



Korea Advanced Institute  
of Science and Technology

December 4, 2007

# The Unified Medical Language System

*What is it and how to use it?*



*Olivier Bodenreider, MD, PhD*

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

# Outline

---

- ◆ Part I: *What is the UMLS?*
  - Introduction
  - Overview through an example
  - The three UMLS Knowledge Sources
    - UMLS Metathesaurus
    - UMLS Semantic Network
    - SPECIALIST Lexicon and lexical tools
- ◆ Part II: *How to use the UMLS?*
  - A UMLS-based algorithm

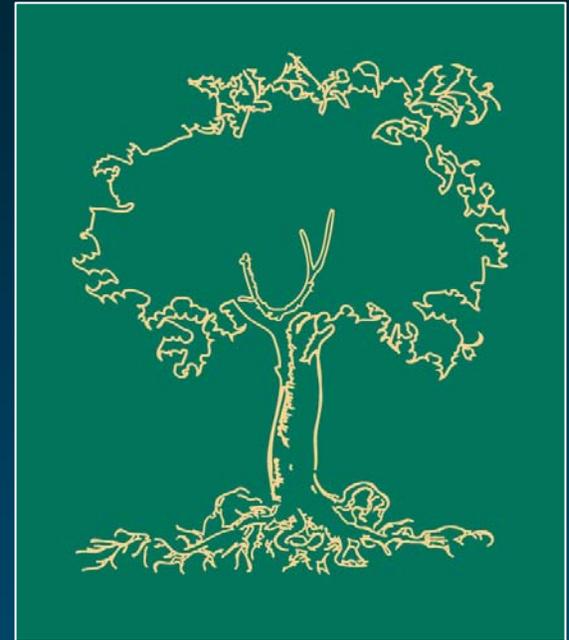
Part I  
What is the UMLS?

*(1) Introduction*

# What does UMLS stand for?

---

- ◆ Unified
- ◆ Medical
- ◆ Language
- ◆ System



UMLS<sup>®</sup>  
Unified Medical Language System<sup>®</sup>  
UMLS Metathesaurus<sup>®</sup>

# Motivation

---

- ◆ Started in 1986
- ◆ National Library of Medicine
- ◆ “Long-term R&D project”
- ◆ Complementary to IAIMS

(Integrated Academic  
Information Management Systems)

«[...] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

- The first is **the variety of ways the same concepts are expressed** in different machine-readable sources and by different people.
- The second is the **distribution** of useful information among many disparate databases and systems.»

# The UMLS in practice

---

## ◆ Database

- Series of relational files

## ◆ Interfaces

- Web interface: Knowledge Source Server (UMLSKS)
- Application programming interfaces (Java and XML-based)

## ◆ Applications

- lvg (lexical programs)
- MetamorphoSys (installation and customization)
- RRF browser (browsing subsets)



The UMLS is *not* an end-user application

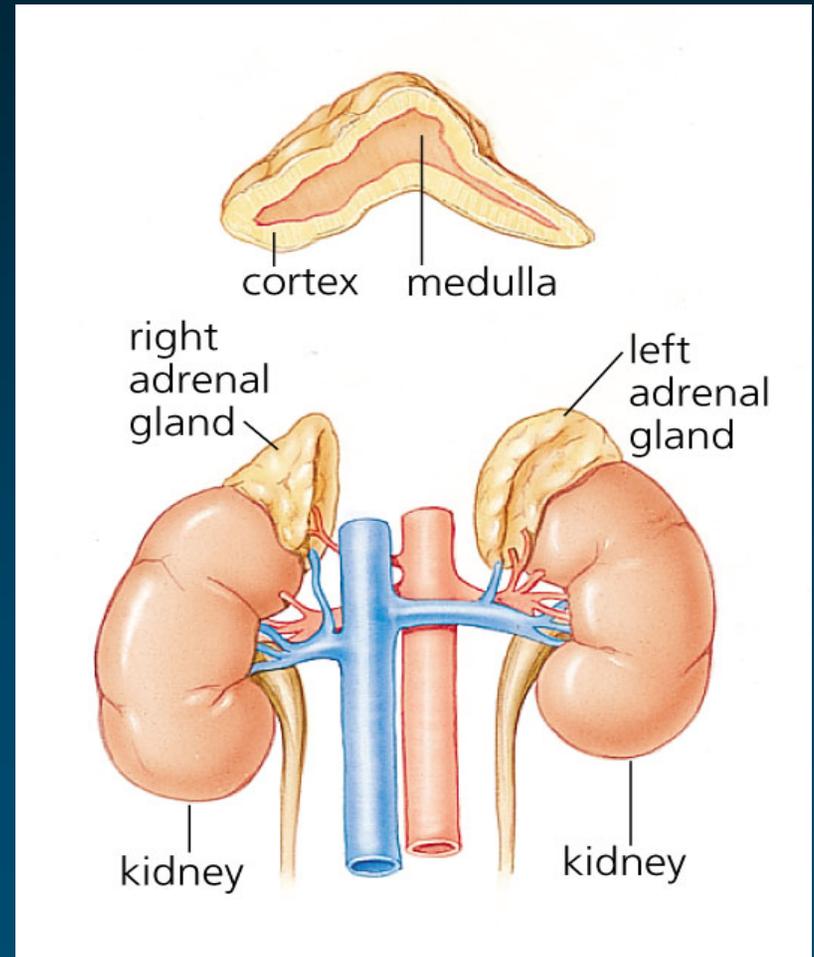
# Part I

## What is the UMLS?

*(2) Overview through an example*

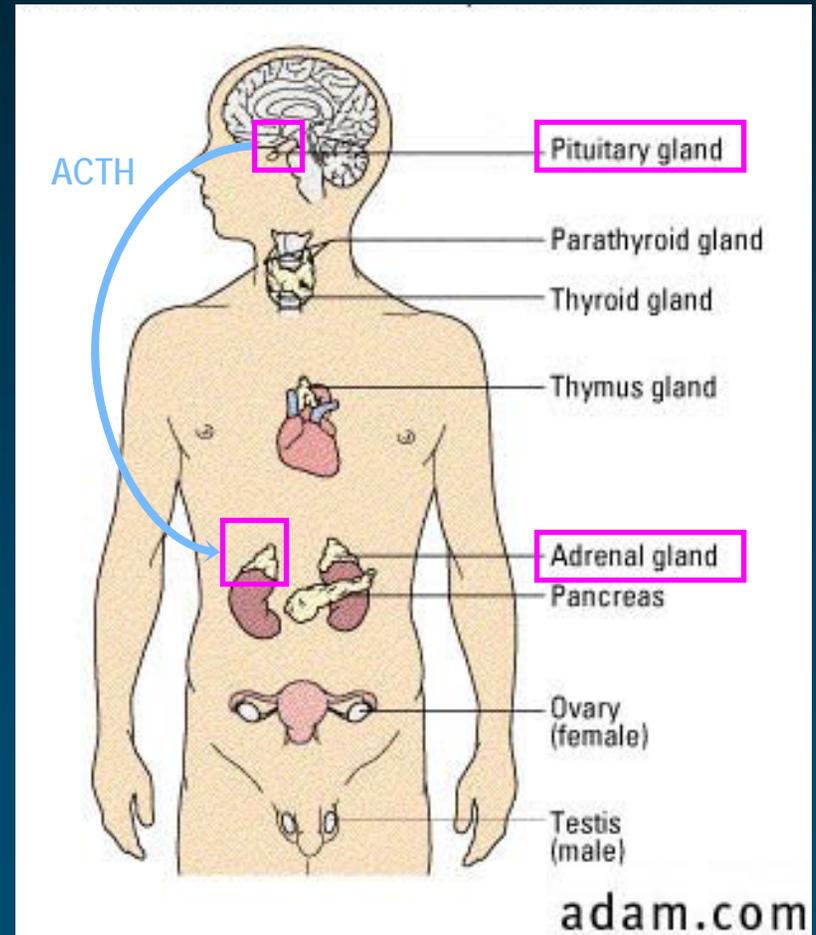
# Addison's disease

- ◆ 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



# 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



# Addison's disease: Symptoms

---

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

# AD in medical vocabularies

---

## ◆ Synonyms: different terms

- Addisonian syndrome
  - Bronzed disease
  - Melasma addisonii
  - Asthenia pigmentosa
  - Primary adrenal deficiency
  - Primary adrenal insufficiency
  - Primary adrenocortical insufficiency
  - Chronic adrenocortical insufficiency
- } eponym
- } symptoms
- } clinical variants

## ◆ Contexts: different hierarchies



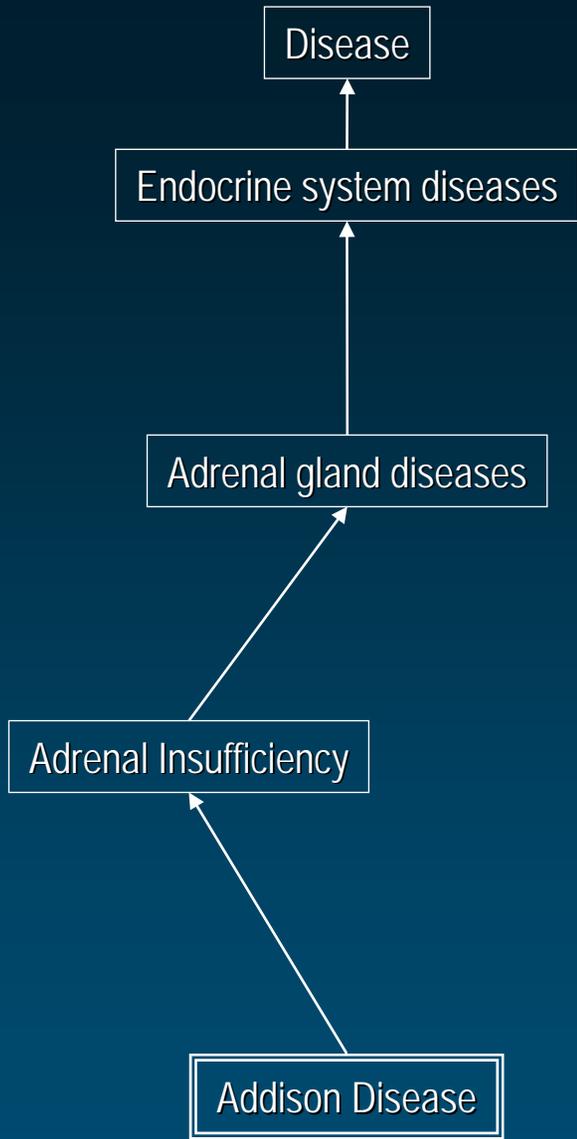
# Organize terms

- ◆ Synonymous terms clustered into a concept
- ◆ Preferred term
- ◆ Unique identifier (CUI)

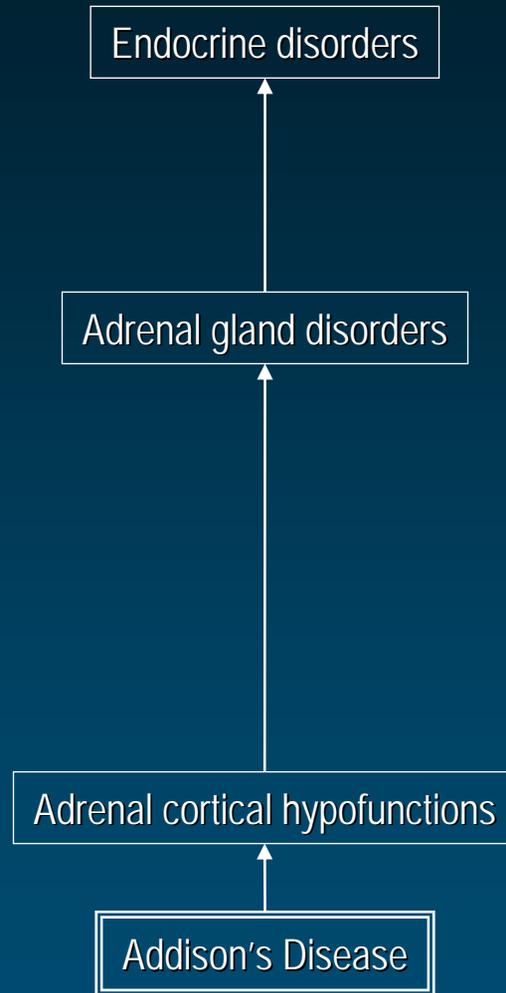
Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

C0001403

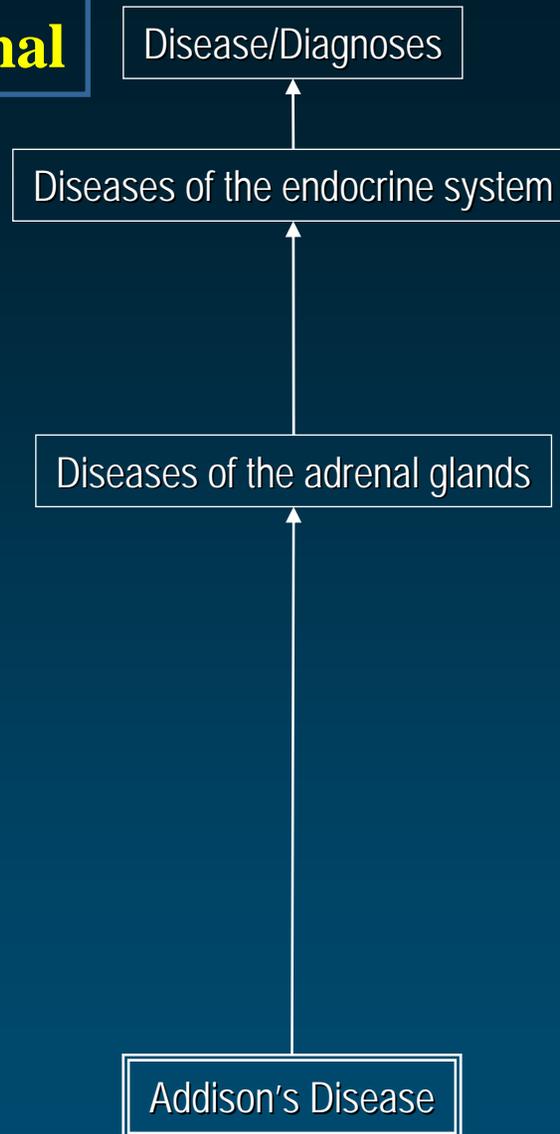
Addison's disease



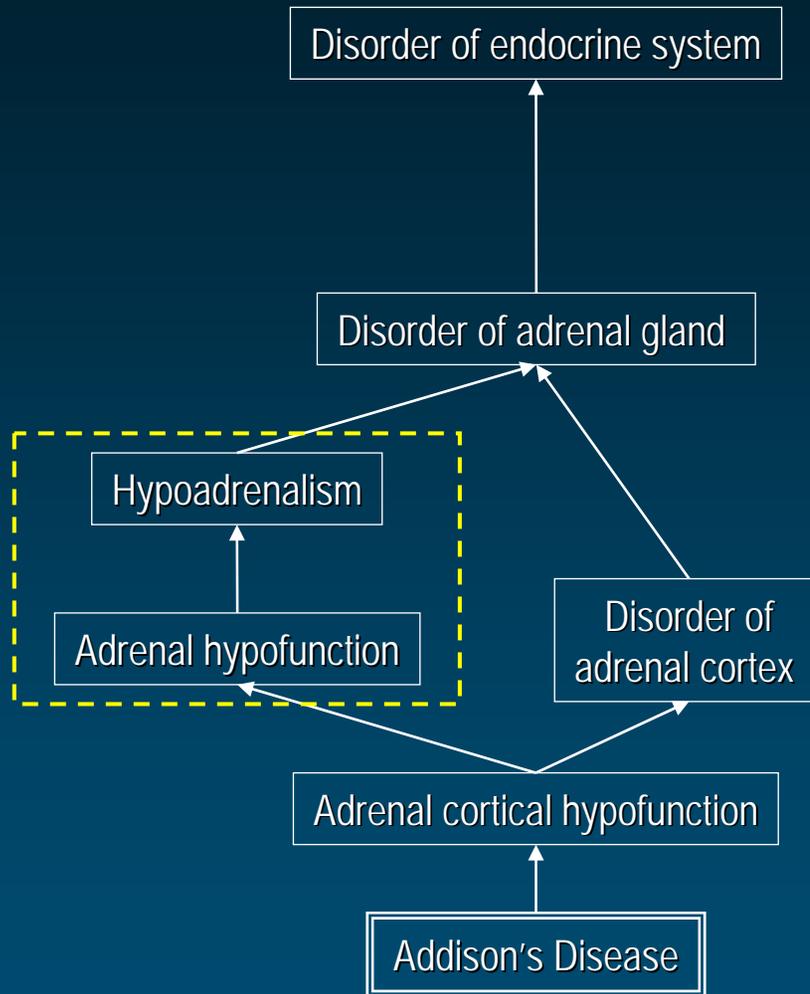
# MedDRA



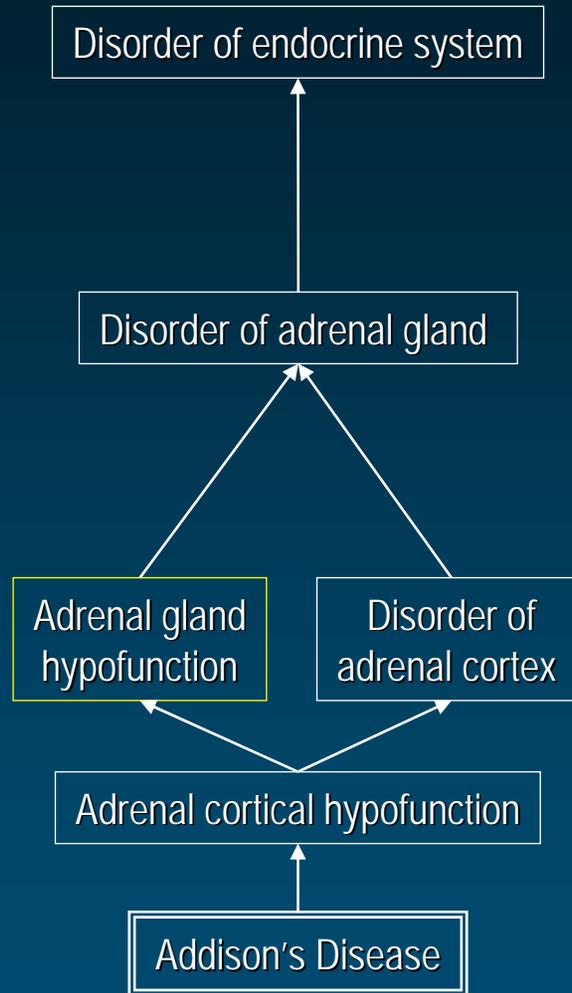
# SNOMED International



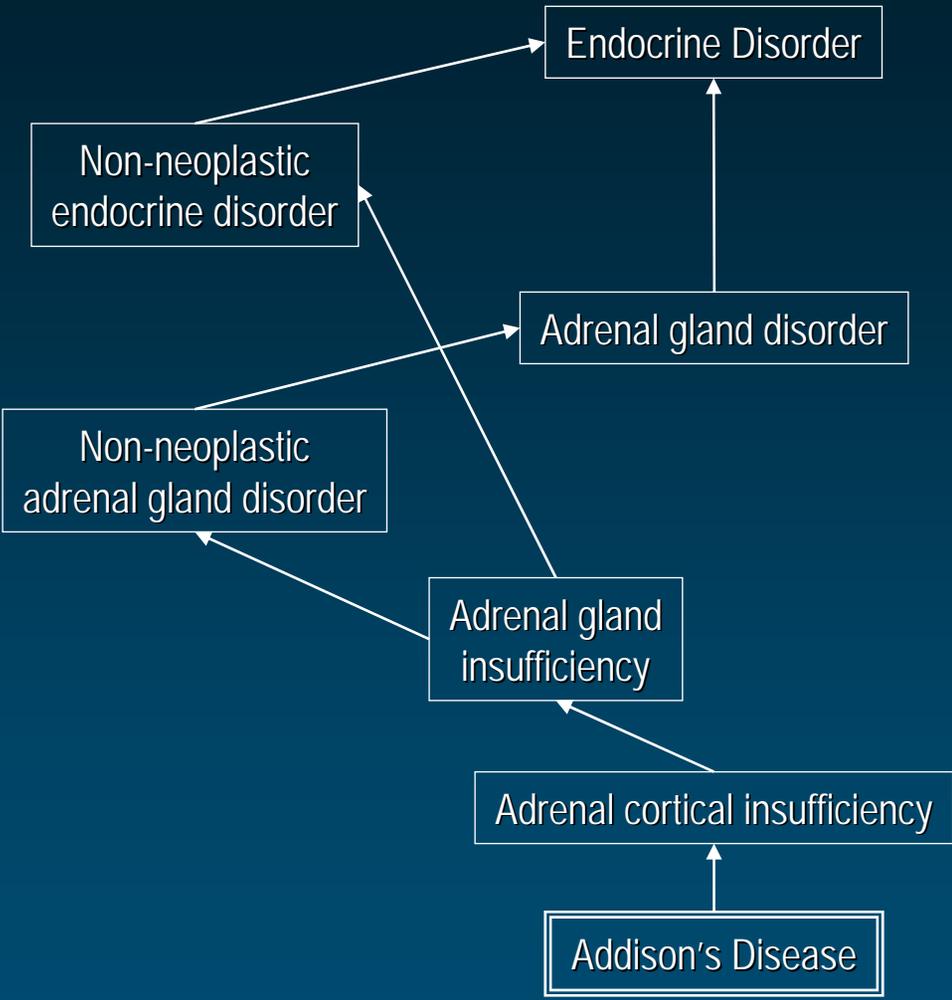
# SNOMED CT (native)



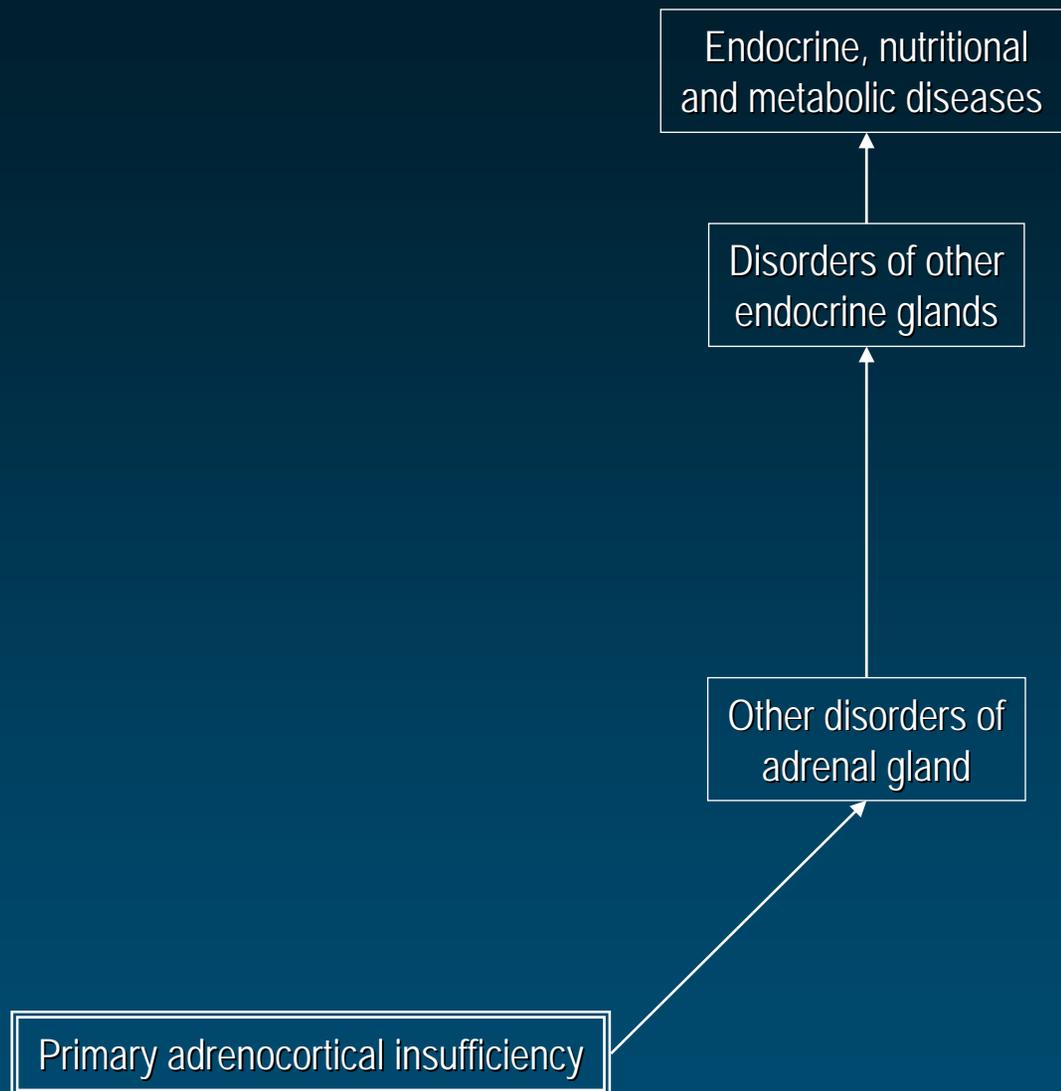
# SNOMED CT (UMLS view)



# NCI Thesaurus

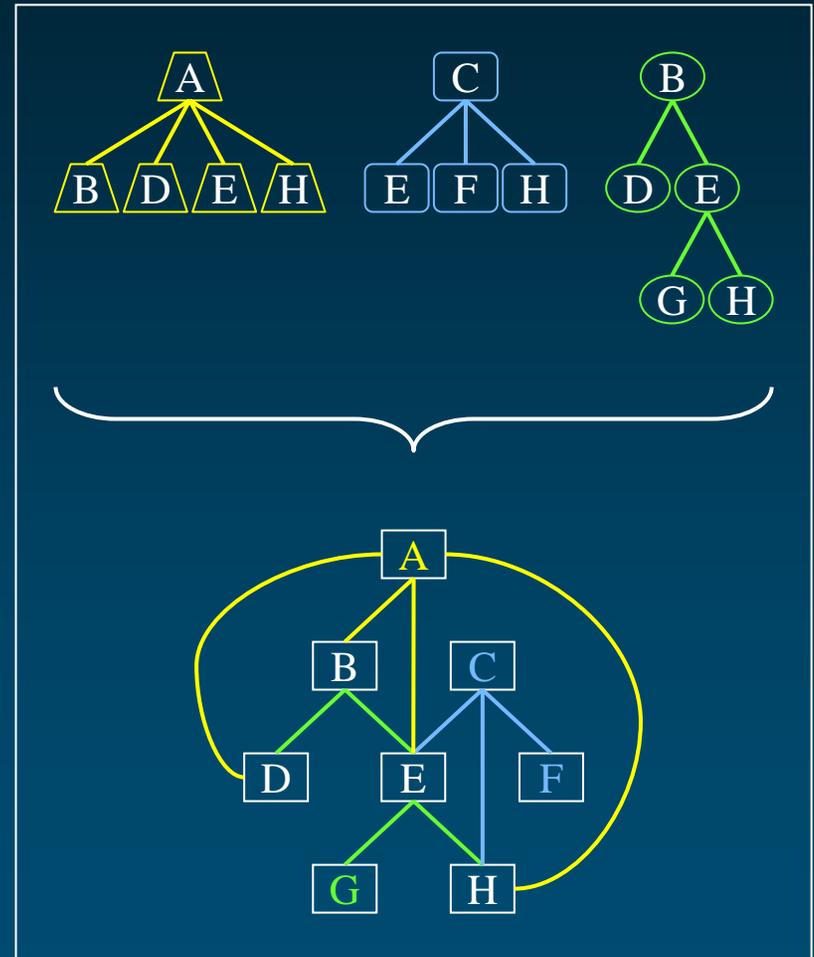


# ICD-10

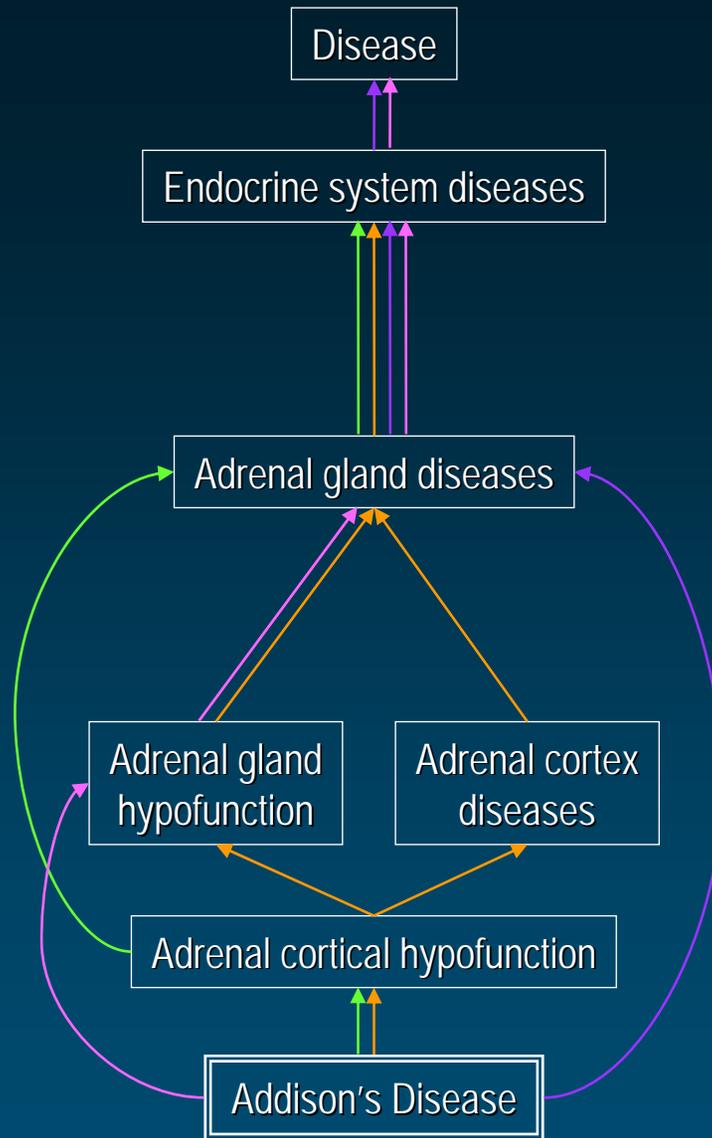


# Organize concepts

- ◆ Inter-concept relationships: hierarchies from the source vocabularies
- ◆ Redundancy: multiple paths
- ◆ One graph instead of multiple trees (multiple inheritance)

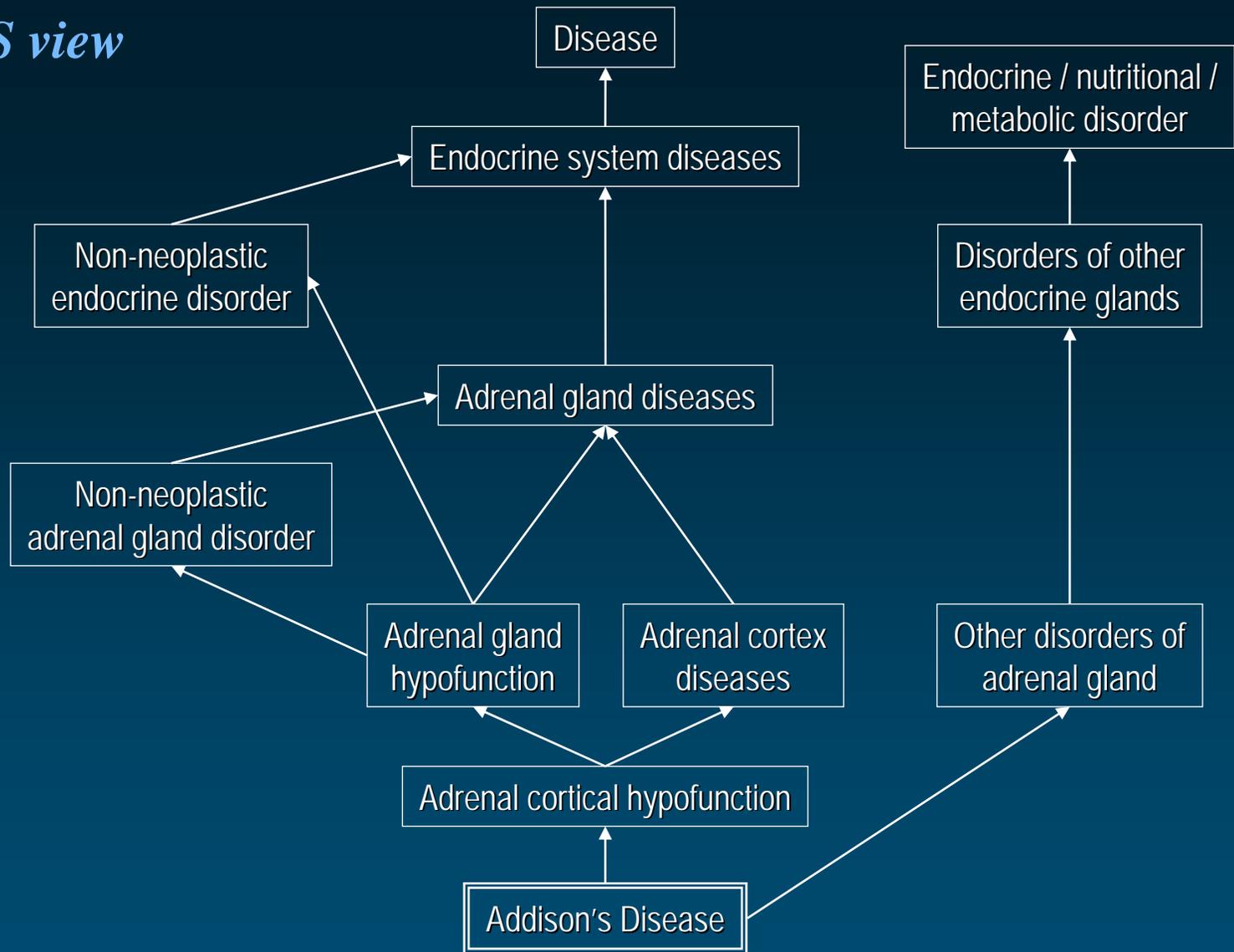


*organize concepts*

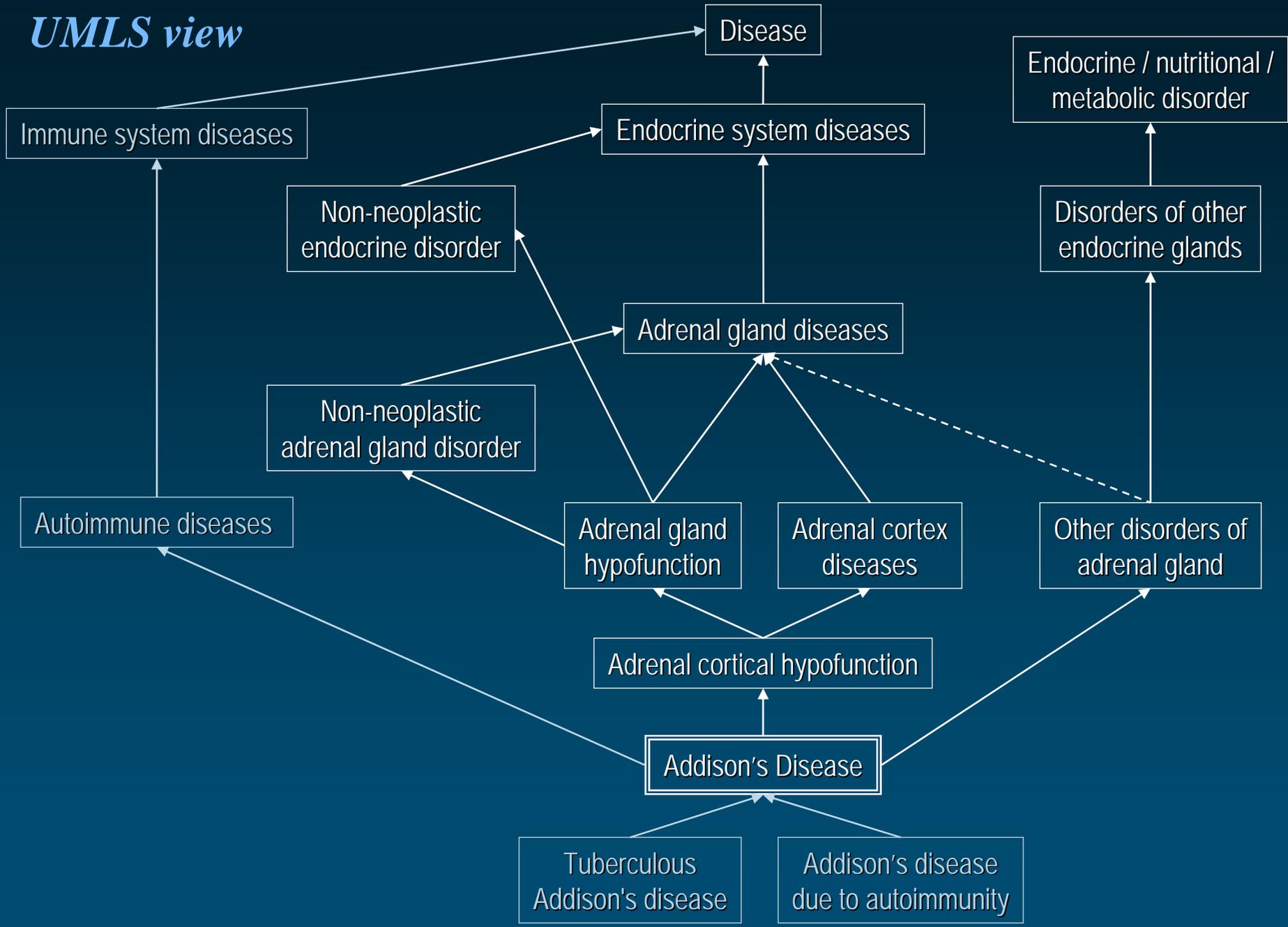


- SNOMED CT**
- SNOMED Intl**
- MeSH**
- MedDRA**

*UMLS view*



*UMLS view*



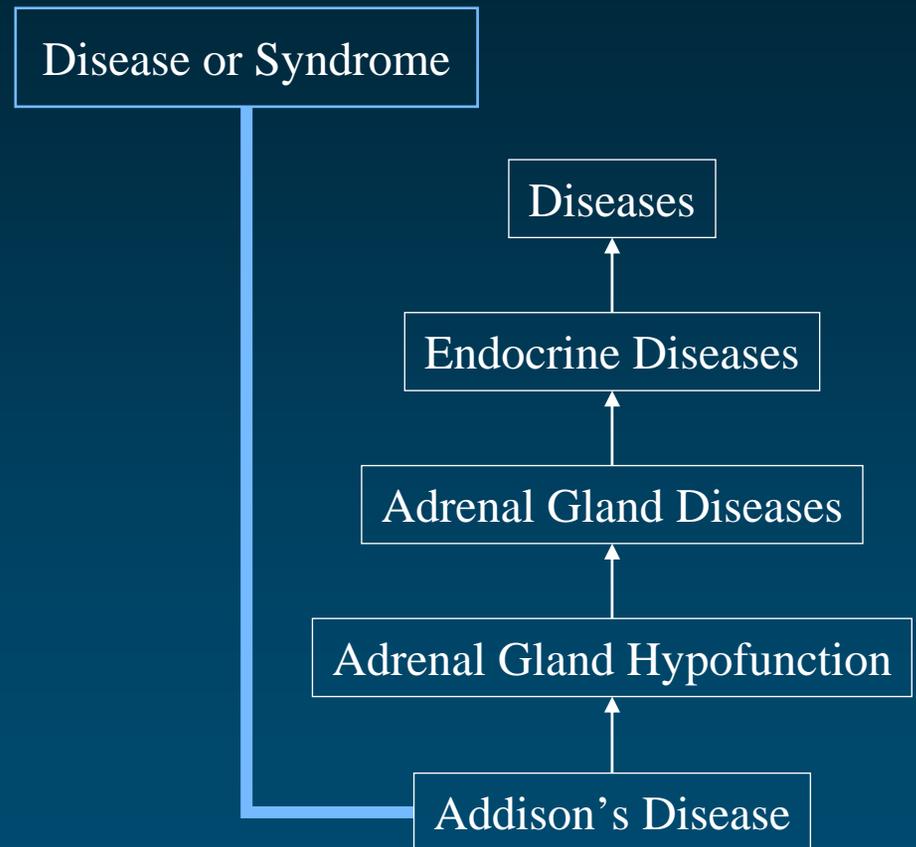
# Relate to other concepts

---

- ◆ Additional hierarchical relationships
  - link to other trees
  - make relationships explicit
- ◆ Non-hierarchical relationships
- ◆ Co-occurring concepts
- ◆ Mapping relationships

# Categorize concepts

- ◆ High-level categories (semantic types)
- ◆ Assigned by the Metathesaurus editors
- ◆ Independently of the hierarchies in which these concepts are located



# How do they do that?

---

- ◆ Lexical knowledge
- ◆ Semantic pre-processing
- ◆ UMLS editors

# Lexical knowledge

---

Adrenal gland diseases

Adrenal disorder

Disorder of adrenal gland

Diseases of the adrenal glands

C0001621

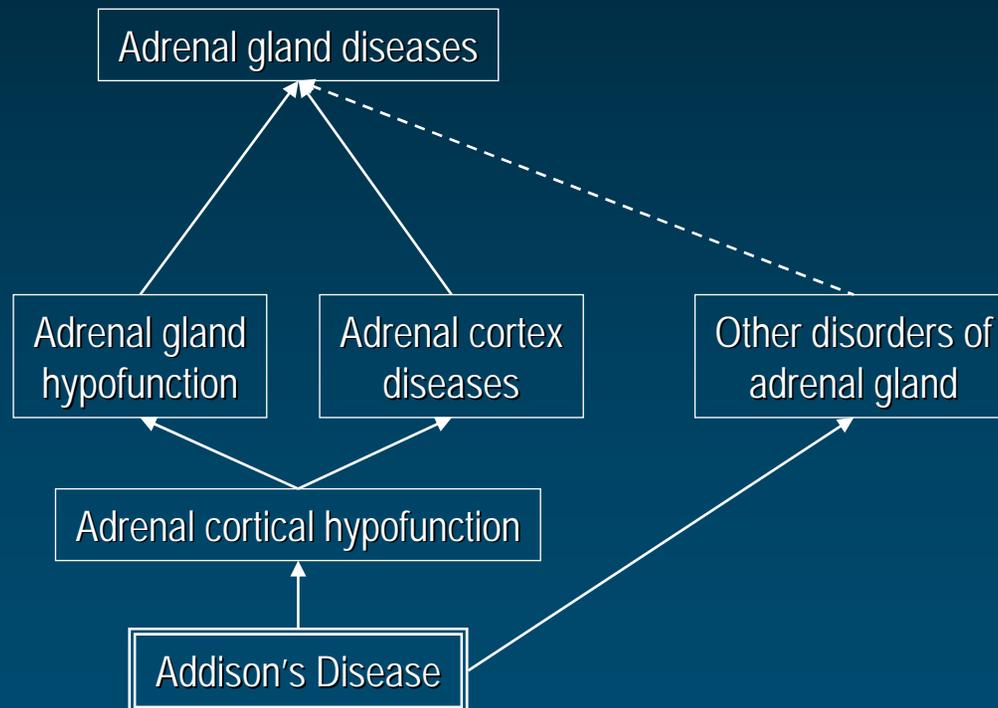
# Semantic pre-processing

---

- ◆ Metadata in the source vocabularies
- ◆ Tentative categorization
- ◆ Positive (or negative) evidence for tentative synonymy relations based on lexical features

# Additional knowledge: UMLS editors

---



# UMLS Summary

---

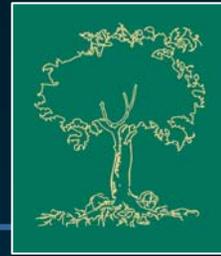
- ◆ Synonymous terms clustered into concepts
- ◆ Unique identifier
  
- ◆ Finer granularity
- ◆ Broader scope
- ◆ Additional hierarchical relationships
- ◆ Semantic categorization

# Part I

## What is the UMLS?

### *(3) UMLS Metathesaurus*

# Unified Medical Language System



## ◆ SPECIALIST Lexicon

- 360,000 lexical items
- Part of speech and variant information

## ◆ Metathesaurus

- 6M names from over 100 terminologies
- 1.5M concepts
- 8M relations

## ◆ Semantic Network

- 135 high-level categories
- 7000 relations among them

Lexical  
resources

Terminological  
resources

Ontological  
resources

# Metathesaurus Basic organization

---

## ◆ Concepts

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

## ◆ Relations

- Concepts are related to other concepts
- Properties are attached to relations, e.g.,
  - Type of relationship
  - Source

# Source Vocabularies

(2007AC)

---

- ◆ 141 source vocabularies
  - 17 languages
- ◆ Broad coverage of biomedicine
  - 6.1M names
  - 1.5M concepts
  - 8M relations
- ◆ Common presentation

# Biomedical terminologies

---

## ◆ General vocabularies

- anatomy (UWDA, Neuronames)
- drugs (RxNorm, First DataBank, Micromedex)
- medical devices (UMD, SPN)

## ◆ Several perspectives

- clinical terms (SNOMED CT)
- information sciences (MeSH, CRISP)
- administrative terminologies (ICD-9-CM, CPT-4)
- data exchange terminologies (HL7, LOINC)

# Biomedical terminologies (cont'd)

---

## ◆ Specialized vocabularies

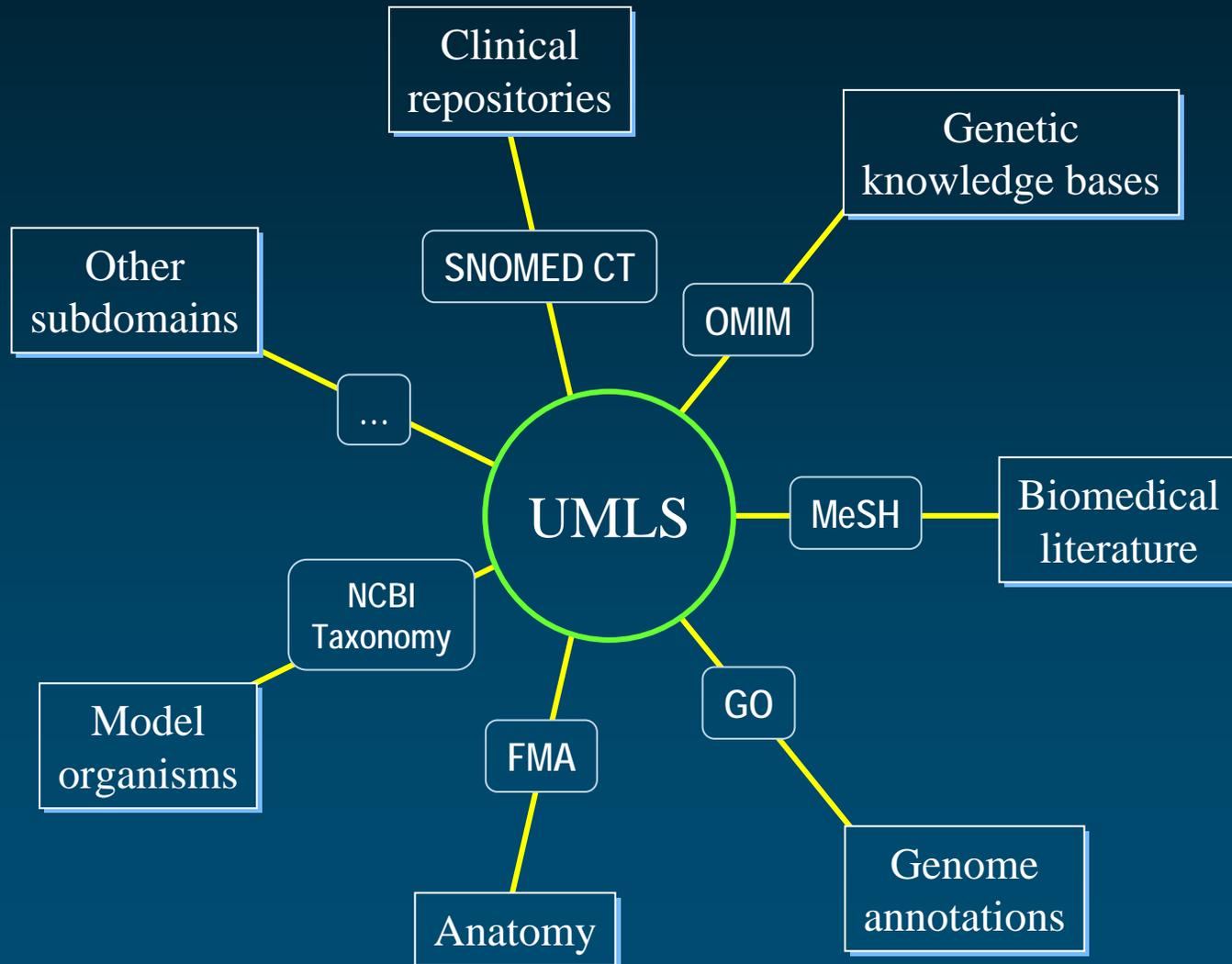
- nursing (NIC, NOC, NANDA, Omaha, PCDS)
- dentistry (CDT)
- oncology (PDQ)
- psychiatry (DSM, APA)
- adverse reactions (COSTART, WHO ART)
- primary care (ICPC)

## ◆ Terminology of knowledge bases (AI/Rheum, DXplain, QMR)

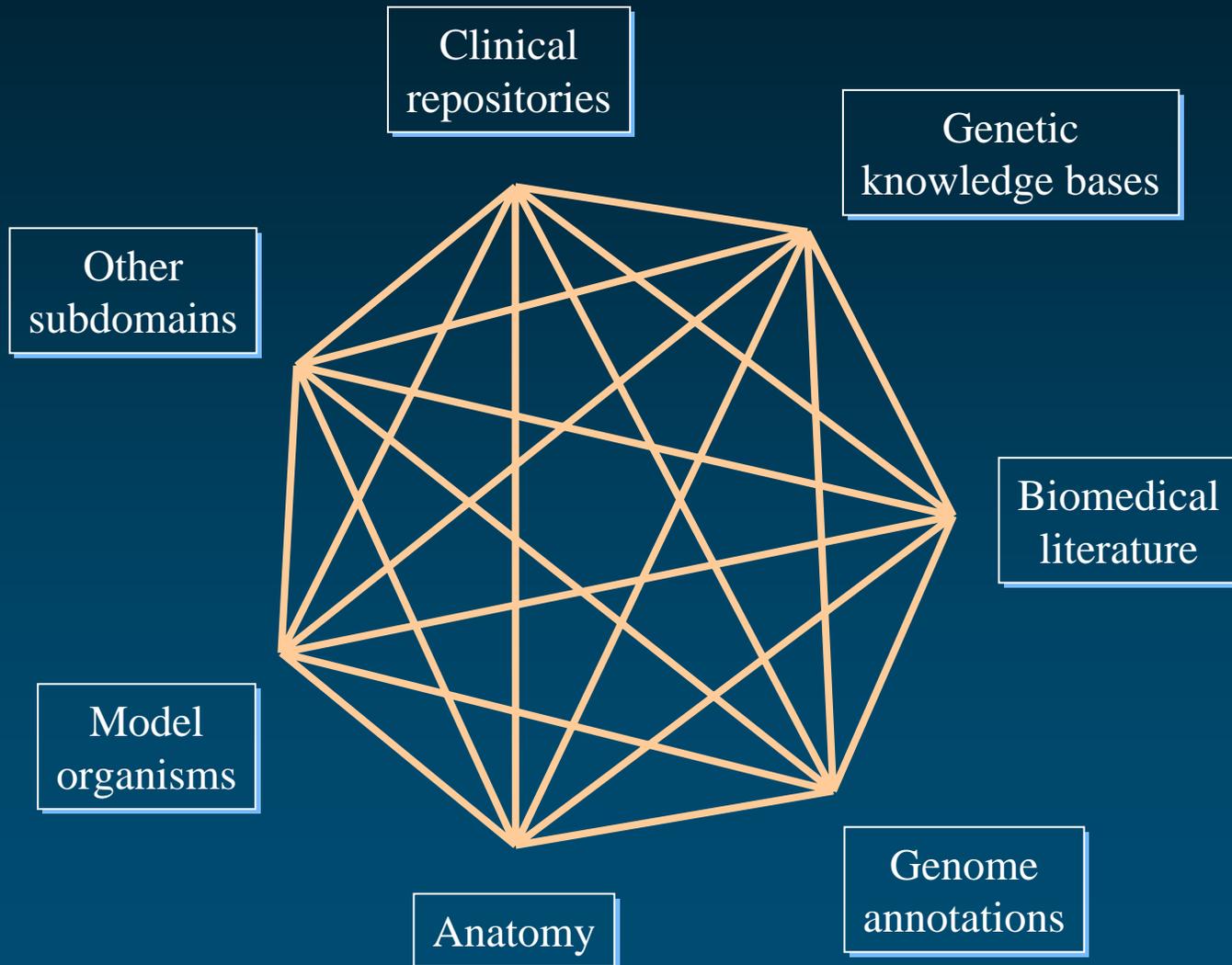
The UMLS serves as a vehicle for the regulatory standards  
(HIPAA, CHI)



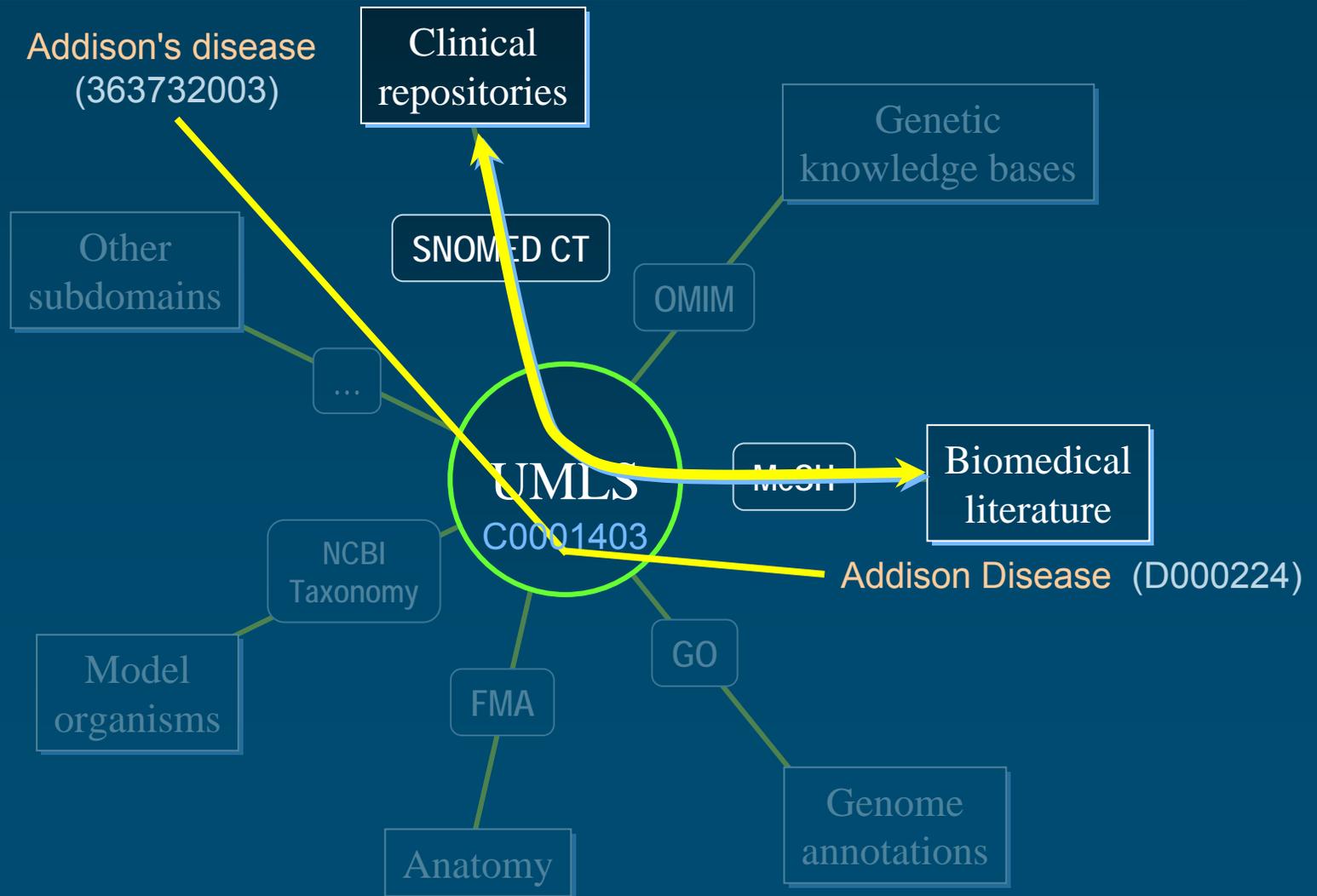
# Integrating subdomains



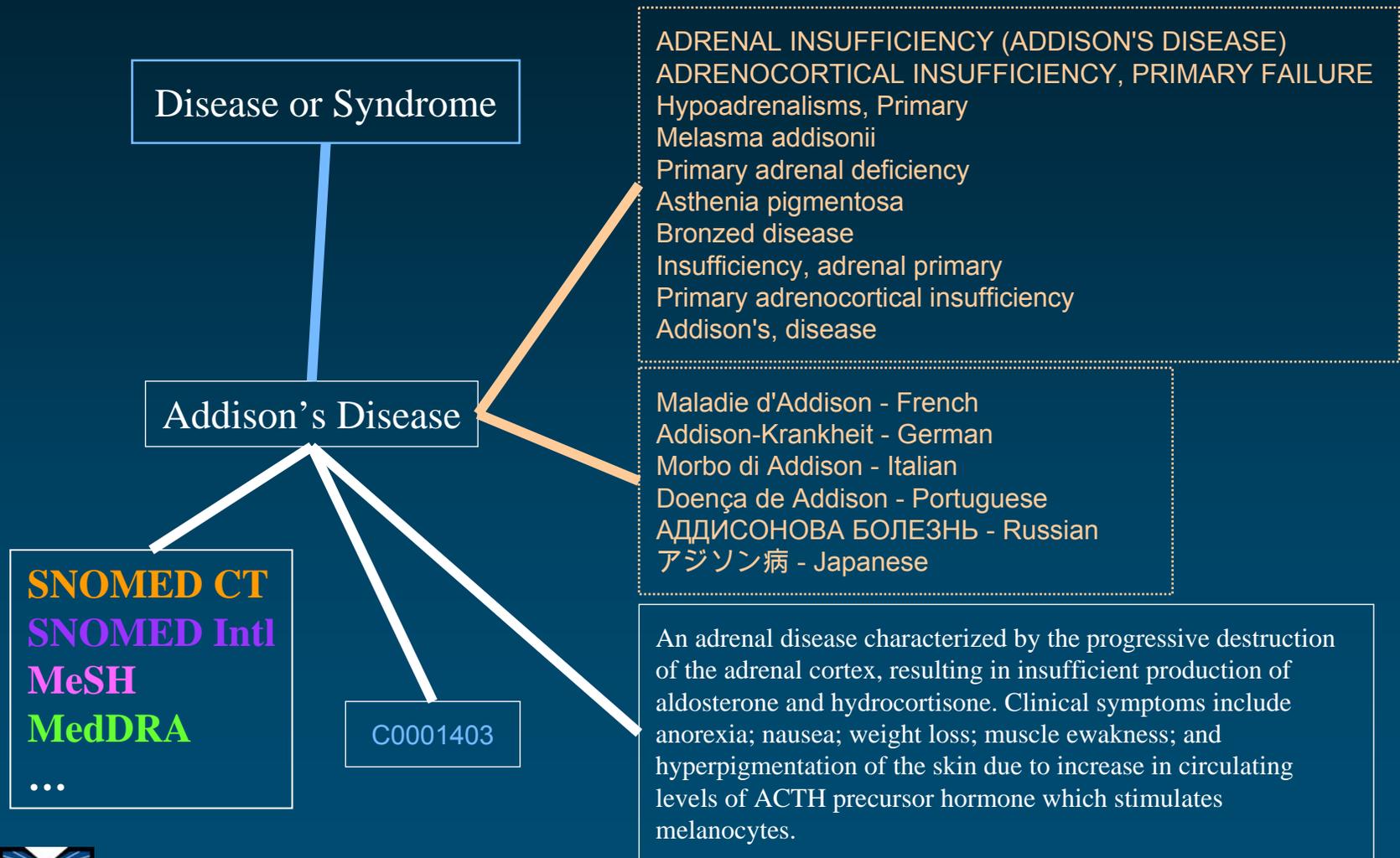
# Integrating subdomains



# Trans-namespace integration



# Addison's Disease: Concept





## Part II

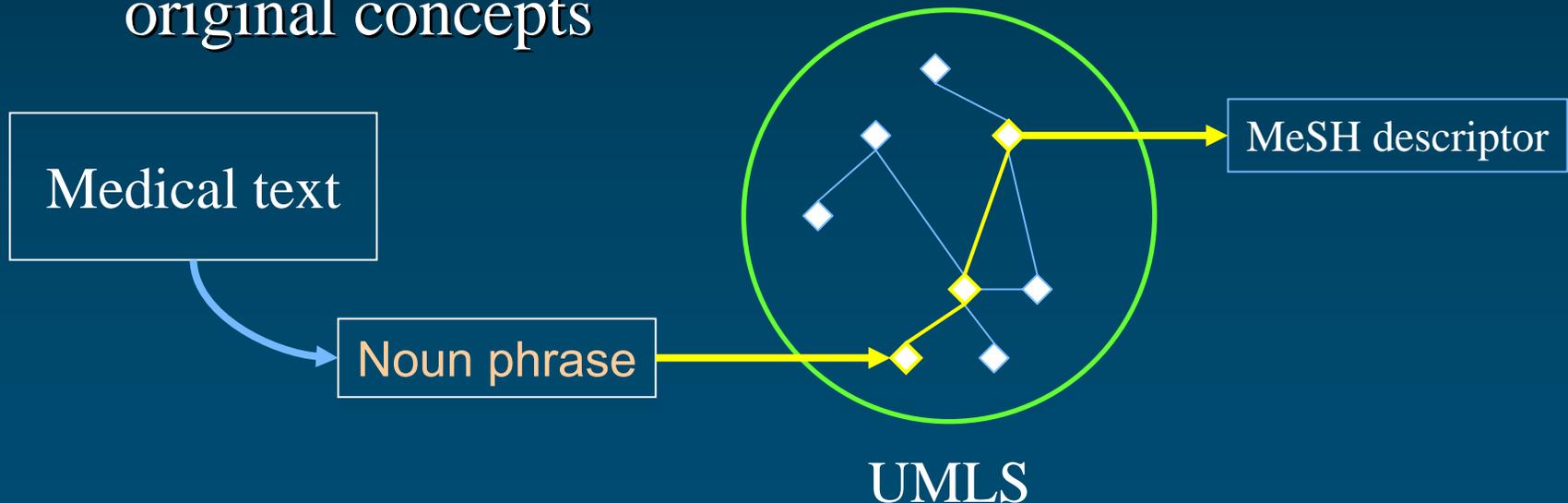
How to use the UMLS?

*A UMLS-based algorithm*

# Indexing Initiative

[Aronson & al., *AMIA*, 2000]

- ◆ For noun phrases extracted from medical texts, map to UMLS concepts
- ◆ Then, select from the MeSH vocabulary the concepts that are the most closely related to the original concepts



# Restrict to MeSH

[Bodenreider & al., *AMIA*, 1998]

- ◆ Based on the principle of semantic locality
- ◆ Use different components of the UMLS
- ◆ 4 techniques of increasing aggressiveness
  - Use Synonymy **MRCONSO**
  - Use Associated expressions (ATXs) **MRATX + MRREL**
  - Explore the Ancestors **MRREL + SN**
  - Explore the Other related concepts **MRREL + SN**

# Restrict to MeSH Synonymy

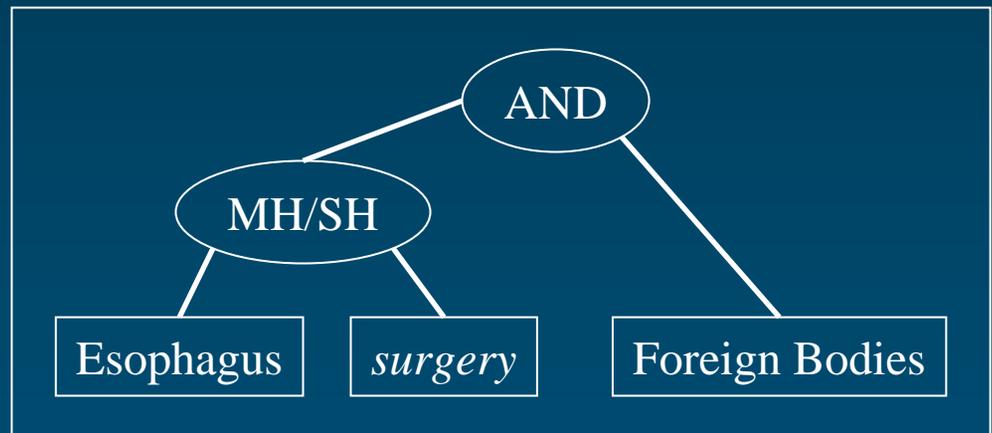
---

- ◆ Term mapped to Source concept
- ◆ For this concept, is there a synonym term that comes from MeSH? (MRCONSO)

# Restrict to MeSH Assoc. expressions

- ◆ If not,
- ◆ Is there an associated expression (ATX) that describes this concept using a combination of MeSH descriptors? (**MRATX/MRMAP + MRREL**)

Endoscopic removal of intraluminal foreign body from oesophagus without incision



# Restrict to MeSH Ancestors

---

- ◆ If not, let us build the graph of the ancestors of this concept
  - using parents and broader concepts (MRREL)
  - all the way to the top
  - excluding ancestors whose semantic types are not compatible with those of the source concept (MRSTY)
- ◆ From the graph, select the concepts that come from MeSH (MRCONSO)
- ◆ Remove those that are ancestors of another concept coming from MeSH

# Restrict to MeSH Other related concepts

---

- ◆ If not, explore the other related concepts (**MRREL**) whose semantic types are compatible with those of the source concept (**MRSTY**)
- ◆ From those, select the concepts that come from MeSH (**MRCNSO**)

# Restrict to MeSH Example

Vein of neck, NOS

There is a MeSH term in the synonyms of SC

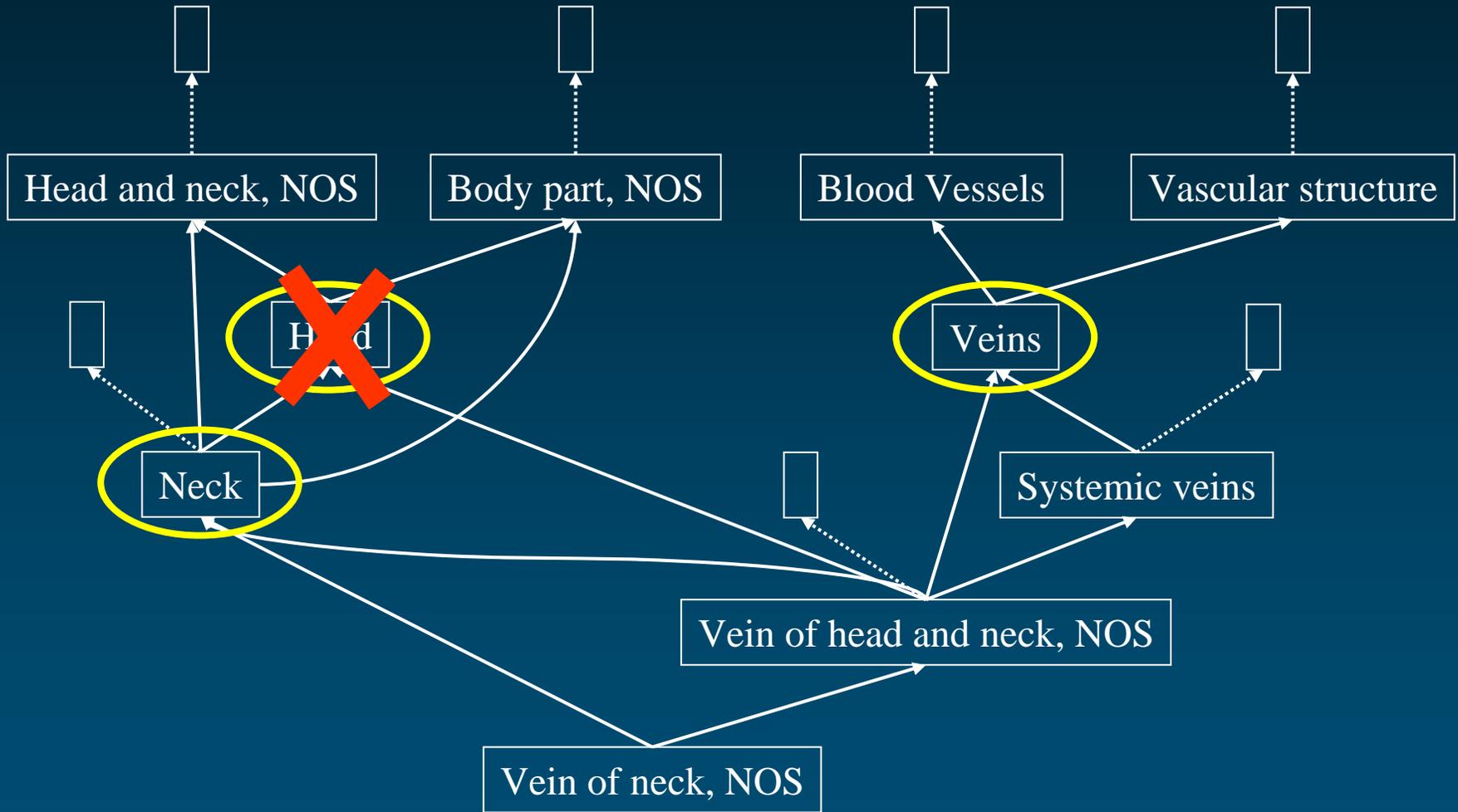
SC is described by a combination of MeSH terms (ATX)

The ancestors of SC contain MeSH terms

MeSH terms from non-hierarchically related concepts

Vein + Neck

# Restrict to MeSH Example



# Restrict to MeSH Quantitative results

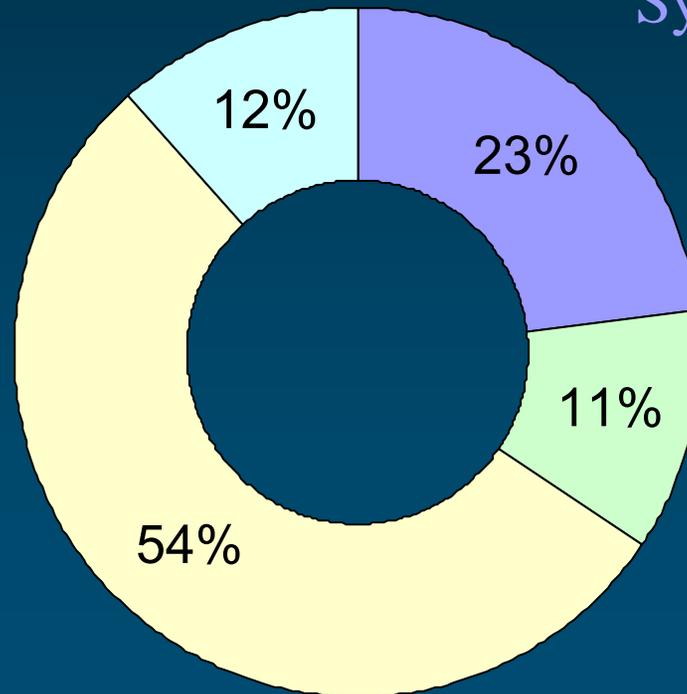
- ◆ 86% of UMLS concepts mapped to MeSH (2007)

Other related concepts

Synonymy

Graph of  
ancestors

Built-in  
mappings



# Restrict to MeSH Qualitative results

---

## ◆ Qualitative evaluation

- 1,036 concepts extracted from 200 MEDLINE citations
- manual review of every mapping or failure

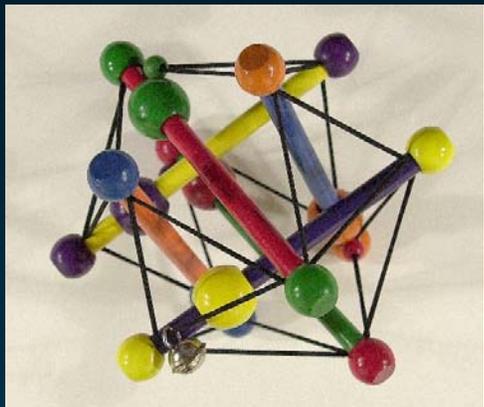
## ◆ 61% Relevant

- Subtotal Gastrectomy → Gastrectomy
- Encephalopathy, NOS → Brain Diseases

## ◆ 28% More or less relevant

- Vitamin A measurement → Laboratory Procedure
- Swelling, NOS → Symptoms

## ◆ 11% Non relevant



# Medical Ontology Research

Contact: [olivier@nlm.nih.gov](mailto:olivier@nlm.nih.gov)

Web: [mor.nlm.nih.gov](http://mor.nlm.nih.gov)



*Olivier Bodenreider*

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

# Bibliography

# References: UMLS home page

---

## ◆ UMLS home page

- [http:// www.nlm.nih.gov/research/umls/](http://www.nlm.nih.gov/research/umls/)

## ◆ UMLS documentation

- Formerly know as the “Green Book”
- Now online documentation
- <http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML>

# References

---

## ◆ Short presentation

- Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue), D267-70.

## ◆ UMLS as a research project

- Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. *Methods Inf Med*, 32(4), 281-91.
- Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc*, 5(1), 1-11.

# References

---

## ◆ Technical papers

- McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods Inf Med*, 34(1-2), 193-201.

## ◆ Comprehensive bibliography 1986-96

<http://www.nlm.nih.gov/pubs/cbm/umlscbm.html>

# Documentation and Support

# UMLS documentation and support

---

## ◆ UMLS homepage

- links to various UMLS resources
- <http://www.nlm.nih.gov/research/umls/>

## ◆ UMLSKS homepage

- links to the User's and Developer's guides
- <http://umlsks.nlm.nih.gov/>

## ◆ UMLS mailing list

- UMLSUSERS-L@LIST.NIH.GOV

## ◆ Email address for support

- [custserv@nlm.nih.gov](mailto:custserv@nlm.nih.gov)

