	structural protein binding	IEA
	protein binding	IFI PubMed
	structural molecule activity	IEA
Process	intracellular protein assembly and/or maintenance	IMP PubMed
	cellular regulation of cell cycle	IEA
	cellular regulation of protein kinase activity	IDA PubMed
	regulation of cell proliferation	IMP PubMed
Component	cellular junction	IMP PubMed
	cytoplasm	IEA
	cytoskeleton	IEA
	cellulose	IEA

Lister Hill National Center for Biomedical Communications 19

To serve as a reference

- ◆ Reference terminology/ontology
 - Universally needed
 - Developed independently of any purposes
 - Reusable by many applications
- ◆ Examples
 - VA National Drug File (NDF)
 - Foundational Model of Anatomy (FMA)
 - SNOMED CT

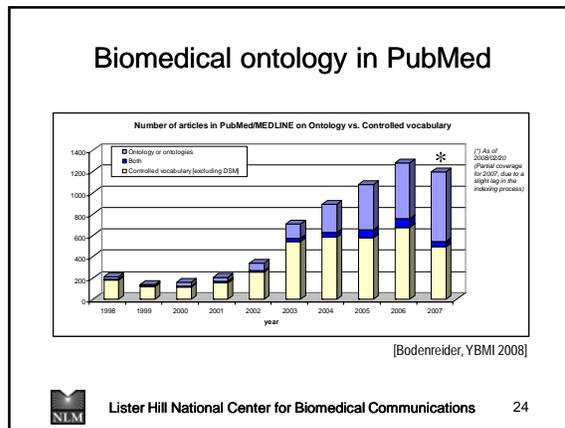
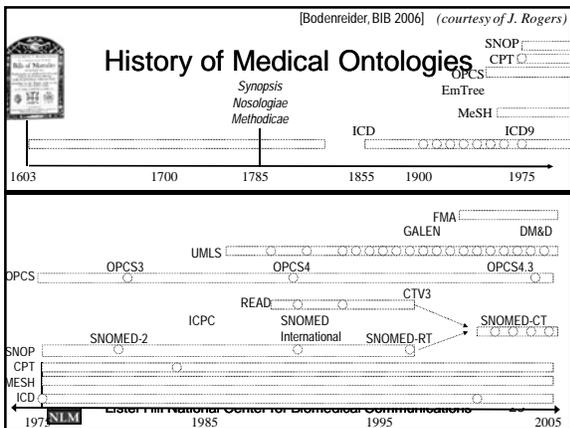
Lister Hill National Center for Biomedical Communications 20

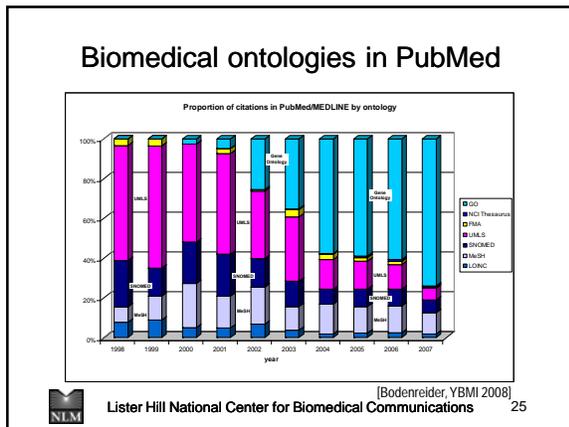


Administrative terminologies

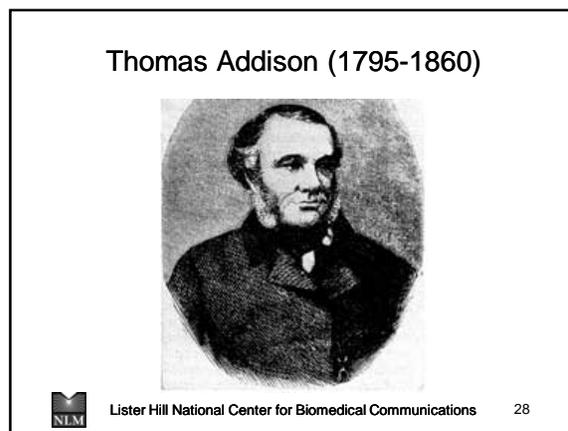
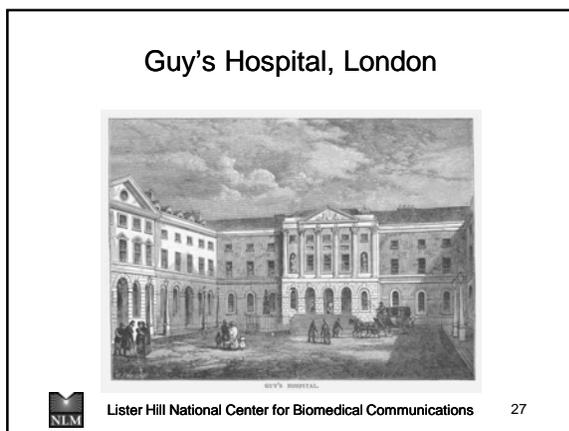
- ◆ Coding patient records
 - International Classification of Primary Care (ICPC)
 - SNOMED
 - Read Codes
- ◆ Reporting claims to health insurance companies
 - Current Procedural Terminology (CPT)
 - International Classification of Diseases (ICD-9 CM)
 - Healthcare Common Procedure Coding System (HCPCS)

Lister Hill National Center for Biomedical Communications 22





Introduction to biomedical terminologies through an example



Addison's disease

- ◆ Addison's disease is a rare endocrine disorder
- ◆ Addison's disease occurs when the adrenal glands do not produce enough of the hormone cortisol
- ◆ For this reason, the disease is sometimes called chronic adrenal insufficiency, or hypocortisolism

Lister Hill National Center for Biomedical Communications 29

Adrenal insufficiency Clinical variants

- ◆ Primary / Secondary
 - Primary: lesion of the adrenal glands themselves
 - Secondary: inadequate secretion of ACTH by the pituitary gland
- ◆ Acute / Chronic
- ◆ Isolated / Polyendocrine deficiency syndrome

Lister Hill National Center for Biomedical Communications 30

Addison's disease: Symptoms

- ◆ Fatigue
- ◆ Weakness
- ◆ Low blood pressure
- ◆ Pigmentation of the skin (exposed and non-exposed parts of the body)
- ◆ ...


Lister Hill National Center for Biomedical Communications
31

AD in medical vocabularies

- ◆ Synonyms: different terms
 - Addisonian syndrome } eponym
 - Bronzed disease } symptoms
 - Addison melanoderma } symptoms
 - Asthenia pigmentosa } symptoms
 - Primary adrenal deficiency } clinical variants
 - Primary adrenal insufficiency } clinical variants
 - Primary adrenocortical insufficiency } clinical variants
 - Chronic adrenocortical insufficiency } clinical variants
- ◆ Contexts: different hierarchies


Lister Hill National Center for Biomedical Communications
32

Internal Classification of Diseases

CHAPTER 4
Endocrine, nutritional and metabolic diseases (E00-E90)

Disorders of other endocrine glands (E20-E35)

E27 Other disorders of adrenal gland

E27.0 Other adrenocortical insufficiency
Overproduction of ACTH, not associated with Cushing's disease
Pregnancy-associated
Excludes1 Cushing's syndrome (E24.)

E27.1 Primary adrenocortical insufficiency
Primary adrenocortical insufficiency
Adrenocortical insufficiency NOS
Autoimmune adrenitis
Excludes1 Addison's nephropathy/adrenoleukodystrophy (I.71.)X2
 nephroses (E26)
 tuberculosis, Addison's disease (A.18.7)
 Waterhouse-Friderichsen syndrome (A.59.1)

E27.2 Adrenocortical crisis
Adrenal crisis
Adrenocortical crisis

E27.3 Drug-induced adrenocortical insufficiency
Corticosteroid (T160-T169) abruptly drug

E27.4 Other and unspecified adrenocortical insufficiency

DRAFT ICD-10 CM Tabular Page 189 Page 2063



Medical Subject Headings



MeSH Tree Structures

[Endocrine Diseases \[C19\]](#)
[Adrenal Gland Diseases \[C19.053\]](#)
 Adrenal Gland Hypofunction [C19.053.264]
 ▶ [Addison's Disease \[C19.053.264.262\]](#)
 ▶ [Adrenocortical Insufficiency \[C19.053.264.270\]](#)
 ▶ [Hypoadosteronism \[C19.053.264.480\]](#)

[Infectious Diseases \[C20\]](#)
[Autoimmune Diseases \[C20.111\]](#)
 ▶ [Addison's Disease \[C20.111.162\]](#)
 ▶ [Adrenal Hypofunction, Autoimmune \[C20.111.172\]](#)
 ▶ [Auto-Immune Reaction, Myxomatous, Juvenile \[C20.111.192\]](#)
 ▶ [Autoimmune Endocrine Pancreas \[C20.111.197\]](#)
 ▶ [Autoimmune Hypoparathyroidism \[C20.111.199\]](#)
 ▶ [Autoimmune Diseases of the Nervous System \[C20.111.232\]](#)


Lister Hill National Center for Biomedical Communications
34

SNOMED CT

386584007 adrenal cortical hypofunction

386584007 **Adrenocortical insufficiency**

22776008 Addison's disease with adrenoleukodystrophy

76715008 Addison's disease due to autoimmunity

186270000 tuberculous Addison's disease

11244009 polyglandular autoimmune syndrome, type 1

Adison's disease - Definition

Concept Status: Current

Descriptions

Adison's disease (disorder)

Adison's disease

enfermedad de Addison

Enfermedad de Addison (trastorno)

Definition: Primitive

Site

Adrenal cortical hypofunction

finding site

Adrenal cortex structure

Qualifiers

severity

onset

recurrence

episodicity

episodicities

special course

courses

Codes

Original SNOMED ID - DB:70620

Root Code (Civ3id) - C1541


Lister Hill National Center for Biomedical Communications
35

Biomedical terms as names for biomedical classes

Terms reflecting valid classes

- Pulmonary anthrax
- BRCA1 protein
- Coronary artery
- Coronary artery bypass
- ...
 - Non-insulin dependent diabetes mellitus
 - Non-Hodgkin lymphoma
 - Non-steroidal anti-inflammatory drugs
 - Non-opioid analgesics
 - Non-invasive medical procedure

 Lister Hill National Center for Biomedical Communications 37

Issues

- ◆ Multiple terms for a class
- ◆ Multiple classes for a term
- ◆ Presence of non-ontological features in terms
- ◆ Composite terms

 Lister Hill National Center for Biomedical Communications 38

Multiple terms for a class

- ◆ Synonymy

<ul style="list-style-type: none"> ▪ Left coronary artery ▪ LCA ▪ Arteria coronaria sinistra 	<ul style="list-style-type: none"> ▪ Addison's disease ▪ Primary adrenocortical insufficiency
---	---
- ◆ "Clinical synonymy" (vs. identity)

<ul style="list-style-type: none"> ▪ Abdominal swelling ▪ Swollen abdomen 	<ul style="list-style-type: none"> ▪ Addison's disease ▪ Primary adrenocortical insufficiency
vs. Waterhouse-Friderichsen Syndrome	

 Lister Hill National Center for Biomedical Communications 39

Multiple classes for a term

- ◆ Polysemy

Cold	}	<ul style="list-style-type: none"> ▪ Cold ▪ Common cold
	}	<ul style="list-style-type: none"> ▪ Cold ▪ Cold temperature
	}	<ul style="list-style-type: none"> ▪ COLD ▪ Chronic Obstructive Airway Disease
- ◆ Truncated terms

Calcium	}	<ul style="list-style-type: none"> ▪ Calcium ▪ Ca⁺⁺ ▪ [Coagulation factor IV]
	}	<ul style="list-style-type: none"> ▪ Calcium ▪ Calcium measurement

 Lister Hill National Center for Biomedical Communications 40

Non-ontological features in terms

- ◆ Epistemological features
 - Gallbladder calculus without mention of cholecystitis
 - Diarrhea of presumed infectious origin
 - Replacement of unspecified heart valve
 - ...

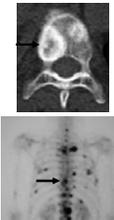
 Lister Hill National Center for Biomedical Communications 41

Ontology vs. Epistemology

<ul style="list-style-type: none"> ◆ Ontology <ul style="list-style-type: none"> • Invariants in reality <ul style="list-style-type: none"> ▪ Classes (universals) ▪ Relations between them • Theory of reality 	<ul style="list-style-type: none"> ◆ Epistemology <ul style="list-style-type: none"> • Knowledge about such entities • Perception of reality
--	--

Bone metastasis

}



Bone metastasis diagnosed by CT scan

Bone metastasis diagnosed by Tc99m bone scintiscan

 42

Composite terms

- ◆ **Sentence-like terms**
 - Several classes and their relations
 - May contain epistemological features
- Tuberculosis of adrenal glands, tubercle bacilli not found (in sputum) by microscopy, but found by bacterial culture




NLM Lister Hill National Center for Biomedical Communications 43

More composite terms

- Nontraffic accident involving being accidentally pushed from motor vehicle, except off-road motor vehicle, while in motion, not on public highway, driver of motor vehicle injured
- Determine whether the elder patient and caretaker have a functional social support network to assist the patient in performing activities of daily living and in obtaining health care, transportation, therapy, medications, community resource information, financial advice, and assistance with personal problems
- Telephone call by a physician to patient or for consultation or medical management or for coordinating medical management with other health care professionals (eg, nurses, therapists, social workers, nutritionists, physicians, pharmacists); complex or lengthy (eg, lengthy counseling session with anxious or distraught patient, detailed or prolonged discussion with family members regarding seriously ill patient, lengthy communication necessary to coordinate complex services of several different health professionals working on different

44

Terminological relations as a surrogate for ontological relations

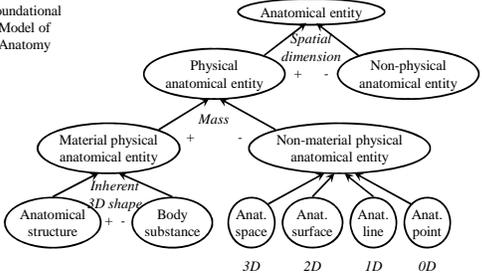
Issues

- ◆ Lack of explicit classificatory principle
- ◆ Underspecification of the relations
- ◆ Thesaurus relations
- ◆ Limited depth in hierarchies “by design”

NLM Lister Hill National Center for Biomedical Communications 46

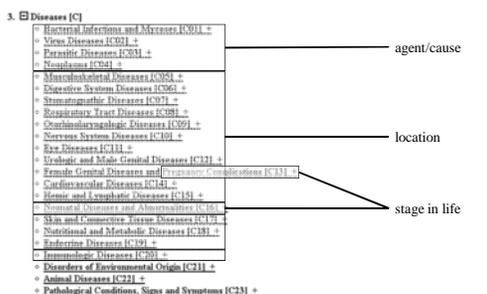
Explicit classificatory principle

Foundational Model of Anatomy



NLM Lister Hill National Center for Biomedical Communications 47

No explicit classificatory principle



NLM Lister Hill National Center for Biomedical Communications 48

1. Certain infectious and parasitic diseases
2. Neoplasms
3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4. Endocrine, nutritional, and metabolic diseases
5. Mental and behavioral disorders
6. Diseases of nervous system
7. Diseases of the eye and adnexa
8. Diseases of the ear and mastoid process
9. Diseases of circulatory system
10. Diseases of respiratory system
11. Diseases of digestive system
12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue
14. Diseases of the genitourinary system
15. Pregnancy, childbirth, and the puerperium
16. Certain conditions originating in the newborn (perinatal) period
17. Congenital malformations, deformations and chromosomal abnormalities
18. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
19. Injury, poisoning and certain other consequences of external causes
20. External causes of morbidity
21. Factors influencing health status and contact with health service



- Attribute
- Body structure
- Clinical finding
- Context-dependent categories
- Environments and geographical locations
- Events
- Observable entity
- Organism
- Pharmaceutical / biologic product
- Physical force
- Physical object
- Procedure
- Qualifier value
- Social context
- Special concept
- Specimen
- Staging and scales
- Substance



Lister Hill National Center for Biomedical Communications 50

Fully specified relations

Viral meningitis in SNOMED CT

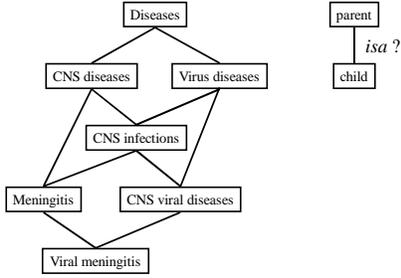
Fully defined by ...

- ⊊ [is a]
 - ⊊ ⊕ viral infections of the central nervous system
 - ⊊ ⊕ infective meningitis
 - ⊊ Causative agent
 - ⊊ ⊕ virus
 - ⊊ Group
 - ⊊ Associated morphology
 - ⊊ ⊕ inflammation
 - ⊊ Finding site
 - ⊊ ⊕ meninges structure



Lister Hill National Center for Biomedical Communications 51

Underspecification of the relations

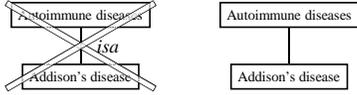



Lister Hill National Center for Biomedical Communications 52

Thesaurus relations

◆ Addison's disease

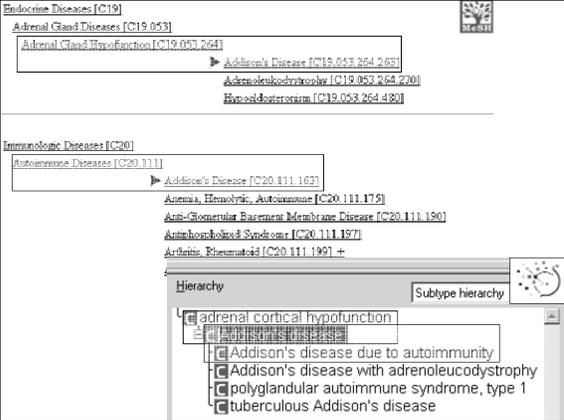
- Due to auto-immunity in 80% of the cases
- Other causes include tuberculosis



Relations used to create hierarchical structures vs. hierarchical relations



Lister Hill National Center for Biomedical Communications 53



Endocrine Diseases [C19]

Adrenal Gland Diseases [C19.053]

Adrenal Gland Hypofunction [C19.053.264]

► Addison's Disease [C19.053.264.263]

Adrenoleucodystrophy [C19.053.264.260]

Hyponatremism [C19.053.264.480]

Immunologic Diseases [C20]

Autoimmune Diseases [C20.111]

► Addison's Disease [C20.111.163]

Anemia, Hemolytic, Autoimmune [C20.111.175]

Anti-Glomerular Basement Membrane Disease [C20.111.190]

Antinuclear Antibody Syndrome [C20.111.197]

Arthritis, Rheumatoid [C20.111.199] +

Hierarchy Subtype hierarchy

- adrenal cortical hypofunction
 - Addison's disease due to autoimmunity
 - Addison's disease with adrenoleucodystrophy
 - polyglandular autoimmune syndrome, type 1
 - tuberculous Addison's disease

Accidents in MeSH

Environment and Public Health [G03]
 Public Health [G03.850]
 ▶ Accidents [G03.850.110]

Accident Prevention [G03.850.110.060] +

Accidental Falls [G03.850.110.085]

Accidents, Aviation [G03.850.110.185]

Accidents, Home [G03.850.110.205]

Accidents, Occupational [G03.850.110.250] +

Accidents, Radiation [G03.850.110.285]

Accidents, Traffic [G03.850.110.320]

Drowning [G03.850.110.500] +

Lister Hill National Center for Biomedical Communications 55

Limited depth in hierarchies “by design”

- ◆ Term identifier (code) used to record the position in the hierarchy
 - Limited number of digits available
 - May hide part of the structure
- ◆ Terminologies: ICD, SNOMED, ...

E84 Cystic fibrosis
 Includes: mucoviscidosis
E84.0 Cystic fibrosis with pulmonary manifestations
 Use additional code to identify any infectious organism present, such as: *Pseudomonas* (R06.5)
E84.1 Meconium ileus in cystic fibrosis
 Excludes1: meconium ileus not due to Cystic fibrosis (P75)
E84.2 Cystic fibrosis with gastrointestinal manifestations
 Excludes2: meconium ileus in cystic fibrosis (E84.1)
E84.5 Cystic fibrosis with other manifestations

Lister Hill National Center for Biomedical Communications 56

Cystic fibrosis in ICD

E84 Cystic fibrosis
 Includes: mucoviscidosis
E84.0 Cystic fibrosis with pulmonary manifestations
 Use additional code to identify any infectious organism present, such as: *Pseudomonas* (R06.5)
E84.1 Meconium ileus in cystic fibrosis
 Excludes1: meconium ileus not due to Cystic fibrosis (P75)
E84.2 Cystic fibrosis with gastrointestinal manifestations
 Excludes2: meconium ileus in cystic fibrosis (E84.1)
E84.5 Cystic fibrosis with other manifestations

Lister Hill National Center for Biomedical Communications 57

Conclusions

Conclusions ☹️

- ◆ Biomedical terms
 - reflect some aspects of biomedical reality
 - Although the primary concern of terminology is naming, not reflecting reality
 - often convey additional features (e.g., epistemology)
- ◆ Biomedical terminology tends to offset part of the complexity
 - but often reflects utility

Lister Hill National Center for Biomedical Communications 59

Conclusions 😊

- ◆ Biomedical terminologies can help populate biomedical ontologies
- ◆ Resources needed
 - Linguistic analysis of terms
 - Statistical analysis of terms in a corpus / annotation database (dependence relations)
 - Manual curation

Lister Hill National Center for Biomedical Communications 60

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration

 Lister Hill National Center for Biomedical Communications 61



Short course – Summer 2008
Biomedical Ontology in Practice

June 9, 2008 – Session #2

Design Principles, Formalisms and Tools for Biomedical Ontologies




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

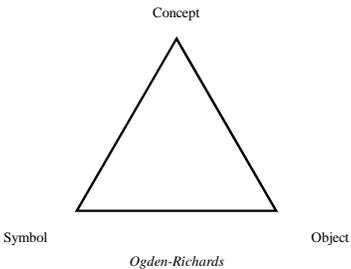
Overview

- ◆ Definitions
 - Ontologies vs. other artifacts
 - Kinds of ontologies
- ◆ Some principles of formal ontology
 - Top-level categories
 - Categories of relationships
- ◆ Formalisms and tools

 Lister Hill National Center for Biomedical Communications 63

Definitions

Introduction



Ogden-Richards

 Lister Hill National Center for Biomedical Communications 65

Definitions

- ◆ The *What* question
 - Objects in the world
 - With their properties
 - With their relations to other objects
 - Also: events, processes, and states
- ◆ Explicit specification of a conceptualization
 - Support software applications
- ◆ Domain ontology reflects
 - Underlying reality
 - Theory of the domain

 Lister Hill National Center for Biomedical Communications 66

Examples of use

- ◆ Natural language processing
- ◆ Access to heterogeneous sources of information (e.g., Semantic Web)
- ◆ Systems engineering

- ◆ Interoperability

 Lister Hill National Center for Biomedical Communications 67

Ontology vs. other artifacts

- ◆ Ontology
 - Defining types of things and their relations
- ◆ Terminology
 - Naming things in a domain
- ◆ Thesaurus
 - Organizing things for a given purpose
- ◆ Classification
 - Placing things into (arbitrary) classes
- ◆ Knowledge bases
 - Assertional knowledge

[Smith, KR-MED 2006]
 [Chute, JAMIA 2000]

 Lister Hill National Center for Biomedical Communications 68

(Controlled) Terminology

- ◆ Objective: naming things
- ◆ Example: Current Procedural Terminology (CPT)
- ◆ Shared understanding
 - Agreement on what terms to use
 - Utility-driven (arbitrary)

Telephone call by a physician to patient or for consultation or medical management or for coordinating medical management with other health care professionals (eg, nurses, therapists, social workers, nutritionists, physicians, pharmacists); complex or lengthy (eg, lengthy counseling session with anxious or distraught patient, detailed or prolonged discussion with family members regarding seriously ill patient, lengthy communication necessary to coordinate complex services of several different health professionals working on different

 Lister Hill National Center for Biomedical Communications 69

Thesaurus

- ◆ Objective: organize things for a purpose
 - e.g., information retrieval
 - Organization by relatedness
- ◆ Example: Medical Subject Headings (MeSH)
 - Indexing/retrieval of biomedical articles
- ◆ Relations used in hierarchies vs. hierarchical relations

 Lister Hill National Center for Biomedical Communications 70

Thesaurus vs. ontology

```

    graph TD
      AD[Autoimmune Diseases] -- "is generally a" --> Add[Addison's disease]
      Add --> TAD[Tuberculous Addison's disease]
      Add --> AddImm[Addison's disease due to autoimmunity]
      AD -.- X TAD
    
```

 Lister Hill National Center for Biomedical Communications 71

Classification

- ◆ Objective: placing things into classes
- ◆ Characteristics
 - Single inheritance (tree)
 - Idiosyncratic inclusion/exclusion criteria

E10 **Insulin-dependent diabetes mellitus**
 [See before E10 for subdivisions]
Includes: diabetes (mellitus):
 - brittle
 - juvenile-onset
 - ketosis-prone
 - type 1
Excludes: diabetes mellitus (m):
 - malnutrition-related (E12.-)
 - neonatal (E70.2)
 - pregnancy, childbirth and the puerperium (928.-)
 glycosuria:
 - NOS (E91)
 - renal (E74.8)
 impaired glucose tolerance (E73.0)
 postsurgical hypoparathyroidism (E85.1)

 Lister Hill National Center for Biomedical Communications 72

Classification

- ◆ **Characteristics (continued)**
 - Everything must be classified, including
 - When there is no specific slot (NEC)
 - When there is insufficient information (NOS)

E84 Cystic fibrosis
Includes: mucoviscidosis

E84.0 Cystic fibrosis with pulmonary manifestations

E84.1 Cystic fibrosis with intestinal manifestations
 Meconium ileus+ (P75*)
Excludes: meconium obstruction in cases where cystic fibrosis is known not to be present (P76.0)

E84.8 Cystic fibrosis with other manifestations
 Cystic fibrosis with combined manifestations

E84.9 Cystic fibrosis, unspecified

Lister Hill National Center for Biomedical Communications 73

Knowledge Bases

- ◆ **Objective:** represent knowledge needed for a given application
- ◆ **Example:** drug knowledge bases
- ◆ **Assertional knowledge**
 - Vs. definitional knowledge in ontologies
 - Often probabilistic
- ◆ **Examples of assertions**
 - Indications of a drug
 - Signs and symptoms of a disease

Lister Hill National Center for Biomedical Communications 74

Fuzzy borders

- ◆ **Some ontologies also collect names**
 - FMA
- ◆ **Some terminologies also provide formal definitions**
 - SNOMED CT
- ◆ **Some terminologies/ontologies include both definitional and assertional knowledge**
 - SNOMED CT

Lister Hill National Center for Biomedical Communications 75

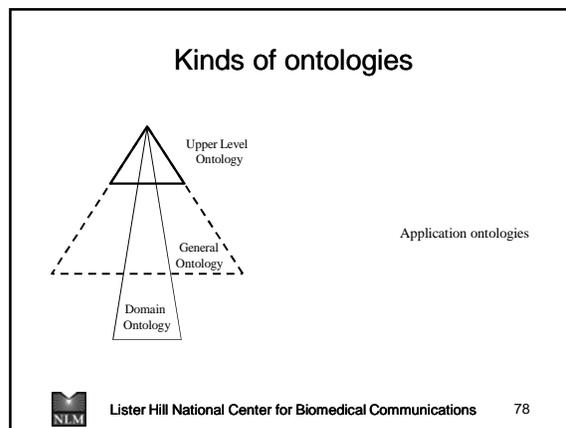
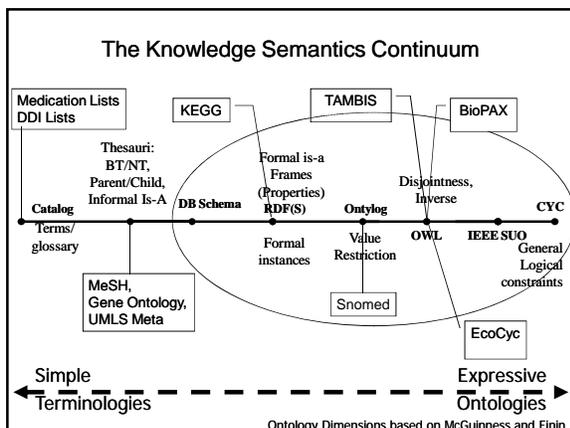
Types of resources

- ◆ **Lexical resources**
 - Collections of lexical items
 - Additional information
 - Part of speech
 - Spelling variants
 - Useful for entity recognition
 - UMLS SPECIALIST Lexicon, WordNet

- ◆ **Ontological resources**
 - Collections of
 - kinds of entities (substances, qualities, processes)
 - relations among them
 - Useful for relation extraction
 - UMLS Semantic Network, BioTop

▲

- ◆ **Terminological resources**
 - Collections lexical items + identifiers
 - Useful for entity resolution
 - UMLS Metathesaurus



Ontology-related issues

- ◆ Creation
- ◆ Merging
- ◆ Alignment
- ◆ Validation



Lister Hill National Center for Biomedical Communications 79

Formal Ontological Principles

Formal ontological distinctions

- ◆ Universal vs. individual
- ◆ Continuant vs. occurrent
- ◆ Independent vs. dependent



Lister Hill National Center for Biomedical Communications 81

Universal vs. Individual

<ul style="list-style-type: none"> ◆ Universal = <i>category</i> ◆ Synonyms <ul style="list-style-type: none"> • Kind, Type, (Class) ◆ Examples <ul style="list-style-type: none"> • eyeball • blood pressure • conference 	 <i>instantiation</i>	<ul style="list-style-type: none"> ◆ Individual = <i>instance</i> ◆ Synonyms <ul style="list-style-type: none"> • Particular, Token ◆ Examples <ul style="list-style-type: none"> • my right eyeball • my blood pressure (132/79) • AMIA Annual Symposium 2003
---	---	---



Lister Hill National Center for Biomedical Communications 82

Continuant vs. Occurrent

<ul style="list-style-type: none"> ◆ Continuant = <i>Continues to exist through time</i> ◆ Synonyms <ul style="list-style-type: none"> • Substance ◆ Examples <ul style="list-style-type: none"> • tennis racquet • mitochondrion • insulin production 	<ul style="list-style-type: none"> ◆ Occurrent = <i>Unfolds through time</i> ◆ Synonyms <ul style="list-style-type: none"> • Process ◆ Examples <ul style="list-style-type: none"> • tennis tournament • metabolism • producing insulin
---	--



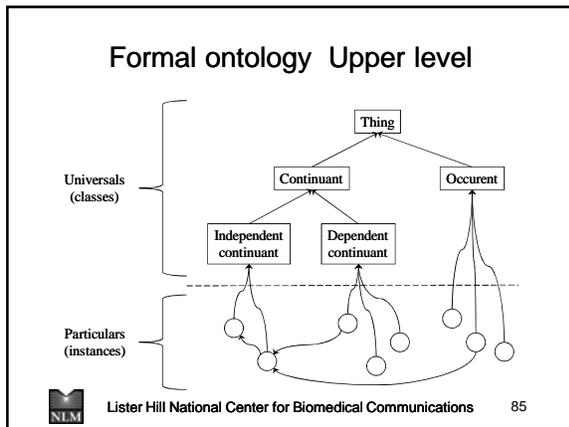
Lister Hill National Center for Biomedical Communications 83

Independent vs. Dependent

<ul style="list-style-type: none"> ◆ Independent = <i>Can exist without support from other entities</i> ◆ Examples <ul style="list-style-type: none"> • virus • molecule • plant 	<ul style="list-style-type: none"> ◆ Dependent = <i>Require support from other entities for its existence</i> ◆ Examples <ul style="list-style-type: none"> • viral infection • DNA binding • food
--	--



Lister Hill National Center for Biomedical Communications 84

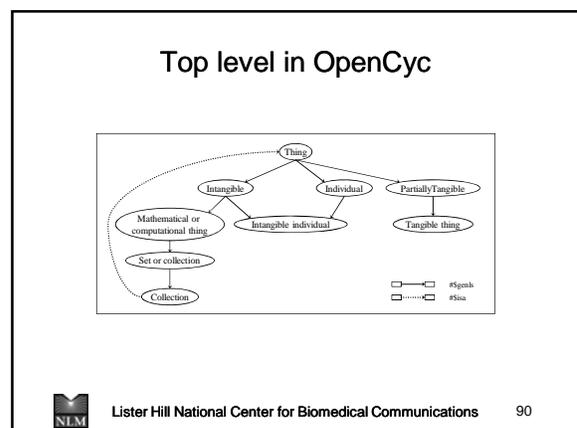


- ### Formal ontological distinctions
- ◆ Basic distinctions in many top-level ontologies
 - Generic: BFO, DOLCE
 - Biomedical: BioTop, UMLS Semantic Network
 - ◆ Condition the relations between various types of entities
 - Relations
 - Between instances (e.g., *part_of* [at time])
 - Between classes (e.g., *isa*, *part_of* [atemporal])
 - Between one instance and one class (*instance_of*)
- [Smith, Genome Biology 2005]
- Lister Hill National Center for Biomedical Communications 86

- ### Formal ontology in practice
- ◆ Provides foundational classes and relations
 - Upper level ontologies
 - Relation ontology
 - ◆ Provides a framework for analyzing entities and relations
- Lister Hill National Center for Biomedical Communications 87

Examples

- ### General ontologies
- ◆ OpenCyc
 - General ontology
 - Cycorp, Inc (D. Lenat & al.)
 - Over 1M hand-coded assertions
 - <http://www.opencyc.org>
 - ◆ WordNet
 - Electronic lexical database
 - Princeton University (G. Miller & al.)
 - Over 100,000 synsets
 - <http://wordnet.princeton.edu/>
- Lister Hill National Center for Biomedical Communications 89



Top level in WordNet

Abstraction
 Activity
 Entity
 Event
 Group
 Location
 Natural phenomenon
 Possession
 Psychological feature
 Shape
 State

Lister Hill National Center for Biomedical Communications 91

GALEN

- ◆ Generalised Architecture for Languages, Encyclopaedias, and Nomenclatures in Medicine
- ◆ European Union project (A. Rector & al.)
- ◆ Over 25,000 concepts (primitives)
- ◆ <http://www.opengalen.org>

Lister Hill National Center for Biomedical Communications 92

Top level in GALEN

Lister Hill National Center for Biomedical Communications 93

UMLS Semantic Network

- ◆ Definitional knowledge in the biomedical domain
- ◆ NLM (A. McCray & al.)
- ◆ Content
 - 135 semantic types
 - 54 types of relationship
 - 6700 semantic relations
- ◆ <http://semanticnetwork.nlm.nih.gov>

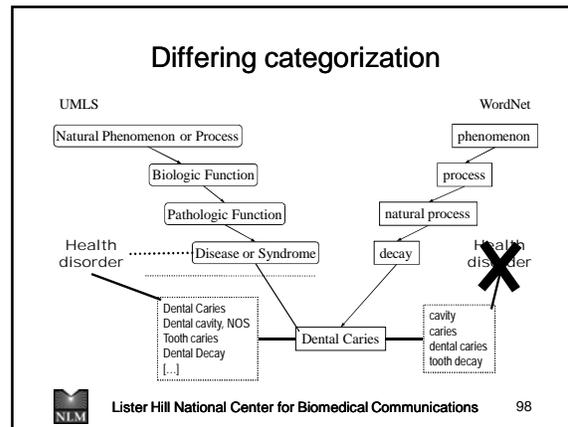
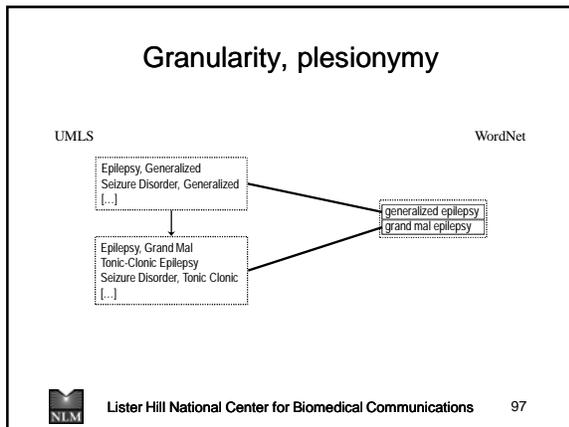
Lister Hill National Center for Biomedical Communications 94

Top level in the Semantic Network

Lister Hill National Center for Biomedical Communications 95

Differences between ontologies

Examples



Formalisms and Tools

- ### Ontology and Formalism
- ◆ Frames
 - ◆ Description logics
 - OWL DL
 - ◆ First-order logic

 - ◆ OBO Format
 - Conversion to OWL DL
- Lister Hill National Center for Biomedical Communications 100

- ### Tools for ontology developers
- ◆ Protégé
 - Publicly available
 - Frames and DL
 - Classifier
 - Supports many file formats (import/export)
 - Large community of users
 - Well supported
 - <http://protege.stanford.edu/>
 - ◆ OBO-Edit
 - Specific to the biomedical/OBO community
 - Simpler than Protégé (“tool for biologists”)
 - <http://oboedit.org/>
- <http://protege.stanford.edu/>
 The OBO Ontology Editor
- Lister Hill National Center for Biomedical Communications 101

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration

Lister Hill National Center for Biomedical Communications 102

 Short course – Summer 2008
Biomedical Ontology in Practice

June 9, 2008 – Session #3

“High-Impact” Biomedical Ontologies
A Structural Perspective

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

Overview

- ◆ Structural perspective [J. Cimino, YBMI 2006]
 - 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)

 Lister Hill National Center for Biomedical Communications 104

International Classification of Diseases



ICD Characteristics (1)

- ◆ Current version: ICD-10
- ◆ Type: Classification
- ◆ Domain: Disorders
- ◆ Developer: World Health Organization (WHO)
- ◆ Funding: WHO
- ◆ Availability
 - Publicly available: No
 - Repositories: UMLS [ICD9-CM in NCBO BioPortal]
- ◆ URL: <http://www.who.int/classifications/icd/en/>

 Lister Hill National Center for Biomedical Communications 106

ICD Characteristics (2)

- ◆ Number of
 - Concepts: 12,318
 - Terms: 1 per concept (tabular)
- ◆ Major organizing principles:
 - Tree (single inheritance hierarchy)
 - No explicit classification criteria
 - Idiosyncratic inclusion/exclusion mechanism
 - .8 slots for Not elsewhere classified (NEC)
 - .9 slots for Not otherwise specified (NOS)
- ◆ Formalism: Proprietary format

 Lister Hill National Center for Biomedical Communications 107

ICD Top level

Chapter	Blocks	Title
I	A00-B99	Certain infectious and parasitic diseases
II	C00-D49	Neoplasms
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00-E90	Endocrine, nutritional and metabolic diseases
V	F00-F99	Mental and behavioural disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and adnexa
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K93	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O99	Pregnancy, childbirth and the puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V01-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U99	Codes for special purposes

 Lister Hill National Center for Biomedical Communications 108

ICD Example

◆ Idiosyncratic inclusion/exclusion criteria

E10 **Insulin-dependent diabetes mellitus**
(See before E10 for subdivisions.)
Includes: diabetes (mellitus):
 - brittle
 - juvenile-onset
 - ketosis-prone
 - type 1
Excludes: diabetes mellitus (m):
 - malnutrition-related (E12.-)
 - neonatal (E70.2)
 - pregnancy, childbirth and the puerperium (O24.-)
 glycosuria:
 - renal (E75.8)
 - impaired glucose tolerance (E73.0)
 postsurgical hypoinsulinaemia (E09.1)

 Lister Hill National Center for Biomedical Communications 109

ICD Example

◆ Not elsewhere classified (NEC)
 ◆ Not otherwise specified (NOS)

E84 **Cystic fibrosis**
Includes: mucoviscidosis
E84.0 **Cystic fibrosis with pulmonary manifestations**
E84.1 **Cystic fibrosis with intestinal manifestations**
 Meconium ileus+ (P75*)
Excludes: meconium obstruction in cases where cystic fibrosis is known not to be present (P76.0)
E84.8 **Cystic fibrosis with other manifestations**
 Cystic fibrosis with combined manifestations
E84.9 **Cystic fibrosis, unspecified**

 Lister Hill National Center for Biomedical Communications 110

Logical Observation Identifiers, Names and Codes (LOINC)



LOINC Characteristics (1)

◆ Current version: 2.22 (Dec. 2007)
 ◆ Type: Controlled terminology*
 ◆ Domain: Laboratory and clinical observations
 ◆ Developer: Regenstrief Institute
 ◆ Funding: NLM
 ◆ Availability
 • Publicly available: Yes
 • Repositories: UMLS
 ◆ URL: www.regenstrief.org/loinc/loinc.htm

 Lister Hill National Center for Biomedical Communications 112

LOINC Characteristics (2)

◆ Number of
 • Concepts: 50k active codes (2.18)
 • Terms: n/a*

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

◆ Formalism: “DL-like”

 Lister Hill National Center for Biomedical Communications 113

LOINC Example

◆ *Sodium:SCnc:-Pt:Ser/Plas:Qn*
 [the molar concentration of sodium is measured in the plasma (or serum), with quantitative result]

Axis	Value
Component	Sodium
Property	SCnc – Substance Concentration (per volume)
Timing	Pt – Point in time (Random)
System	Ser/Plas – Serum or Plasma
Scale	Qn – Quantitative
Method	--

 Lister Hill National Center for Biomedical Communications 114

SNOMED Clinical Terms

SNOMED CT Characteristics (1)

- ◆ Current version: January 31, 2008 (2 annual releases)
- ◆ Type: Reference terminology / ontology
- ◆ Domain: Clinical medicine
- ◆ Developer: IHTSDO
- ◆ Funding: IHTSDO
- ◆ Availability
 - Publicly available: Yes* (in member countries)
 - Repositories: UMLS
- ◆ URL: <http://www.ihtsdo.org/>

Lister Hill National Center for Biomedical Communications 116

SNOMED CT Characteristics (2)

- ◆ Number of
 - Concepts: 311,313 active concepts (Jan. 31, 2008)
 - Terms: 794,061 active “descriptions”
- ◆ Major organizing principles:
 - Utility for clinical medicine (e.g., assertional + definitional knowledge)
 - Model of meaning (incomplete)
 - Rich set of associative relationships
 - Small proportion of defined concepts (many primitives)
- ◆ Formalism: Description logics (KRSS)

Lister Hill National Center for Biomedical Communications 117

SNOMED CT Top level

Lister Hill National Center for Biomedical Communications 118

SNOMED CT Example

Lister Hill National Center for Biomedical Communications 119

Foundational Model of Anatomy

Gene Ontology Characteristics (2)

- ◆ Number of
 - Concepts: 22,546 (Jan. 2, 2007)
 - Terms: 2.15 per concept
- ◆ Major organizing principles:
 - 3 major hierarchies
 - Molecular function
 - Cellular component
 - Biological process
 - Relations (within hierarchies): *isa, part_of, regulates*
 - No relations between concepts across hierarchies
- ◆ Formalism: OBO format



Lister Hill National Center for Biomedical Communications 127

Gene Ontology Top level (MF)

- ```

all : all [250418 gene products] &
 GO:0008150 : biological_process [166605 gene products]
 GO:0005775 : cellular_component [169814 gene products]
 GO:0003674 : molecular_function [168558 gene products] &
 GO:0016209 : antibody_activity [566 gene products]
 GO:0015457 : auxiliary_transport_protein_activity [161 gene products]
 GO:0005488 : binding [46697 gene products]
 GO:0003824 : catalytic_activity [51856 gene products]
 GO:0030188 : chaperone_regulator_activity [73 gene products]
 GO:0042056 : chemoattractant_activity [14 gene products]
 GO:0045499 : chemorepellent_activity [9 gene products]
 GO:0030234 : enzyme_regulator_activity [2370 gene products]
 GO:0016530 : metallochaperone_activity [47 gene products]
 GO:0060089 : molecular_transducer_activity [7873 gene products]
 GO:0003774 : motor_activity [527 gene products]
 GO:0045735 : nutrient_reservoir_activity [49 gene products]
 GO:0031386 : protein_tag [18 gene products]
 GO:0005198 : structural_molecule_activity [4324 gene products]
 GO:0030528 : transcription_regulator_activity [10429 gene products]
 GO:0045182 : translation_regulator_activity [893 gene products]
 GO:0005215 : transporter_activity [10583 gene products]

```



Lister Hill National Center for Biomedical Communications 128

### Gene Ontology Top level (CC)

- ```

all : all [250418 gene products] &
  GO:0008150 : biological_process [166605 gene products]
  GO:0005775 : cellular_component [169814 gene products] &
    GO:0005623 : cell [111066 gene products]
    GO:0044464 : cell_part [111049 gene products]
    GO:0031975 : envelope [3316 gene products]
    GO:0031012 : extracellular_matrix [573 gene products]
    GO:0044420 : extracellular_matrix_part [292 gene products]
    GO:0005576 : extracellular_region [5001 gene products]
    GO:0044421 : extracellular_region_part [3365 gene products]
    GO:0032991 : macromolecular_complex [14668 gene products]
    GO:0031974 : membrane_endothelial_lumen [5290 gene products]
    GO:0043226 : organelle [79653 gene products]
    GO:0044422 : organelle_part [16645 gene products]
    GO:0055044 : synplast [3 gene products]
    GO:0045202 : synapse [454 gene products]
    GO:0044856 : synapse_part [210 gene products]
    GO:0019012 : viron [227 gene products]
    GO:0044423 : viron_part [186 gene products]
    GO:0003674 : molecular_function [168558 gene products]
    
```



Lister Hill National Center for Biomedical Communications 129

Gene Ontology Top level (BP)

- ```

all : all [250418 gene products] &
 GO:0008150 : biological_process [166605 gene products] &
 GO:0022610 : biological_adhesion [1586 gene products]
 GO:0065007 : biological_regulation [31031 gene products]
 GO:0001906 : cell_killing [177 gene products]
 GO:0009987 : cellular_process [79087 gene products]
 GO:0032502 : developmental_process [13678 gene products]
 GO:0051234 : establishment_of_localization [15270 gene products]
 GO:0040007 : growth [4139 gene products]
 GO:0002376 : immune_system_process [2517 gene products]
 GO:0051179 : localization [17811 gene products]
 GO:0040011 : locomotion [1251 gene products]
 GO:0008152 : metabolic_process [61127 gene products]
 GO:0051704 : multi-organism_process [4780 gene products]
 GO:0032501 : multiorganismal_process [20567 gene products]
 GO:0048519 : negative_regulation_of_biological_process [5037 gene products]
 GO:0043473 : pigmentation [235 gene products]
 GO:0048518 : positive_regulation_of_biological_process [6585 gene products]
 GO:0050789 : regulation_of_biological_process [28645 gene products]
 GO:0000003 : reproduction [6343 gene products]
 GO:0022414 : reproductive_process [3525 gene products]
 GO:0050896 : response_to_stimulus [16487 gene products]
 GO:0048511 : rhythmic_process [404 gene products]
 GO:0016032 : viral_reproduction [536 gene products]

```



Lister Hill National Center for Biomedical Communications 130

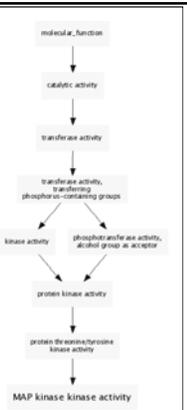
### Gene Ontology Ex

- ```

all : all [250418 gene products]
  GO:0003674 : molecular_function [168558 gene products]
  GO:0003824 : catalytic_activity [51856 gene products]
  GO:0016740 : transferase_activity [15763 gene products]
  GO:0016772 : transferase_activity_transferring_phosphor-containing_groups
  GO:0016301 : kinase_activity [6093 gene products]
  GO:0004672 : protein_kinase_activity [3504 gene products]
  GO:0004712 : protein_serine/threonine/tyrosine_kinase_activity
  GO:0004708 : MAP_kinase_kinase_activity
  GO:0016773 : phosphotransferase_activity_alcohol
  GO:0004672 : protein_kinase_activity [3504 gene products]
  GO:0004712 : protein_serine/threonine/tyrosine_kinase_activity
  GO:0004708 : MAP_kinase_kinase_activity
    
```



Lister Hill National Center for Biomedical



RxNorm

RxNorm Characteristics (1)

- ◆ Current version: June 2, 2007 (monthly releases)
- ◆ Type: Controlled terminology
- ◆ Domain: Drug names
- ◆ Developer: NLM
- ◆ Funding: NLM
- ◆ Availability
 - Publicly available: Yes*
 - Repositories: UMLS
- ◆ URL: <http://www.nlm.nih.gov/research/umls/rxnorm/>



Lister Hill National Center for Biomedical Communications 133

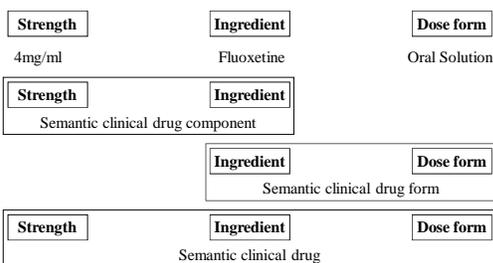
RxNorm Characteristics (2)

- ◆ Number of
 - Concepts: 93k
 - Terms: 105k
- ◆ Major organizing principles:
 - Generic vs. brand
 - Combinations of Ingredient / Form / Dose
 - No hierarchical structure
 - Links to all major US drug information sources
 - No clinical information
- ◆ Formalism: UMLS RRF format



Lister Hill National Center for Biomedical Communications 134

RxNorm Normalized form



Lister Hill National Center for Biomedical Communications 135

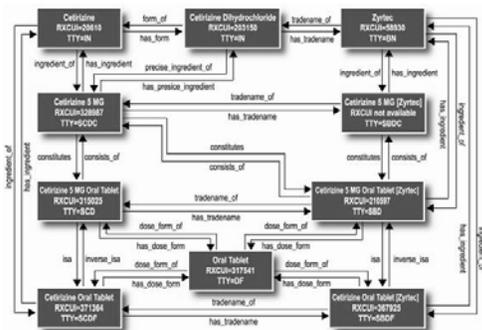
Rx Norm Generic vs. Brand

- ◆ Generic
 - Ingredient (IN)
 - Clinical drug form (SCDF)
 - Clinical drug component (SCDC)
 - Clinical drug (SCD)
 - ◆ Brand
 - Brand name (BN)
 - Branded drug form (SBDF)
 - Branded drug component (SBDC)
 - Branded drug (SBD)
- tradenname_of*



Lister Hill National Center for Biomedical Communications 136

RxNorm Relations among drug entities



Medical Subject Headings (MeSH)



NCI thesaurus Characteristics (1)

- ◆ Current version: 08.04d (~monthly releases)
- ◆ Type: Controlled terminology / ontology
- ◆ Domain: Cancer
- ◆ Developer: NCI Center for Bioinformatics
- ◆ Funding: NCI
- ◆ Availability
 - Publicly available: Yes
 - Repositories: UMLS / OBO / NCBO BioPortal
- ◆ URL: <http://nciterns.nci.nih.gov/>


Lister Hill National Center for Biomedical Communications
145

NCI thesaurus Characteristics (2)

- ◆ Number of
 - Concepts: 58,868 (2007_05E)
 - Terms: 2.68 per concept
- ◆ Major organizing principles:
 - Subsumption hierarchy
 - Rich set of associative relationships
 - Small proportion of defined concepts (many primitives)
 - Links to many external resources
- ◆ Formalism: OWL Lite


Lister Hill National Center for Biomedical Communications
146

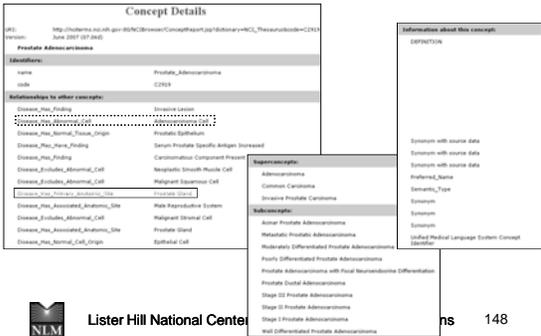
NCI thesaurus Top level

NCI_Thesaurus Taxonomy

- Abnormal Cell
- Activity
- Anatomic Structure, System, or Substance
- Biochemical Pathway
- Biological Process
- Chemotherapy Regimen or Agent Combination
- Conceptual Entity
- Diagnostic, Therapeutic, and Research Equipment
- Diagnostic or Prognostic Factor
- Disease, Disorder or Finding
- Drug, Food, Chemical or Biomedical Material
- Experimental Organism Anatomical Concept
- Experimental Organism Diagnosis
- Gene
- Gene Product
- Molecular Abnormality
- NCI Administrative Concept
- Organism
- Property or Attribute
- Retired Concept


Lister Hill National Center for Biomedical Communications
147

NCI thesaurus Example




Lister Hill National Center for Biomedical Communications
148

Unified Medical Language System (UMLS)



UMLS Characteristics (1)

- ◆ Current version: 2008AA (2-3 annual releases)
- ◆ Type: Terminology integration system
- ◆ Domain: Biomedicine
- ◆ Developer: NLM
- ◆ Funding: NLM (intramural)
- ◆ Availability
 - Publicly available: Yes* (cost-free license required)
 - Repositories: UMLS
- ◆ URL: <http://umlsks.nlm.nih.gov/>


Lister Hill National Center for Biomedical Communications
150

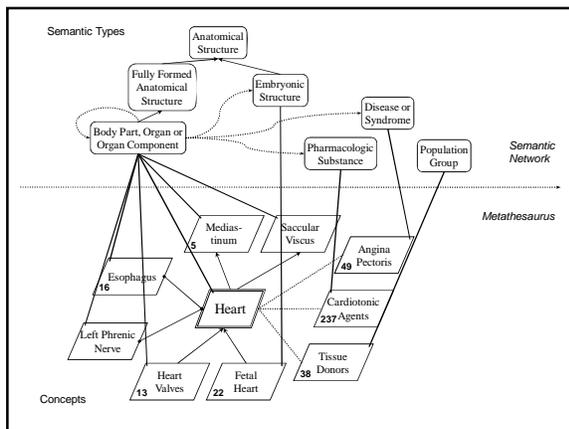
UMLS Characteristics (2)

- ◆ **Number of**
 - Concepts: 1.5M (2008AA)
 - Terms: ~6M
- ◆ **Major organizing principles (Metathesaurus):**
 - Concept orientation
 - Source transparency
 - Multi-lingual through translation
- ◆ **Formalism: Proprietary format (RRF)**

Lister Hill National Center for Biomedical Communications 151

UMLS Integrating subdomains

Lister Hill National Center for Biomedical Communications 152



Addison's Disease: Concept

Lister Hill National Center for Biomedical Communications 154

Metathesaurus Concepts (2007AB)

- ◆ **Concept (~ 1.4M) CUI**
 - Set of synonymous concept names
- ◆ **Term (~ 5.3 M) LUI**
 - Set of normalized names
- ◆ **String (~ 5.9M) SUI**
 - Distinct concept name
- ◆ **Atom (~ 7.2M) AUI**
 - Concept name in a given source

A0066000	Headache (MeSH)	A0065992	Headache (ICD-10)	
				S0046854
A0066007	Headaches (MedDRA)	A12003304	Headaches (OMIM)	
				S0046855
				L0018681
A0540936	Cephalodynia (MeSH)	S0475647		
				L0380797
				C0018681

Lister Hill National Center for Biomedical Communications 155

Recap

Name	Scope	# concepts	Median	Subs. Hier	Version
SNOMED CT	Clinical medicine (patient records)	310,314	2	yes	July 31, 2007
LOINC	Clinical observations and laboratory tests	46,406	3	no	Version 2.21 (no "natural language" names)
FMA	Human anatomical structures	~72,000	?	yes	(not yet in the UMLS)
Gene Ontology	Functional annotation of gene products	22,546	1	yes	Jan. 2, 2007
RxNorm	Standard names for prescription drugs	93,426	1	no	Aug. 31, 2007
NCI Thesaurus	Cancer research, clinical care, public information	58,868	2	yes	2007_05E
ICD-10	Diseases and conditions (health statistics)	12,318	1	no	1998 (tabular)
MeSH	Biomedicine (descriptors for indexing the literature)	24,767	5	no	Aug. 27, 2007
UMLS	Terminology integration in the life sciences	1.4 M	2	n/a	2007AC (English only)

Lister Hill National Center for Biomedical Communications [Bodenreider, YBMI 2008]

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration

 Lister Hill National Center for Biomedical Communications 157



Short course – Summer 2008
Biomedical Ontology in Practice

June 9, 2008 – Session #4

Biomedical Ontologies in Action

A Functional Perspective on Biomedical Ontologies




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

Overview

- ◆ Functional perspective [Bodenreider, YBMI 2008]
 - 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 reasoning (data selection and aggregation, decision support, natural language processing applications, knowledge discovery).

 Lister Hill National Center for Biomedical Communications 159

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
 - Exceptions: GALEN, LOINC
- ◆ Not all surface forms represented
 - Often insufficient for NLP applications
 - Large variation in number of terms per concept across ontologies

 Lister Hill National Center for Biomedical Communications 162

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)



 Lister Hill National Center for Biomedical Communications 163

GO Annotations for Aldh2 (mouse)

GO Annotations in Tabular Form (Text View) (GO Graph) 

Category	Classification Term	Evidence
Molecular Function	aldehyde dehydrogenase (NAD) activity	IEA
Molecular Function	oxidoreductase activity	IEA
Molecular Function	oxidoreductase activity	IEA
Cellular Component	mitochondrion	IDA
Biological Process	metabolic process	IEA
Biological Process	oxidation reduction	IEA

<http://www.informatics.jax.org/>

 Lister Hill National Center for Biomedical Communications 164

GO ALD4 in Yeast

GO Annotations

Manually curated

Molecular Function

Manually curated

Biological Process

Manually curated

Cellular Component

Manually curated

High-throughput

All ALD4 GO evidence and references

View Computational GO annotations for ALD4

Molecular Function

- aldehyde dehydrogenase (NAD) activity (IDA, IMP, ISS)
- aldehyde dehydrogenase [NAD(P)+] activity (IDA)

Biological Process

- ethanol metabolic process (IMP)

Cellular Component

- mitochondrial nucleoid (IDA)
- mitochondrion (IMP, ISS)
- mitochondrion (IDA)

<http://db.yeastgenome.org/>

 Lister Hill National Center for Biomedical Communications 165

GO Annotations for ALDH2 (Human)

GO ID	Function	Interpro	IEA	IPPO1	UniProt
GO:0013451	oxidoreductase activity	IrIrrpro	IEA	IPPO15590	UniProt_9803
GO:0013451	oxidoreductase activity	IrIrrpro	IEA	IPPO13180	UniProt_9803
GO:0013451	oxidoreductase activity	IrIrrpro	IEA	IPPO13162	UniProt_9803
GO:0013451	oxidoreductase activity	IrIrrpro	IEA	IPPO13161	UniProt_9803
GO:0013451	oxidoreductase activity	IrIrrpro	IEA	IPPO13161	UniProt_9803
GO:0004029	aldehyde dehydrogenase (NAD) activity	IrIrrpro	IEA	K6A-0563	UniProt_9803
GO:0004029	aldehyde dehydrogenase (NAD) activity	IrIrrpro	TAS	130E115	PINC_9603
GO:0004030	aldehyde dehydrogenase [NAD(P)+] activity	IrIrrpro	TAS	8981321	PINC_9603
GO:0003055	electron carrier activity	IrIrrpro	TAS	8981321	UniProt_9803
GO:0004029	aldehyde dehydrogenase (NAD) activity	Enzyme	IEA	1.2.1.3	UniProt_9803

<http://www.ebi.ac.uk/GOA/>

 Lister Hill National Center for Biomedical Communications 166

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
 - Automatic indexing systems



 Lister Hill National Center for Biomedical Communications 167

MeSH MEDLINE indexing

Abstracts 2008 Jun;106(6):1813-9 Related Articles, Links

Full Text Search Alerts

Free cortisol in sepsis and septic shock.

Brundel S, Karlsson S, Penttilä V, Lehto P, Vuorjoki M, Ruokonen E. Finnsepsis Study Group.

▶ [Cochrane Review](#)
 Department of Intensive Care, Kuopio University Hospital, PL 16222 Kuopio, Finland. Stepani.Brendel@kuh.fi

BACKGROUND: Severe sepsis activates the hypothalamic-pituitary 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 to 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 electrochemoluminescence immunoassay. The Coatless 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 (723 ± 395 nmol/L vs 793 ± 439 nmol/L, $P = 0.44$). Non-survivors 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
- Biological Markers/blood
- Female
- Finland/epidemiology
- Hospital Mortality
- Humans
- Hydrocortisone/blood*
- Hydrocortisone/therapeutic use
- Kaplan-Meiers Estimate

- Male
- Predictive Value of Tests
- Prospective Studies
- Sepsis/blood*
- Sepsis/drug therapy
- Susceptibility
- Severity of Illness Index
- Shock, Septic/blood*
- Shock, Septic/drug therapy
- Shock, Septic/mortality
- Treatment Outcome

Substances:

- Biological Markers
- Hydrocortisone

Lister Hill National Center for Biomedical Communications 169

MeSH MEDLINE indexing

Expert Opin Investig Drugs, 2008 Apr;17(4):497-509

Expert Opinion

Replacement therapy for Addison's disease: recent developments.

Leivis K. Hasekke ES

University of Bergen, Institute of Medicine, Section of Endocrinology, 5021 Bergen, Norway.
Knutan.lovas@helse-bergen.no

BACKGROUND: The hormone deficiencies in Addison's disease (primary adrenal insufficiency) are conventionally treated with oral glucocorticoid and mineralocorticoid replacement but the available therapies do not restore the physiological hormone levels and biorythm. Despite such treatment these patients self-report impaired health-related quality of life (HRQL) and recent research has indicated increased mortality. **OBJECTIVE/METHODS:** We review the literature and recent developments in replacement therapy. **RESULTS/CONCLUSION:** Patients with Addison's disease require mineralocorticoid replacement, i.e., fludrocortisone 0.05 - 0.20 mg once daily. Starting doses of glucocorticoids should be 15 - 20 mg for hydrocortisone or 20 - 30 mg for cortisone acetate, divided into two or three doses, and preferably weight-adjusted. There are indications that the systemic glucocorticoids have undesirable metabolic long-term effects, which make them less suitable as first-line treatment. Time-release hydrocortisone tablets and continuous subcutaneous hydrocortisone infusion are promising new treatment modalities. Studies of replacement with the adrenal androgen dehydroepiandrosterone (DHEA) in adrenal failure have shown inconsistent benefit on HRQL. DHEA, or possibly testosterone replacement is likely to be beneficial for selected groups of patients with Addison's disease but this remains to be shown. We here give our opinion of the best treatment and future direction of research in this area.

Lister Hill National Center for Biomedical Communications 172

MeSH MEDLINE indexing

MeSH Terms

- Addison Disease/blood
- Addison Disease/drug therapy*
- Androgens/administration & dosage*
- Androgens/therapeutic use
- Dosage Forms
- Drug Administration Routes
- Drug Administration Schedule
- Glucocorticoids/administration & dosage*
- Glucocorticoids/adverse effects
- Glucocorticoids/blood
- Glucocorticoids/deficiency
- Hormone Replacement Therapy*
- Humans
- Mineralocorticoids/administration & dosage*
- Mineralocorticoids/adverse effects
- Mineralocorticoids/blood
- Mineralocorticoids/deficiency
- Quality of Life
- Treatment Outcome

Substances:

- Androgens
- Dosage Forms
- Glucocorticoids
- Mineralocorticoids

Lister Hill National Center for Biomedical Communications 171

ICD9-CM Coding clinical data

- ◆ ICD9-CM
 - Used for coding clinical data e.g., for billing purposes
- ◆ Other uses of ICD
 - Morbidity and mortality reporting worldwide



Lister Hill National Center for Biomedical Communications 172

Knowledge management

Accessing biomedical information

Resources for biomedical search engines

- ◆ Synonyms
- ◆ Hierarchical relations
- ◆ High-level categorization
- ◆ Co-occurrence information
- ◆ Translation




Lister Hill National Center for Biomedical Communications 174

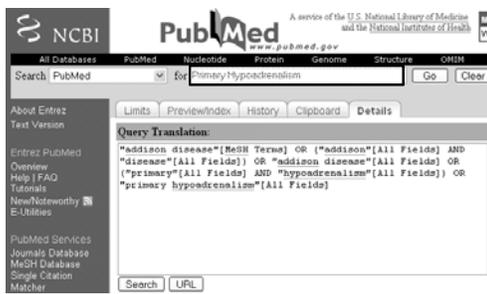
MeSH "synonyms" MEDLINE retrieval

- ◆ MeSH entry terms
 - Used as equivalent terms for retrieval purposes
 - Not always synonymous
- ◆ Increase recall without hurting precision

MeSH Heading	Addison Disease
Entry Term	Addison's Disease
Entry Term	Primary Adrenal Insufficiency
Entry Term	Primary Adrenocortical Insufficiency

 Lister Hill National Center for Biomedical Communications 175

MeSH "synonyms" MEDLINE retrieval



 Lister Hill National Center for Biomedical Communications 176

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 Addison disease

```

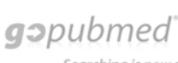
    graph TD
      A[Adrenal insufficiency] --> B[Addison disease]
    
```

 Lister Hill National Center for Biomedical Communications 177

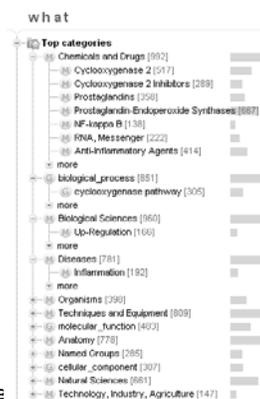


 Lister Hill National Center for Biomedical Communications 178

Co-indexing

 Searching is now sorted!
<http://www.gpubmed.com/>

cox-2 →



 Lister Hill National Center for E

Knowledge management

Mapping across biomedical ontologies

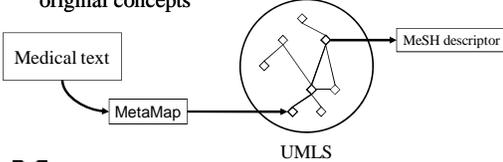
Reusing information

- ◆ Clinical information coded with SNOMED CT
 - Mapped to ICD9-CM and CPT for billing purposes
 - Mapped to ICD-O for epidemiology purposes
- ◆ Existing mapping tables crated by terminology developers as an incentive to use SNOMED CT

 Lister Hill National Center for Biomedical Communications 181

Reusing tools

- ◆ For noun phrases extracted from medical texts, map to UMLS concepts [Aronson & al., AMIA, 2000]
- ◆ Then, select from the MeSH vocabulary the concepts that are the most closely related to the original concepts



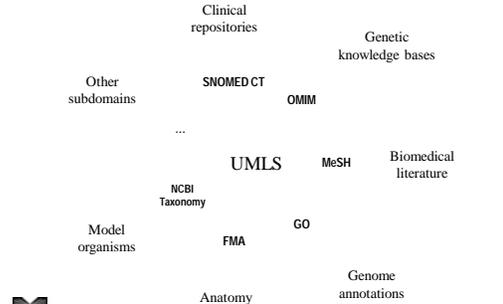
 Lister Hill National Center for Biomedical Communications 182

Terminology integration systems

- ◆ Terminology integration systems (UMLS, RxNorm) help bridge across vocabularies
- ◆ Uses
 - Information integration
 - Ontology alignment
 - Medication reconciliation

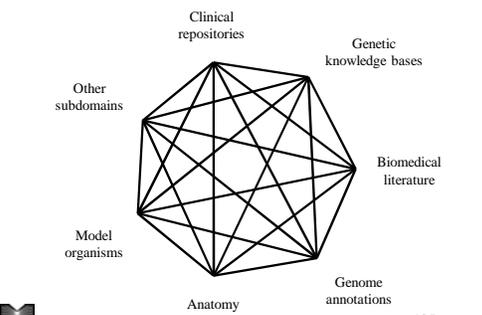
 Lister Hill National Center for Biomedical Communications 183

Integrating subdomains



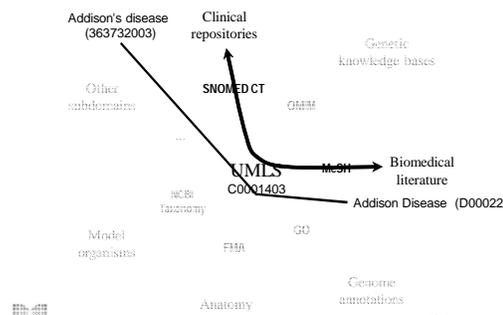
 Lister Hill National Center for Biomedical Communications 184

Integrating subdomains



 Lister Hill National Center for Biomedical Communications 185

Trans-namespace integration



 Lister Hill National Center for Biomedical Communications 186

Data integration, exchange and semantic interoperability

Data integration, exchange and semantic interoperability
Information exchange and semantic operability

“Standards”

- ◆ Ontologies help standardize patients data
 - Facilitate the exchange of data across institutions
 - Help connect “islands of data” (silos)
- ◆ LOINC
 - Exchange of laboratory data
 - In conjunction with HL7 messaging

 Lister Hill National Center for Biomedical Communications 189

Semantic interoperability projects BRIDG

- ◆ Biomedical Research Integrated Domain Group
 - Information model for clinical research
 - Interoperability between clinical trials information systems
 - Ontologies provide value sets to the information model

 Lister Hill National Center for Biomedical Communications 190

Semantic interoperability projects CDA

- ◆ Clinical Document Architecture (CDA R2)
 - Formal representation of clinical statements
 - Clinical observations
 - Medication administration
 - Adverse events
 - Associate an information model (HL7 RIM) with terminologies (LOINC, SNOMED CT, RxNorm)

 Lister Hill National Center for Biomedical Communications 191

Semantic interoperability projects caCORE

- ◆ Cancer Common Ontologic Representation Environment
 - Infrastructure developed to support an interoperable biomedical information system for cancer research
 - Uses the NCI Thesaurus as a component

 Lister Hill National Center for Biomedical Communications 192

Data integration, exchange and semantic interoperability

Information and data integration

Approaches to data integration

◆ **Warehousing**

- Sources to be integrated are transformed into a common format and converted to a common vocabulary
- Normalization through ontologies (e.g., GO annotations)

◆ **Mediation**

- Local schema (of the sources)
- Global schema (in reference to which the queries are made)
- Ontologies help define the global schema and map between local and global schemas (OntoFusion, ARIANE)

Lister Hill National Center for Biomedical Communications 194

Ontologies and integration

- ◆ Terminology integration systems help bridge across terminologies and the domains they represent
- ◆ Mappings across ontologies enable the integration of namespaces in the Semantic Web

Lister Hill National Center for Biomedical Communications 195

Trans-namespace integration

Lister Hill National Center for Biomedical Communications#196 196

Decision support and reasoning

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

Lister Hill National Center for Biomedical Communications 198

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



Lister Hill National Center for Biomedical Communications 199

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 NCI)



Lister Hill National Center for Biomedical Communications 200

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



Lister Hill National Center for Biomedical Communications 201

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)



Lister Hill National Center for Biomedical Communications 202

Barriers to usability of biomedical ontologies

Availability

- ◆ Many ontologies are freely available
- ◆ The UMLS is freely available for research purposes
 - Cost-free license required
- ◆ Licensing issues can be tricky
 - SNOMED CT is freely available in member countries of the IHTSDO
- ◆ Being freely available
 - Is a requirement for the Open Biomedical Ontologies (OBO)
 - Is a de facto prerequisite for Semantic Web applications



Lister Hill National Center for Biomedical Communications 204

Discoverability

- ◆ **Ontology repositories**
 - UMLS: 143 source vocabularies (biased towards healthcare applications)
 - NCBO BioPortal: ~100 ontologies (biased towards biological applications)
 - Limited overlap between the two repositories
- ◆ **Need for discovery services**



Lister Hill National Center for Biomedical Communications 205

Formalism

- ◆ **Several major formalism**
 - Web Ontology Language (OWL) – NCI Thesaurus
 - OBO format – most OBO ontologies
 - UMLS Rich Release Format (RRF) – UMLS, RxNorm
- ◆ **Conversion mechanisms**
 - OBO to OWL
 - LexGrid (import/export to LexGrid internal format)



Lister Hill National Center for Biomedical Communications 206

Ontology integration

- ◆ **Post hoc integration, from the bottom up**
 - UMLS approach
 - Integrates ontologies “as is”, including legacy ontologies
 - Facilitates the integration of the corresponding datasets
- ◆ **Coordinated development of ontologies**
 - OBO Foundry approach
 - Ensures consistency *ab initio*
 - Excludes legacy ontologies



Lister Hill National Center for Biomedical Communications 207

Quality

- ◆ **Quality assurance in ontologies is still imperfectly defined**
 - Difficult to define outside a use case or application
- ◆ **Several approaches to evaluating quality**
 - Collaboratively, by users (Web 2.0 approach)
 - Marginal notes enabled by BioPortal
 - Centrally, by experts
 - OBO Foundry approach
- ◆ **Important factors besides quality**
 - Governance
 - Installed base / Community of practice



Lister Hill National Center for Biomedical Communications 208

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration



Lister Hill National Center for Biomedical Communications 209



Short course – Summer 2008
Biomedical Ontology in Practice

June 10, 2008 – Session #1

Interfaces to Biomedical Ontologies



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

Overview

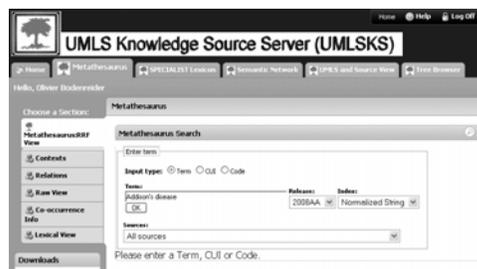
- ◆ Graphical interfaces
 - UMLS Knowledge Source Server
 - NCBO BioPortal
 - NCI Thesaurus
 - MeSH browser
 - Foundational Model of Anatomy (FMA) Explorer
 - Gene Ontology AmiGO
 - ICD-10 online
 - RxNav (RxNorm)
 - [...]

- ◆ Application Programming Interfaces

 Lister Hill National Center for Biomedical Communications 211

Graphical interfaces

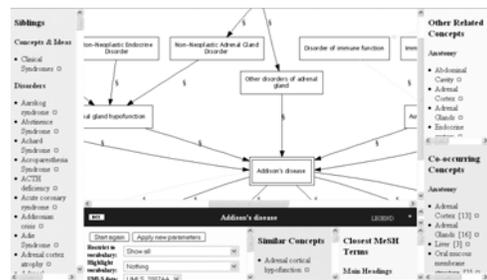
UMLS Knowledge Source Server



<http://umlsks.nlm.nih.gov/>

 Lister Hill National Center for Biomedical Communications 213

UMLS Semantic Navigator



<http://mor.nlm.nih.gov/perl/semnav.pl>

 Lister Hill National Center for Biomedical Communications 214

NCBO BioPortal

Name	Format	Current Version/Control Location	Action
African Traditional Medicine	OBO	1.0.1 NCBO Library	[Download] [Visualize] [Search]
Amino Acid	OWL Full	1.2 NCBO Library	[Download] [Visualize] [Search]
Amphibian gross anatomy	OBO	1.7 NCBO Library	[Download] [Visualize] [Search]
Animal natural history and life history	Protégé	See Remote Site Remote	[Download] [Visualize] [Search]
Basic Vertebrate Anatomy	OWL Full	1.1 NCBO Library	[Download] [Visualize] [Search]
Biological imaging methods	OBO	1.1 NCBO Library	[Download] [Visualize] [Search]
Biological process	OBO	1.208 NCBO Library	[Download] [Visualize] [Search]
Biomedical Resource Ontology	OWL Lite	1.1 NCBO Library	[Download] [Visualize] [Search]
BP2NL.es	OWL DL	1.3.1 NCBO Library	[Download] [Visualize] [Search]

<http://www.bioontology.org/tools/portal/bioportal.html>

 Lister Hill National Center for Biomedical Communications 215

NCI Thesaurus (EVS Server)

<http://ncit.nci.nih.gov/NCIBrowser/SearchConcept.do>

 Lister Hill National Center for Biomedical Communications 216

ICD-10

<http://www.who.int/classifications/apps/icd/icd10online/>
 Lister Hill National Center for Biomedical Communications 223

RxNav (RxNorm)

<http://mor.nlm.nih.gov/download/rxnav/>

Lister Hill National Center for Biomedical Communications 224

Application Programming Interfaces

Application Programming Interface

- ◆ Expose resources in such a way that they can be integrated in programs
 - Programming “against” a resource
- ◆ Standard protocols for communication
 - Web services (SOAP, REST)
- ◆ Standard libraries for programming
- ◆ Focus on content, not message

Lister Hill National Center for Biomedical Communications 226

UMLSKS Web Service API

- ◆ UMLSKS <http://umlsks.nlm.nih.gov/>
 - Developer's Guide > Webservice Operations
- ◆ WSDL available
- ◆ API give access to all 3 knowledge sources
- ◆ Licensing issues
 - Granting ticket and Single-use tickets

Lister Hill National Center for Biomedical Communications 227

UMLSKS Web Service API Example

```

ConceptIdGroup findCUIByNormString
(ConceptIdNormStringRequest request);
    
```

Argument: **ConceptIdNormStringRequest**

This class contains the arguments that further restrict the behavior of the call.

```

setCuiTicket (String #)
- Single-use ticket returned by the AuthorizationPort webservice
setRelease (String #)
- UMLS release of interest
setSearchString (String #)
- Input search string
setTable (String[] array)
- set of source abbreviations to search
setLanguage (String #)
- language restriction
setIncludeSuppressed (Boolean #)
- true if suppressed strings are included in the search
setCUI (long #)
- CUI flag for the content view to search
    
```

Return: **ConceptIdGroup**

Lister Hill National Center for Biomedical Communications 228

Other APIs to terminology systems

- ◆ NCBO BioPortal
http://www.bioontology.org/docs/bioportal/development/web_services.html
- ◆ OLS - Ontology Lookup Service
<http://www.ebi.ac.uk/ontology-lookup/WSDLDocumentation.do>
- ◆ RxNorm
<http://mor.nlm.nih.gov/download/rxnav/RxNormAPI.html>



Lister Hill National Center for Biomedical Communications 229

Applications based on WS APIs

- ◆ UMLSKS API
 - UMLSKS
<http://umlsks.nlm.nih.gov/>
- ◆ RxNorm API
 - RxNav
<http://mor.nlm.nih.gov/download/rxnav/rxnav.jnlp>
 - MyMedicationList
<http://mml.nlm.nih.gov/MyMedicationList.jnlp>



Lister Hill National Center for Biomedical Communications 230

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration



Lister Hill National Center for Biomedical Communications 231



Short course – Summer 2008
Biomedical Ontology in Practice

June 10, 2008 – Session #2

Searching and Analyzing Biomedical Concepts




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

Exercise 1

- ◆ What are the Clinical Drug Components for Zyrtec? (RxNav)



Lister Hill National Center for Biomedical Communications 233

Exercise 2

- ◆ What are the parts of the Aorta? (FMA)



Lister Hill National Center for Biomedical Communications 234

Exercise 3

- ◆ What are the parents of Hodgkin's disease in SNOMED CT?
 - Try SNOMEDCTID: 118599009
- ◆ What is its associated morphology?



Lister Hill National Center for Biomedical Communications 235

Exercise 4

- ◆ What are the various meanings of IL-2? (UMLS)



Lister Hill National Center for Biomedical Communications 236

Exercise 5

- ◆ What are the pharmacologic actions of Zyrtec? (MeSH)



Lister Hill National Center for Biomedical Communications 237

Exercise 6

- ◆ What are some synonyms for Schwannoma? (NCI Thesaurus)



Lister Hill National Center for Biomedical Communications 238

Solutions

Exercise 1

- ◆ What are the Clinical Drug Components for Zyrtec? (RxNav)

The screenshot shows the RxNav interface with a search for 'Zyrtec'. The results are displayed in a hierarchical tree structure. The root node is 'Ingredient' (Cetirizine), which has a child 'Ingredient Variant' (Cetirizine Hydrochloride). This variant has a child 'Brand Name' (Zyrtec). The 'Ingredient' node also has a child 'Clinical Drug Component' (Cetirizine 1 MQLM, Cetirizine 10 MQLM, Cetirizine 4 MQLM). The 'Clinical Drug Component' node has a child 'Branded Drug Component' (Cetirizine 1 MQLM, Purified, Cetirizine 10 MQLM, Purified, Cetirizine 4 MQLM, Purified). The interface includes a search bar at the top and a navigation pane on the left.



Lister Hill National Center for Biomedical Communications 240

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration

 Lister Hill National Center for Biomedical Communications 247



Short course – Summer 2008
Biomedical Ontology in Practice

June 10, 2008 – Session #3 / June 11, 2008 – Session #1

Contrasting and Critiquing Biomedical Ontologies





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

Exercise #1

- ◆ Hodgkin’s disease
 - NCI Thesaurus
 - SNOMED CT

 Lister Hill National Center for Biomedical Communications 249

Exercise #2

- ◆ Prostate
 - FMA
 - SNOMED CT

 Lister Hill National Center for Biomedical Communications 250

Exercise #3

- ◆ Cetirizine
 - MeSH
 - SNOMED CT

 Lister Hill National Center for Biomedical Communications 251

Solutions

Solutions

Exercise #1

Exercise #1

- ◆ **Hodgkin’s disease**
 - NCI Thesaurus
 - Using the NCI browser (EVS)
 - <http://nciterns.nci.nih.gov/>
 - SNOMED CT
 - Using the online browser from U. Sydney
 - <http://www.cs.usyd.edu.au/~hltru/scf/A3.cgi>

 Lister Hill National Center for Biomedical Communications 254

Hodgkin’s disease in NCIt (1)

URI: http://nciterns.nci.nih.gov/BD/NCITBrowser/ConceptReport.jsp?dictionary=NCI_Thesaurus&code=C9357
Version: April 2009 (08.043)

Hodgkin Lymphoma

Identifiers:

name	Hodgkin_s_Lymphoma
code	C9357

Relationships to other concepts:

Disease_Has_Primary_Anatomic_Site	Hematopoietic and Lymphatic System
Disease_Has_Normal_Tissue_Origin	Lymphoid Tissue
Disease_Excludes_Normal_Cell_Origin	Myeloid Cell
Disease_Excludes_Normal_Cell_Origin	Plasma Cell
Disease_Has_Abnormal_Cell	Reed-Sternberg Cell
Disease_Has_Associated_Anatomic_Site	Hematopoietic and Lymphatic System
Disease_Has_Normal_Cell_Origin	Mature Lymphocyte
Disease_Has_Primary_Anatomic_Site	Lymphatic System

Superconcepts

- Common Hematopoietic Neoplasm
- Lymphoma

 Lister Hill National Center for Biomedical Communications 255

Hodgkin’s disease in NCIt (1)

Information about this concept:

ALT_DEFINITION

NCI-GLOSSIA malignant disease of the lymphatic system that is characterized by painless enlargement of lymph nodes, the spleen, or other lymphatic tissue. It is sometimes accompanied by symptoms such as fever, weight loss, fatigue, and night sweats.

DEFINITION

NCIIA lymphoma, previously known as Hodgkin’s disease, characterized by the presence of Reed-Sternberg cells. There are two distinct subtypes: nodular lymphocyte predominant Hodgkin lymphoma and classical Hodgkin lymphoma. Hodgkin lymphoma has a bimodal age distribution, and involves primarily lymph nodes. Current therapy for Hodgkin lymphoma has resulted in an excellent outcome and cure for the majority of patients.

ICD-O-3_Code

9450/3

Preferred_Name

Hodgkin Lymphoma

Semantic_Type

Neoplastic Process

Synonym

HL

Synonym

Hodgkin Lymphoma

Synonym

Hodgkin’s Disease

Synonym

Hodgkin’s Lymphoma

Synonym

Hodgkin’s disease

Unified Medical Language System Concept Identifier

C0033829

Comments on Hodgkin’s disease in NCIt (1)

- ◆ **Search term: “Hodgkin’s disease”**
 - Not found, although “Hodgkin’s disease” is listed as a synonym
 - Search on “hodgkin”, select “Hodgkin lymphoma”
- ◆ **Parent classes**
 - Common hematopoietic neoplasm
 - Not an ontological category
 - Would be better represented through an associative relation (e.g., along the lines of “has_prevalence high prevalence”)
 - *Isa* overloading

 Lister Hill National Center for Biomedical Communications 257

Comments on Hodgkin’s disease in NCIt (2)

- ◆ **Associative relations**
 - For cancers, anatomy and morphology are foundational relations
 - Here
 - Anatomy : *Disease_Has_Primary_Anatomic_Site* Hematopoietic and Lymphatic System
 - Morphology: not directly represented (indirectly through *Disease_Has_Normal_Cell_Origin* Mature Lymphocyte)

 Lister Hill National Center for Biomedical Communications 258

Hodgkin's disease in SNOMED CT (1)

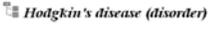
CONCEPT		
Concept ID	Fully Specified Name	Concept Status
118590009	Hodgkin's disease (disorder)	Current (0)

PARENTS	
Concept ID	FSN for Parent Concept (This Concept)
1 118800007	↳ Malignant lymphoma (disorder)

ATTRIBUTES		
Concept ID	FSN for Target Concept	Relationship Type
1 209520004	↳ Episodicities (qualifier value)	↳ Episodicity (attribute)
2 128930002	↳ Hodgkin lymphoma - category (morphologic abnormality)	↳ Associated morphology (attribute)

 Lister Hill National Center for Biomedical Communications 259

Hodgkin's disease in SNOMED CT (2)



CONCEPT			
Concept ID	Fully Specified Name	Concept Status	CTV3ID
118590009	Hodgkin's disease (disorder)	Current (0)	B61..

DESCRIPTIONS and SYNONYMS			
Description ID	Term	Description Status	Description Type
1 177017015	Hodgkin's disease (clinical)	Current (0)	Preferred (1)
2 1220409010	Malignant Hodgkin's lymphoma	Current (0)	Synonym (2)
3 1220408019	HD - Hodgkin's disease	Current (0)	Synonym (2)

 Lister Hill National Center for Biomedical Communications 260

Comments on Hodgkin's disease in SNOMED CT (1)

- ◆ Search term: "Hodgkin's disease"
 - Not found, although "Hodgkin's disease" is listed as a synonym
 - Search result: "Hodgkin lymphoma, nodular sclerosis, grade 1 (morphologic abnormality)"
 - Search on "lymphoma", navigate down from "Malignant lymphoma"
 - "hodgkin's disease" is ambiguous
 - Hodgkin lymphoma, no ICD-O subtype (morphologic abnormality)
 - Hodgkin's disease (disorder)
 - "Malignant lymphoma, Hodgkin's"
 - NB: lymphoma is always malignant
- ◆ Parent classes
 - Malignant lymphoma (clinical) [OK]

 Lister Hill National Center for Biomedical Communications 261

Comments on Hodgkin's disease in SNOMED CT (2)

- ◆ Associative relations
 - For cancers, anatomy and morphology are foundational relations
 - Here
 - Anatomy : not directly represented (indirectly through descendant concepts, e.g., Hodgkin's disease of intrathoracic lymph nodes)
 - Morphology: *Associated morphology* Hodgkin lymphoma - category

 Lister Hill National Center for Biomedical Communications 262

Hodgkin's disease NCI vs. SNOMED CT (1)

- ◆ Shared synonyms: NCI 1/2, SNOMED CT 1/3
 - Hodgkin's disease
- ◆ Shared relations
 - *Isa*
 - NCI: Lymphoma
 - Definition: "malignant (clonal) proliferation of B-lymphocytes or T-lymphocytes which involves the lymph nodes, bone marrow and/or extranodal sites. This category includes Non-Hodgkin lymphomas and Hodgkin lymphomas."
 - SNOMED CT: Malignant lymphoma
 - Same UMLS concept (CUI: C0024299)

 Lister Hill National Center for Biomedical Communications 263

Hodgkin's disease NCI vs. SNOMED CT (2)

- ◆ Shared relations: Associative relations
 - Anatomy
 - In NCI, but not in SNOMED CT
 - Morphology
 - In SNOMED CT, but not in NCI
 - Only indirectly, though cell type
 - Cell type
 - Only in NCI

 Lister Hill National Center for Biomedical Communications 264

Solutions

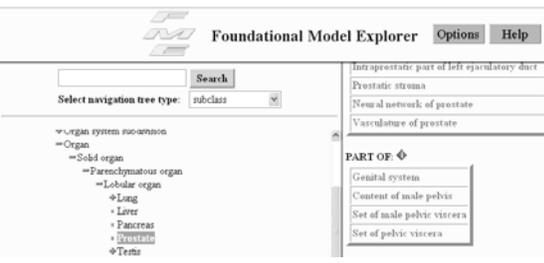
Exercise #2

Exercise #2

- ◆ Prostate
 - FMA
 - Using the Foundational Model Explorer
<http://siq.biostr.washington.edu/projects/fm/FME/>
 - SNOMED CT
 - Using the online browser from U. Sydney
<http://www.cs.usyd.edu.au/~hlru/scf/A3.cgi>

 Lister Hill National Center for Biomedical Communications 266

Prostate in FMA (1)



 Lister Hill National Center for Biomedical Communications 267

Prostate in FMA (2)



 Lister Hill National Center for Biomedical Communications 268

Comments on Prostate in FMA

- ◆ No synonyms in English
 - Latin and Spanish synonyms
- ◆ Hierarchies
 - *Isa*: Lobular organ
 - *Part_of*: Set of pelvic viscera
- ◆ Associative relations
 - *Lymphatic drainage*
 - No spatial relations

 Lister Hill National Center for Biomedical Communications 269

Prostate in SNOMED CT (1)

435 results found for prostate:

#	Concept ID	Fully Specified Name	Preferred Terms and Synonyms
1	9713002	Prostatitis (disorder)	Inflammation of prostate- Prostatitis [P]- Prostatitis, NOS
2	11441004	Prostatism (disorder)	Prostatism [P]- Prostatism, NOS
3	41216001	Prostatic structure (body structure)	Prostatic structure [P]- Prostate- Prostate, NOS
4	181422007	Entire prostate (body structure)	Entire prostate [P]- Prostate

CONCEPT

Concept ID	Fully Specified Name
181422007	Entire prostate (body structure)

DESCRIPTIONS AND SYNONYMS

Description ID	Term
1 280451017	Entire prostate
2 280452012	Prostate

PARENTS

Concept ID	FSH for Parent Concept (This Concept IS A)
1 310536002	Male internal genital organ (body structure)
2 41216001	Prostatic structure (body structure)
3 300443000	Entire male genital organ (body structure)

 Lister Hill National Center for Biomedical Communications 270

Prostate in SNOMED CT (2)

ATTRIBUTES		
Concept ID	FSN for Target Concept	Relationship Type
1 118760003	↳ Entire viscus (body structure)	↳ Part of (attribute)
2 245461005	↳ Entire urinary tract (body structure)	↳ Part of (attribute)
3 362265004	↳ Entire male internal genitalia (body structure)	↳ Part of (attribute)
4 362267007	↳ Entire lower male genitourinary tract (body structure)	↳ Part of (attribute)
5 362717004	↳ Entire minor pelvis (body structure)	↳ Part of (attribute)
6 362206001	↳ Entire lower genitourinary tract (body structure)	↳ Part of (attribute)
7 361340001	↳ Entire male genital system (body structure)	↳ Part of (attribute)
8 302553009	↳ Entire abdomen (body structure)	↳ Part of (attribute)



Lister Hill National Center for Biomedical Communications 271

Comments on Prostate in SNOMED CT

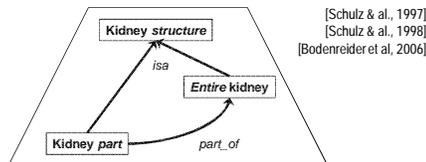
- ◆ “Ambiguous” term
 - Entire prostate
 - Prostatic structure
- ◆ Structure-Entire-Part representation of anatomical entities in SNOMED CT
 - Reification of *part_of*
 - Enables mereological inference through *isa* hierarchy
 - Not intuitive



Lister Hill National Center for Biomedical Communications 272

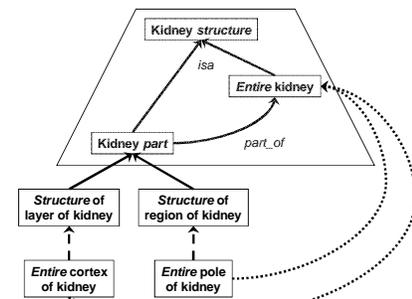
Structure-Entire-Part (SEP) triples

- ◆ S – The entity or any of its parts
- ◆ E – The entire anatomical entity
- ◆ P – Any parts of the anatomical entity



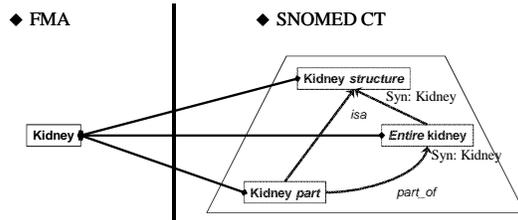
Lister Hill National Center for Biomedical Communications 273

Mereological inference through isa



Lister Hill National Center for Biomedical Communications 274

FMA mapping goes to Entire



Lister Hill National Center for Biomedical Communications 275

Prostate FMA vs. SNOMED CT

- ◆ Shared synonyms: FMA 1/1, SNOMED CT 1/2
 - Prostate
- ◆ Shared relations
 - *Isa*: no
 - FMA
 - Lobular organ
 - SNOMED CT
 - Prostatic structure
 - Male internal genital organ
 - Entire male genital organ



Lister Hill National Center for Biomedical Communications 276

Prostate FMA vs. SNOMED CT

◆ Shared relations

- Part of: almost
 - FMA
 - Genital system
 - Content of male pelvis
 - Set of male pelvic viscera
 - Set of pelvic viscera
 - SNOMED CT
 - Entire minor pelvis
 - Entire male genital system
 - ...

Lister Hill National Center for Biomedical Communications 277

Solutions

Exercise #3

Exercise #3

◆ Cetirizine

- MeSH
 - Using the MeSH browser <http://www.nlm.nih.gov/mesh/MBrowser.html>
- SNOMED CT
 - Using the online browser from U. Sydney <http://www.cs.usyd.edu.au/~hlru/sct/A3.cgi>

Lister Hill National Center for Biomedical Communications 279

Cetirizine in MeSH (1)

Entry Term	2-(4-((4-Chlorophenyl)phenylmethyl)-1-piperazinyl)ethoxy)acetic Acid
Entry Term	Alerisim
Entry Term	Alud Brand of Cetirizine Dihydrochloride
Entry Term	Alpharma Brand of Cetirizine Dihydrochloride
Entry Term	AWD pharma Brand of Cetirizine Dihydrochloride
Entry Term	Ampharma Brand of Cetirizine Dihydrochloride
Entry Term	Banco Brand of Cetirizine Dihydrochloride
Entry Term	Cetaleg

• • •

Entry Term	Volvic
Entry Term	Wolf Brand of Cetirizine Dihydrochloride
Entry Term	Worwig Brand of Cetirizine Dihydrochloride
Entry Term	Zetr
Entry Term	Zartek
Entry Term	Zytec

Cetirizine in MeSH (2)

Heterocyclic Compounds [D03]
 Heterocyclic Compounds, 1-Ring [D03.383]
 Piperazines [D03.383.606]
 Hydroxyzine [D03.383.606.515] ▶ Cetirizine [D03.383.606.515.200]

Pharm. Action [Anti-Allergic Agents](#)
 Pharm. Action [Histamine H1 Antagonists, Non-Sedating](#)

Lister Hill National Center for Biomedical Communications 281

Comments on Cetirizine in MeSH

- ◆ 45 entry terms
 - Various generic and brand names
 - Chemical formula
 - Code (P-071)
- ◆ Hierarchy
 - *Isa*: Piperazines [chemistry]
- ◆ Pharmacologic action
 - Anti-Allergic Agents
 - Histamine H1 Antagonists, Non-Sedating

Lister Hill National Center for Biomedical Communications 282

Cetirizine in SNOMED CT (1)

15 results found for cetirizine:

Previous Next

#	Concept ID	Fully Specified Name	Preferred Terms and Synonyms
1	108655000	% Cetirizine (product)	Cetirizine (P)
2	372523007	% Cetirizine (substance)	Cetirizine (P)

Cetirizine (substance)

Concept ID	Fully Specified Name	Concept Status	CTV3ID	SNOMED ID	Is Primitive
372523007	Cetirizine (substance)	Current (C)	XJVVUJ	F-61523	Primitive (T)

DESCRIPTIONS and SYNONYMS

Description ID	Term	Description Status	Description Type	Language Code	Initial Capital Status	
1	1211057019	Cetirizine	Current (C)	Preferred (T)	en	Capitalization meaningless (C)

PARENTS

Concept ID	FSN for Parent Concept (This Concept IS A)
1	372624008 % Non-sedating antihistamine (substance)

CHILDREN

Concept ID	FSN for Child Concept
1	108655004 % Cetirizine hydrochloride (substance)

Cetirizine in SNOMED CT (2)

Concept ID	Fully Specified Name	Concept Status	CTV3ID
108655000	Cetirizine (product)	Current (C)	y01Dq

DESCRIPTIONS and SYNONYMS

Description ID	Term	Description Status	Description Type	
1	173189012	Cetirizine	Current (C)	Preferred (T)

PARENTS

Concept ID	FSN for Parent Concept (This Concept IS A)
1	349956006 % Non-sedating antihistamine (product)

ATTRIBUTES

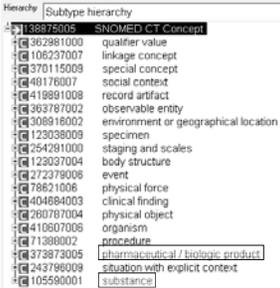
Concept ID	FSN for Target Concept	Relationship Type
1	372523007	% Cetirizine (substance) % Has active ingredient (attribute) (Cetirizine hydrochloride)

CHILDREN

Concept ID	FSN for Child Concept
1	320818006 % Cetirizine hydrochloride 10mg tablet (product)
2	320820009 % Cetirizine hydrochloride 1mg/mL, oral liquid (product)
3	371746005 % Cetirizine hydrochloride 5mg tablet (product)
4	375571002 % Cetirizine hydrochloride 5mg tablet (product)
5	375572009 % Cetirizine hydrochloride 10mg tablet (product)
6	375573004 % Cetirizine hydrochloride 5mg/5 mL syrup (product)
7	400482001 % Cetirizine hydrochloride + pseudoephedrine hydrochloride (product)
8	409491005 % Cetirizine hydrochloride 5mg chewable tablet (product)
9	409492003 % Cetirizine hydrochloride 10mg chewable tablet (product)

Comments on Cetirizine in SNOMED CT

- ◆ **Ambiguous term**
 - Cetirizine (product)
 - Cetirizine (substance)
- ◆ **Hierarchy**
 - *Isa*: Non-sedating antihistamine (substance) [pharmacologic action]
- ◆ **No associative relations**



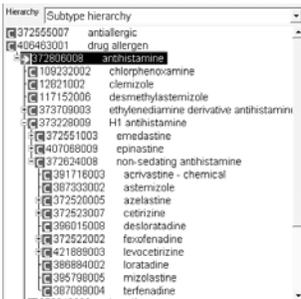
Lister Hill National Center for Biomedical Communications 285

Cetirizine MeSH vs. SNOMED CT (1)

- ◆ **Shared synonyms: MeSH 1/45, SNOMED CT 1/1**
 - Cetirizine
- ◆ **Shared relations: none**
 - MeSH:
 - *Isa*: <chemistry>
 - *Associative*: <pharmacologic action>
 - SNOMED CT
 - *Isa*: < pharmacologic action>
 - *Associative*: none

Lister Hill National Center for Biomedical Communications 286

Cetirizine MeSH vs. SNOMED CT (2)



Lister Hill National Center for Biomedical Communications 287

Summary

- ◆ **Differing representations**
 - Not necessarily inconsistent
 - Consistency may be difficult to assess automatically
- ◆ **Often due to idiosyncratic representation in one ontology**
- ◆ **Hindrance to ontology alignment and evaluation methods relying on shared relations**

Lister Hill National Center for Biomedical Communications 288

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration

 Lister Hill National Center for Biomedical Communications 289



Short course – Summer 2008
Biomedical Ontology in Practice

June 9, 2008 – Session #2

Extending Biomedical Ontologies



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

Overview

- ◆ Corpus terminology
- ◆ Identify terms in biomedical text (in reference to the UMLS)
- ◆ Identify additional terms
- ◆ Place these terms in UMLS hierarchies

[Bodenreider, ACL 2002]

 Lister Hill National Center for Biomedical Communications 291

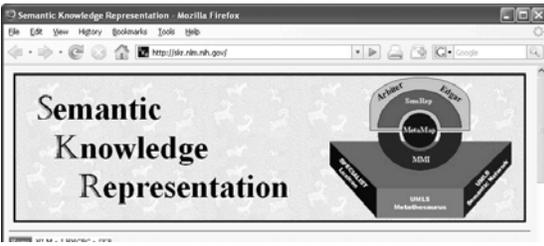
Tiny corpus One MEDLINE abstract

Full Text
Free cortisol in sepsis and septic shock. PMID: 18499615

Brendel S, Kuitonen S, Penttilä V, Lehto P, Vuorinen M, Ruuskanen E. Finreopsis Study Group.
Department of Intensive Care, Kuopio University Hospital, PL 16222 Kuopio, Finland. Stjepan.Brendel@kuh.fi

BACKGROUND: Severe sepsis activates the hypothalamic-pituitary 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 to 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 electrochemoluminescence immunoassay. The CoviLens 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 23%. 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, $P = 0.44$). Non-survivors 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:** Classically, 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.

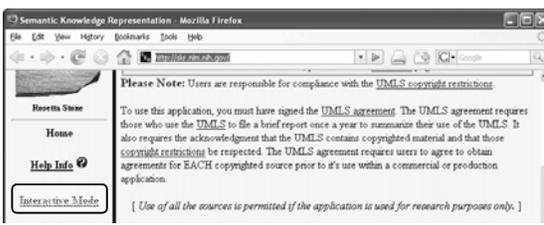
Identify UMLS concepts with MetaMap



<http://skr.nlm.nih.gov/>

 Lister Hill National Center for Biomedical Communications 293

Interactive mode



Please Note: Users are responsible for compliance with the UMLS copyright restrictions.

To use this application, you must have signed the UMLS agreement. The UMLS agreement requires those who use the UMLS to file a brief report once a year to summarize their use of the UMLS. It also requires the acknowledgment that the UMLS contains copyrighted material and that those copyright restrictions be respected. The UMLS agreement requires users to agree to obtain agreements for EACH copyrighted source prior to its use within a commercial or production application.

[Use of all the sources is permitted if the application is used for research purposes only.]

 Lister Hill National Center for Biomedical Communications 294

Interactive MetaMap

Interactive Mode

Please NOTE:
The Interactive mode is only intended for the testing of the various program and their options.

Lister Hill National Center for Biomedical Communications 295

Paste abstract

Interactive MetaMap

Users are responsible for compliance with the [UMI-2 copyright restrictions](#)

Test to be Processed:

BACKGROUND: Severe sepsis activates the hypothalamopituitary axis, increasing cortisol production. In some studies, hydrocortisone substitution based on an adrenocorticotrophic hormone-stimulation test or baseline cortisol measurements 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 Coxsone method was used for

Lister Hill National Center for Biomedical Communications 296

Select options

Lister Hill National Center for Biomedical Communications 297

Run MetaMap

Lister Hill National Center for Biomedical Communications 298

Output

```

Processing 00000000.nx.1: BACKGROUND: Severe sepsis activates the hypothalamopituitary axis, increasing cortisol production.

Phrases: "Severe sepsis"
>>>> Phrases
severe sepsis
<<<<< Phrases
>>>> Candidates
Meta Candidates (8):
1000 C1719472:Severe Sepsis [Disease or Syndrome]
861 C0036690:Sepsis (Septicemia) [Disease or Syndrome]
861 C0243026:Sepsis (Systemic Infection) [Disease or Syndrome]
861 C1090821:Sepsis [Invertebrate]
789 C0333346:Sepsis [Functional Concept]
694 C0205082:Severe [Qualitative Concept]
694 C1249279:SEVERE (Severe Adverse Event) [Finding]
694 C1561581:Severe (Allergy Severity - Severe) [Finding]
<<<<< Candidates
>>>> Mappings
Meta Mapping (1000):
1000 C1719472:Severe Sepsis [Disease or Syndrome]
<<<<< Mappings
    
```

Lister Hill National Center for Biomedical Communications 299

Suggest term candidates

- ◆ Not recognized by MetaMap at all
- ◆ Partially identified by MetaMap
- ◆ Missing terms in a concept

Lister Hill National Center for Biomedical Communications 300

Suggest placement in UMLS

- ◆ Use a browser
- ◆ Identify close parent
- ◆ Examine its children
- ◆ Assess placement by comparing with potential siblings

 Lister Hill National Center for Biomedical Communications 301

Possible new terms (1)

- ◆ Hypothalamopituitary axis
 - Concept exists: C0678897, but missing exact (neoclassical) synonym
 - hypothalamic pituitary axis
 - hypothalamus hypophysis axis
 - hypothalamus-pituitary axis
- ◆ American College of Chest Physicians
 - Similar to other American Colleges (e.g., American College of Physicians ())
 - Integrate as a child of Professional Organization or Group (C1522486)
 - NB: instance, cannot be a child of ACP

 Lister Hill National Center for Biomedical Communications 302

Possible new terms (2)

- ◆ Free cortisol
 - Identified as a substance (C0443476), not a laboratory procedure / test result
 - Cortisol, free measurement (C0236401)
- ◆ Coolens method
 - Missing term / concept
 - Method for estimating (not measuring directly) the free fraction of cortisol

 Lister Hill National Center for Biomedical Communications 303

Possible new terms (3)

- ◆ Electrochemiluminescence immunoassay
 - Missing concept
 - Create as a child of Chemiluminescence assay (C0201709)
- ◆ Nonsurvivors
 - Survivors exists as a concept (C0206194)
 - Create as a child of Patients (C0030705)

 Lister Hill National Center for Biomedical Communications 304

Agenda

Monday, June 9	Introduction to Biomedical Ontologies	Design Principles, Formalisms and Tools for Biomedical Ontologies	Biomedical Ontologies - Content and structure - Function
Tuesday, June 10	Interfaces to Biomedical Ontologies	Searching and Analyzing Biomedical Concepts	Contrasting Biomedical Ontologies
Wednesday, June 11	Critical Analysis of Biomedical Ontologies	Extending Biomedical Ontologies	Using Biomedical Ontologies for Data Integration

 Lister Hill National Center for Biomedical Communications 305



Short course – Summer 2008
Biomedical Ontology in Practice

June 11, 2008 – Session #3

Using Biomedical Ontologies for Data Integration



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

Overview

- ◆ Motivation
- ◆ Some practical considerations and issues
 - Integration approaches
 - Concept repositories
 - Using existing mappings
 - Creating mappings through the UMLS
 - Comparing semantic descriptions
- ◆ Thinking outside the integration box

 Lister Hill National Center for Biomedical Communications 307

Motivation

Motivation Translational research

- ◆ “Bench to Bedside”
- ◆ Integration of clinical and research activities and results
- ◆ Supported by research programs
 - NIH Roadmap
 - Clinical and Translational Science Awards (CTSA)
- ◆ Requires the effective integration and exchange and of information between
 - Basic research
 - Clinical research

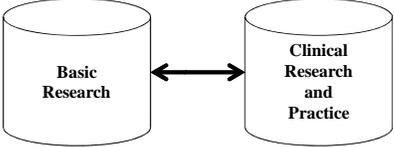
 Lister Hill National Center for Biomedical Communications 309

Translational research NIH Roadmap



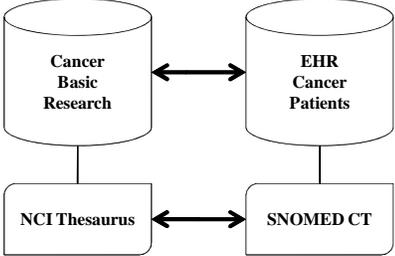
 Lister Hill National Center for Biomedical Communications 310

Motivation Translational research



 Lister Hill National Center for Biomedical Communications 311

Terminology and translational research



 Lister Hill National Center for Biomedical Communications 312

Some practical considerations
and issues
Integration approaches

Approaches to data integration

- ◆ **Warehousing**
 - Sources to be integrated are transformed into a common format and converted to a common vocabulary
 - Normalization through ontologies (e.g., GO annotations)
- ◆ **Mediation**
 - Local schema (of the sources)
 - Global schema (in reference to which the queries are made)
 - Ontologies help define the global schema and map between local and global schemas (OntoFusion, ARIANE)

 Lister Hill National Center for Biomedical Communications 314

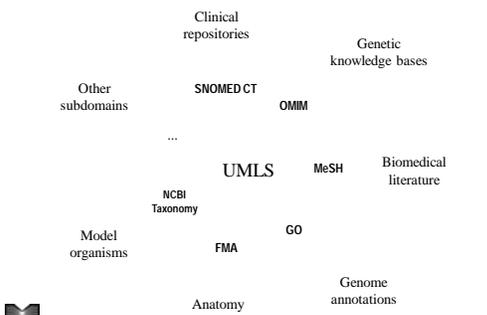
Some practical considerations
and issues
Concept repositories

(Integrated) concept repositories

- ◆ **Unified Medical Language System**
<http://umlsks.nlm.nih.gov>
- ◆ **NCBO's BioPortal**
<http://www.bioontology.org/tools/portal/bioportal.html>
- ◆ **Open Biomedical Ontologies (OBO)**
<http://obofoundry.org/>
- ◆ **caDSR**
http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr

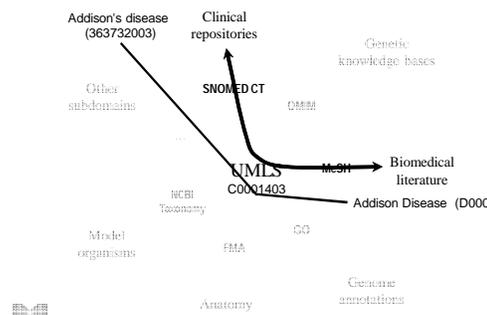
 Lister Hill National Center for Biomedical Communications 316

Integrating subdomains



 Lister Hill National Center for Biomedical Communications 317

Trans-namespace integration



 Lister Hill National Center for Biomedical Communications 318

Some practical considerations
and issues

Mappings

Mappings

The diagram shows two hierarchical tree structures representing ontologies. The left tree is labeled 'NCI Thesaurus' and the right tree is labeled 'SNOMED CT'. A dashed arrow labeled 'UMLS' points from the NCI Thesaurus tree to the SNOMED CT tree, indicating a mapping between the two.

Lister Hill National Center for Biomedical Communications 320

Mappings

- ◆ Created manually
 - UMLS
- ◆ Created automatically
 - BioPortal
- ◆ Key to enabling semantic interoperability
- ◆ Enabling resource for the Semantic Web

Lister Hill National Center for Biomedical Communications 321

Quality of mappings

- ◆ Created for a purpose
 - Reusability issues
- ◆ Generally unidirectional
 - Mapping from ontology 1 to ontology 2
 - Not necessarily reversible

Lister Hill National Center for Biomedical Communications 322

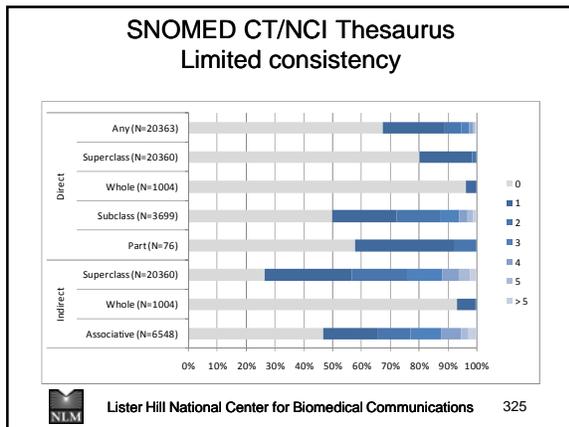
Some practical considerations
and issues

Comparing semantic descriptions

Semantic descriptions Consistent

The diagram illustrates semantic relationships between several concepts. At the top, 'S:Pancreatic structure' and 'N:Pancreas' are connected by a dashed arrow labeled 'owl:sameAs'. Below this, 'S:Disorder of pancreas' and 'N:Pancreatic disorder' are also connected by a dashed arrow labeled 'owl:sameAs'. 'S:Disorder of pancreas' is connected to 'S:Disorder of endocrine pancreas' by a solid arrow labeled 'S:is_a'. 'N:Pancreatic disorder' is connected to 'N:Endocrine pancreas disorder' by a solid arrow labeled 'N:subClassOf'. 'S:Disorder of endocrine pancreas' and 'N:Endocrine pancreas disorder' are connected by a dashed arrow labeled 'owl:sameAs'. At the bottom, both 'S:Disorder of endocrine pancreas' and 'N:Endocrine pancreas disorder' are connected to the code 'C0030286' by solid arrows labeled 'U:hasCUI'.

Lister Hill National Center for Biomedical Communications 324



Comparing formal definitions

- ◆ Relatively small proportion of relata in common between equivalent concepts from NCI and SNOMED CT
- ◆ Large number of primitive concepts in NCI and SNOMED CT (70-80%)
- ◆ Insufficient for effectively comparing definitions
 - Could not be used for validating the mapping provided by the UMLS

[Bodenreider, KRMed 2008]

Lister Hill National Center for Biomedical Communications 326

Exercises

Exercise #1

- ◆ Check the equivalence (shared relata) between these 2 concepts:
 - NCI Thesaurus: N:Endocrine pancreas disorder
 - SNOMED CT: S:Disorder of endocrine pancreas

Lister Hill National Center for Biomedical Communications 328

Exercise #2

- ◆ Find a correspondence in SNOMED CT for the LOINC term: *Sodium:SCnc:-Pt:Ser/Plas:Qn* [the molar concentration of sodium is measured in the plasma (or serum), with quantitative result]

Axis	Value
Component	Sodium
Property	SCnc - Substance Concentration (per volume)
Timing	Pt - Point in time (Random)
System	Ser/Plas - Serum or Plasma
Scale	Qn - Quantitative
Method	--

Lister Hill National Center for Biomedical Communications 329

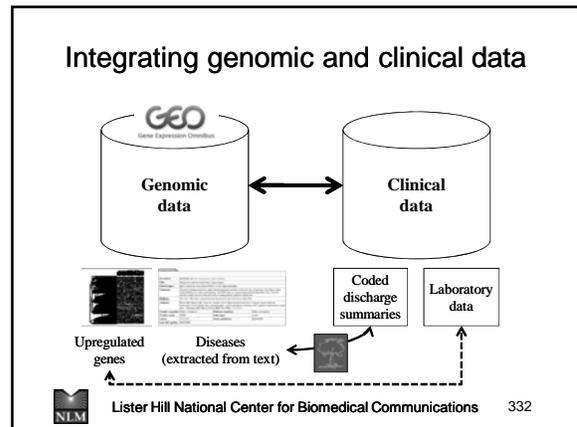
Comments on exercise #2

- ◆ Difficult in the absence of a search mechanism on the values of the relations
- ◆ Large number of underspecified descriptions in SNOMED CT
- ◆ 2 separate concepts for plasma and serum concentrations of sodium in SNOMED CT
- ◆ Property, time and scale not represented in SNOMED CT

Lister Hill National Center for Biomedical Communications 330

Thinking outside the integration box

The Butte approach



References

- ◆ Dudley J, Butte AJ "Enabling integrative genomic analysis of high-impact human diseases through text mining." *Pac Symp Biocomput* 2008; 580-91
- ◆ Chen DP, Weber SC, Constantinou PS, Ferris TA, Lowe HJ, Butte AJ "Novel integration of hospital electronic medical records and gene expression measurements to identify genetic markers of maturation." *Pac Symp Biocomput* 2008; 243-54
- ◆ Butte AJ, "Medicine. The ultimate model organism." *Science* 2008; 320: 5874: 325-7

Lister Hill National Center for Biomedical Communications 333

Medical Ontology Research

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

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