Unified Medical Language System and Dictionary Generation for Phenotyping

2021 Verity Mini-Course

Tianrun Cai, MD
Associate Medical Informatician
Verity Bioinformatics Core
Brigham and Women's Hospital
Harvard Medical School

Outline

- A Brief Tour of the Unified Medical Language System (UMLS)
- Build a dictionary using UMLS and online knowledge sources

Background

- Electronic Medical Record (EMR) data
  - Structured data (billing codes):
    - Diagnosis, procedure, medication, laboratory test
  - Unstructured data – clinical notes
- Medical research
  - Structured data only is sometimes insufficient
  - Additional information from notes

Background

- Medical research
  - Natural language processing(NLP)
    - **Language understanding** and generation by a computer
    - Application in medicine
      - Literature indexing – PubMed
      - Data extraction
      - Clinical decision support
      - Phenotyping

Background

- Medical research
  - Traditional way
    - Manually come up a term list

      E.g., **joint pain**, **joint pains**, **painful joints**, **arthralgias**, rheumatoid arthritis, morning stiffness, CRP, etc. (for rheumatoid arthritis)
    - Process notes get term count
    - Grouping term by knowledge
    - Modeling

Why Is UMLS Useful

Note 1

HISTORY: Multiple sclerosis, gastroesophageal reflux disease, **joint pains** in her knees, overweight, **inflammatory arthritis** with mildly increased **CRP**.
Reason for Visit
Joint Pain
Reason for Visit History
Diagnoses
Inflammatory arthritis - Primary M19.90
…

Note 2

Problem list:
Arthralgias
…

Patient 1

Why Is UMLS Useful

	Inflammatory arthritis	CRP	Joint Pain
patient 1	2	1	3
patient 2	...	...	...

Example Data

Background

- Medical research
  - Traditional way
    - Disadvantage
      - Missing terms

        E.g., 35 different expressions for ‘joint pain’ in English – joint ache, ache in joint
  - Very useful resource for NLP in medicine

    Unified Medical Language System (UMLS)
    - Automatically find potential synonyms

UMLS

❖ What's Unified Medical Language System (UMLS)

- The Unified Medical Language System (UMLS) is a set of files and software available from the U.S. National Library of Medicine (NLM) that brings together many biomedical vocabularies and standards for drugs, disorders, procedures, lab tests, medical devices, organisms, anatomy, genes, and more.
- **In brief: Standardized medical vocabulary system organized by concepts**

UMLS

- **The Unified Medical Language System**
  - **Metathesaurus**
  - **Semantic Network**
  - SPECIALIST lexicon and tools

Metathesaurus

- **The Unified Medical Language System (UMLS)**
  - **Metathesaurus**
    - Vocabulary database
    - 218 source vocabularies
      - ICD10 (International Classification of Diseases, 10th Revision), ICD9
      - LOINC (Logical Observation Identifiers Names and Codes)
      - CPT (Current Procedural Terminology)
      - RxNorm (Normalized names for clinical drugs)
    - Multiple languages (25)

      --English, Spanish, Chinese, French, German…
    - 16 million terms and 4.4 million concepts

Metathesaurus

- **Metathesaurus** (continue)
  - Concept - A meaning of a term
    - Concept Unique Identifier (CUI)
  - Semantic type – category of a concept

EX:

C0003862 – joint pain

Semantic type - sign or symptom

Synonyms - Painful joints, joint pains, joint pain, arthralgia.. (35 in English, 137 in all languages)

Metathesaurus

- **Metathesaurus** (continue)
  - More than 20 relationships between different concepts: **Child**, **parent**, **narrower, broader**, sibling, etc.
    - EX: child concepts of "joint pain" include (57 children):

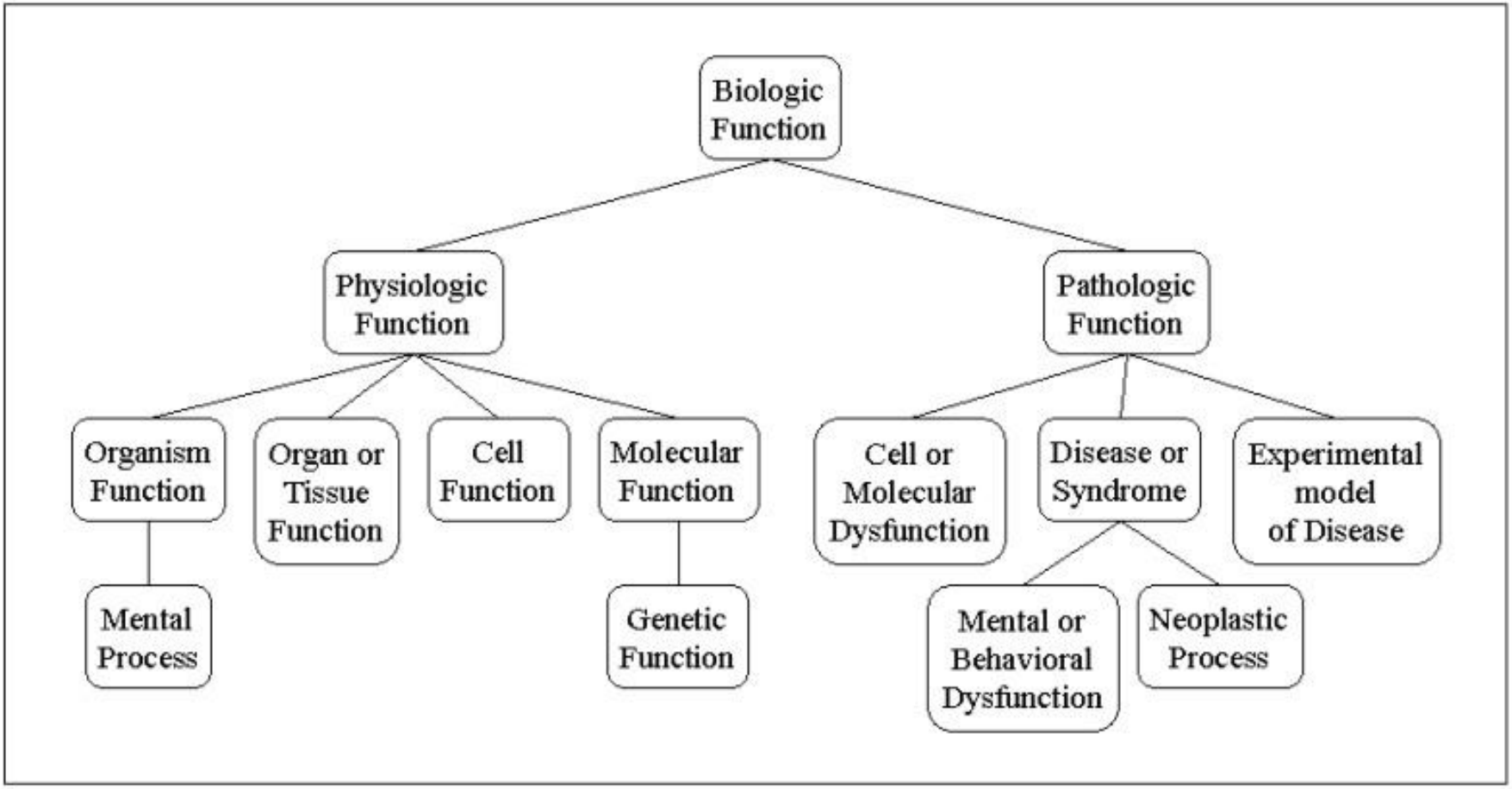
C0019559	Hip Joint Pain
C0037011	Shoulder Pain
C0162296	Polyarthralgia
C0221785	Wrist Arthralgia

UMLS

- **Semantic Types and Semantic Network**
  - 127 Semantic Types with hierarchy
    - **Diagnosis**, **sign or symptom**, **hormone**, **virus,**
      Disease or syndrome, body part, organ or organ part…
  - 54 Semantic Relationships – treat, cause…
    e.g. **virus** cause **disease or syndrome**
    drugs treat **disease or syndrome**
  - All the different semantic relationships build a semantic network

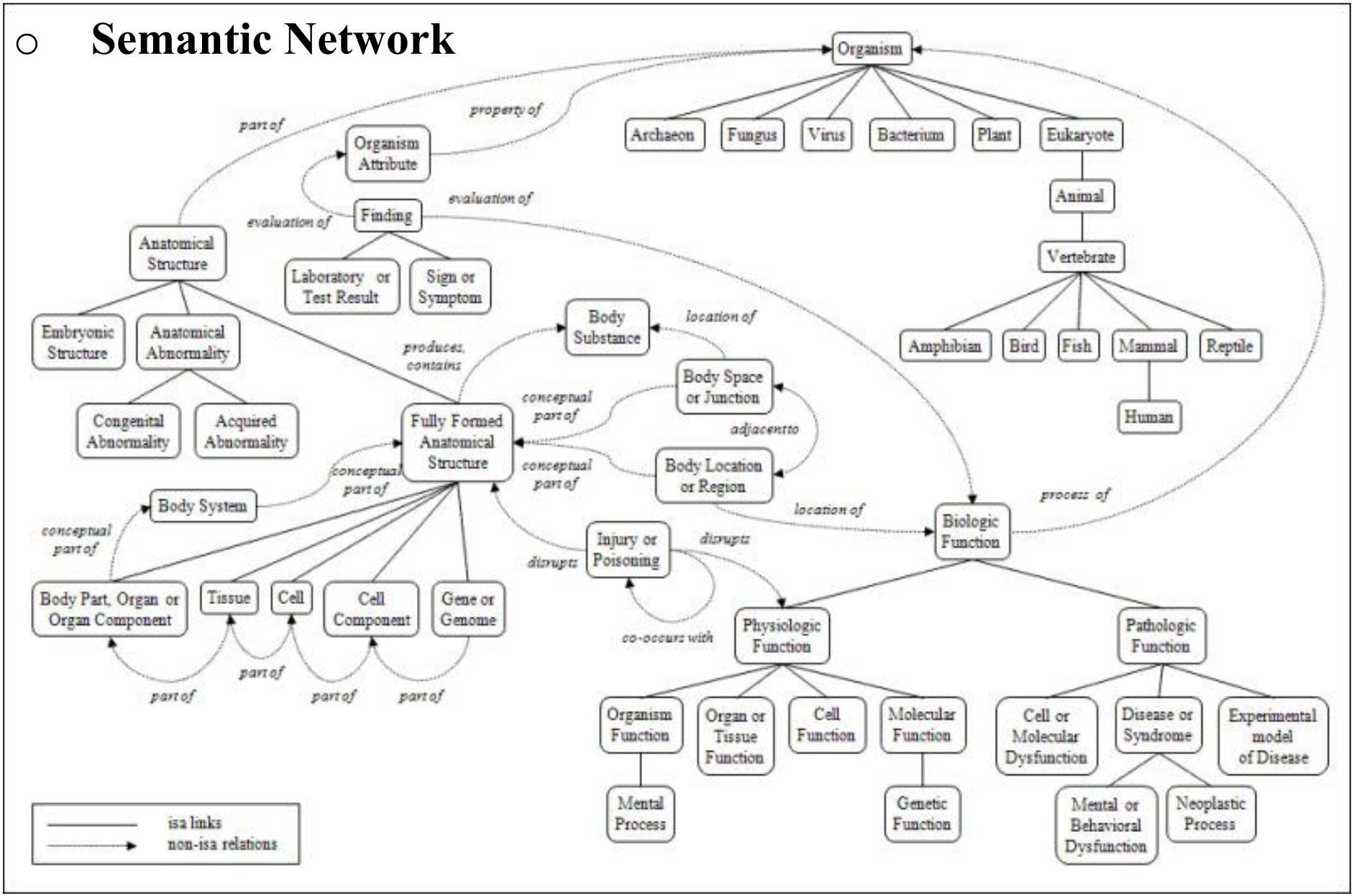
UMLS

- **Semantic Types and Semantic Network semantic Network**



UMLS

- **Semantic Network**



UMLS

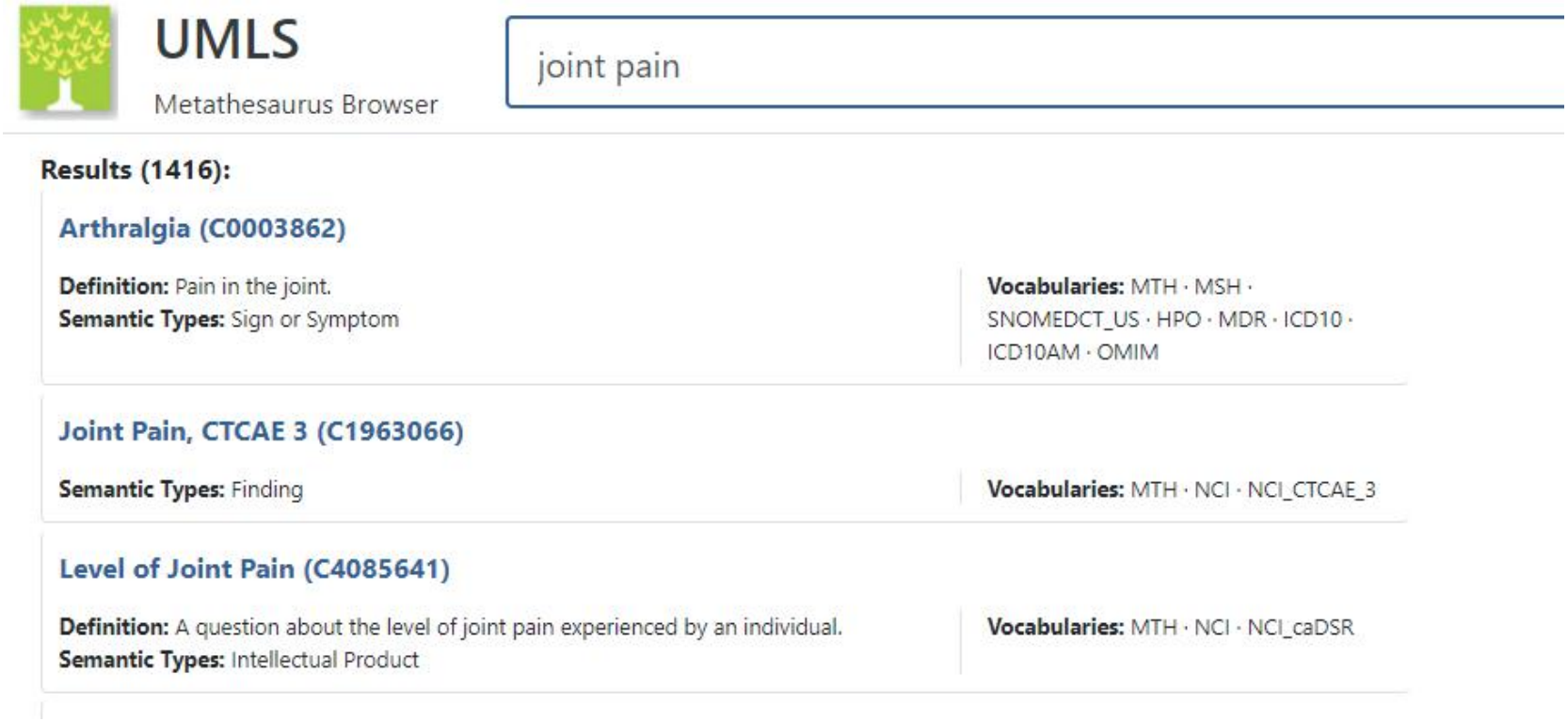
- **Semantic Types and Semantic Network**
  - Concept categorization or grouping
  - Concept inclusion or exclusion
    - 45 semantic types

How to use UMLS

- **UMLS Terminology Services account**
  - **License**
- **Online:**

  UMLS Metathesaurus Browser

  https://uts.nlm.nih.gov/uts/umls/home



How to use UMLS

- **UMLS REST API**
  https://documentation.uts.nlm.nih.gov/rest/home.html
  - Get CUIs and CUI names from terms
  - Get codes from terms
  - Map codes to CUIs
- **Local version**

https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

  - Offline
  - High-throughput by programming language

Dictionary Generation for Phenotyping

- Collect articles from 5 online knowledge sources
  - Wikipedia, Medline plus, Merck Manuel, Mayo clinic, Medscape

Rheumatoid arthritis

From Wikipedia, the free encyclopedia

For juvenile rheumatoid arthritis, see juvenile idiopathic arthritis.

Rheumatoid arthritis (**RA**) is a long-term autoimmune disorder that primarily affects joints.[1] It typically results in warm, swollen, and painful joints.[1] Pain and stiffness often worsen following rest.[1] Most commonly, the wrist and hands are involved, with the same joints typically involved on both sides of the body.[1] The disease may also affect other parts of the body, including skin, eyes, lungs, heart, nerves and blood.[1] This may result in a low red blood cell count, inflammation around the lungs, and inflammation around the heart.[1] Fever and low energy may also be present.[1] Often, symptoms come on gradually over weeks to months.[2]

While the cause of rheumatoid arthritis is not clear, it is believed to involve a combination of genetic and environmental factors.[1] The underlying mechanism involves the body's immune system attacking the joints.[1] This results in inflammation and thickening of the joint capsule.[1]

Dictionary Generation for Phenotyping

- Run a NER (Named Entity Recognition) software for mapping terms to CUIs
  - Identify

    Rheumatoid arthritis *(RA) is a long-term autoimmune disorder that primarily affects joints.[1] It typically results in warm, swollen, and painful joints.*
  - Mapping

Original terms	Mapped CUI	CUI name
painful joints	C0003862	Joint pain
Rheumatoid arthritis	C0003873	Rheumatoid arithris

Dictionary Generation for Phenotyping

- Run a NER (Named Entity Recognition) software for mapping terms to CUIs (continue)
  - Majority voting – exclude rare concepts
    - Appear in less than 3 articles

CUI	meds	merk	medl	mayo	wiki	Count
C0036916	0	1	0	1	1	3
~~C0013516~~	~~1~~	~~1~~	~~0~~	~~0~~	~~0~~	~~2~~
~~C0677635~~	~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~
C0003376	1	1	1	0	1	4

  - Build a dictionary using the remaining concepts
    - Using CUI to find all synonyms in UMLS

Dictionary Generation for Phenotyping

- **Final dictionary –** (155 CUIs, 5685 terms)

arthralgia | C0003862
joint pain| C0003862
arthritis | C0003864
rheumatoid arthritis | C0003873
stiffness | C0427008
il 6 receptor | C0063717
fever | C0015967
crp | C0006560
…

QUESTIONS

THANK YOU