Master en Santé publique (M2) – SITIS

Institut de Santé Publique d'Epidémiologie et de Développement

Université de Bordeaux, France

October 11, 2019

Interoperability and biomedical terminologies

Three examples of coverage and alignment studies

Olivier Bodenreider

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

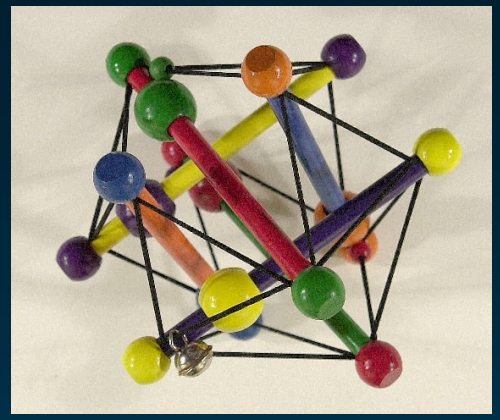
U.S. National Library of Medicine

Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.

Outline

- Interoperability of disease concepts in clinical and research ontologies – Contrasting coverage and structure in the Disease Ontology and SNOMED CT
  - Raje S, Bodenreider O.
  - Stud Health Technol Inform (Proc Medinfo) 2017:925-929.
- Comparing the representation of medicinal products in RxNorm and SNOMED CT – Consequences on interoperability
  - Nikiema J-N, Bodenreider O.
  - Proceedings of the 8th International Conference on Biomedical Ontology (ICBO 2019) 2019.
- Automatic construction of UMLS Metathesaurus with deep learning
  - Yip J, Bodenreider O



Medical Ontology Research

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

Olivier Bodenreider

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

U.S. National Library of Medicine

MEDINFO 2017
Hangzhou, China
August 23, 2017 – Session KM#1

Interoperability of Disease Concepts in Clinical and Research Ontologies

Contrasting Coverage and Structure in the Disease Ontology and SNOMED CT

Satyajeet Raje, Olivier Bodenreider

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

NIH U.S. National Library of Medicine

Motivation

Interoperability is critical for translational research

- Disease Ontology (DO)
  - Part of OBO; widely used in research projects
  - 6931 active disease concepts
  - Not integrated in UMLS (Some concepts mapped to SNOMED CT via "obo:hasDbXref")
  - August 2016 release used in this study (Available in OWL)
- SNOMED CT
  - Largest clinical terminology
  - About 100,000 concepts in the "Clinical findings" hierarchy
  - Integrated in UMLS
  - March 2016 US edition used in this study. (Converted to OWL)

Objectives

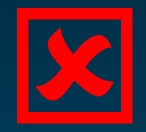
- To investigate the coverage of disease concepts between the Disease Ontology (DO) and SNOMED CT
  - To identify and characterize the concepts present in DO but not covered by SNOMED CT
  - To analyze the differences in hierarchical structure between the two ontologies

Methods

- Establishing a reference set of mappings
  - Apply semantic constraints on existing mappings from DO to SNOMED CT
  - Find additional mappings lexically
- Characterizing unmapped DO concepts
  - Distribution of mapped vs. unmapped concepts by top-level hierarchies in DO
  - Analysis of connected components of unmapped concepts
  - Manual review of semantic “differentiae” for unmapped concepts

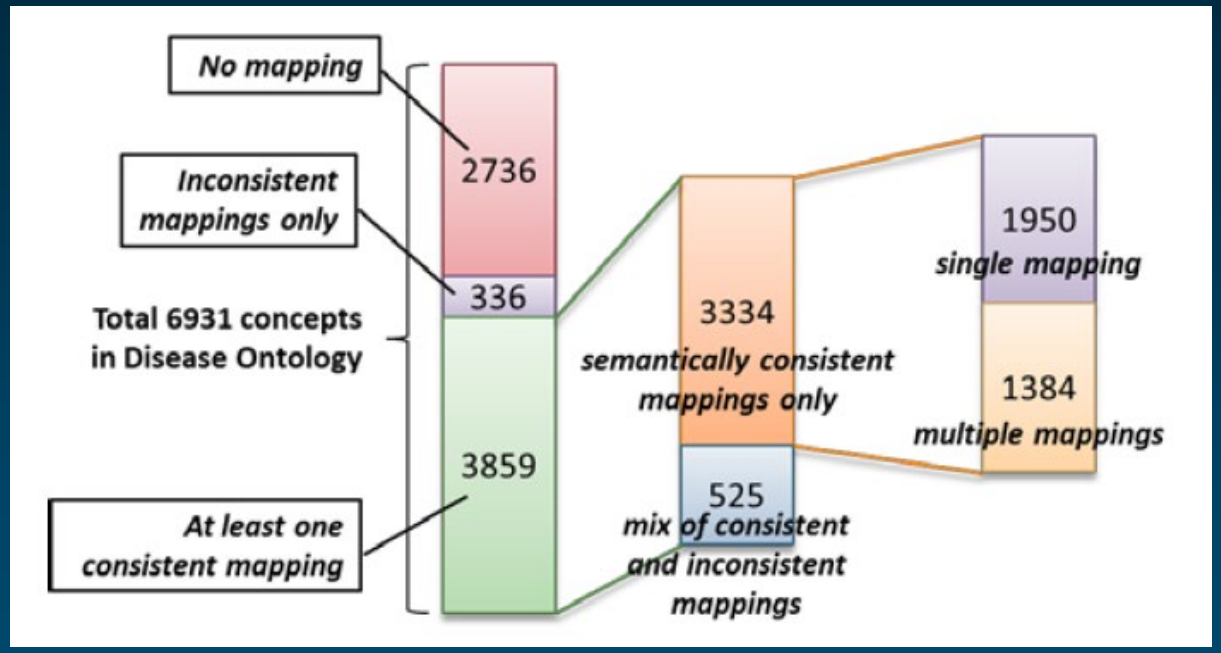
Establishing a reference set of mappings

- Semantic constraints on DO mappings ("xrefs")
  - Filter out mappings outside "Clinical findings"
    - transitional cell carcinoma → Transitional cell carcinoma (morphologic abnormality)
  - Resolve mappings to obsolete SNOMED CT concepts
    - Tympanosclerosis → SCTID: 111540000 *** obsolete ***
      → SCTID: 23606001 Tympanosclerosis (disorder)
- Additional lexical mappings
  - Leverage synonymy in the UMLS
    - rheumatic heart disease → Rheumatic heart disease (disorder)



Reference set of mappings

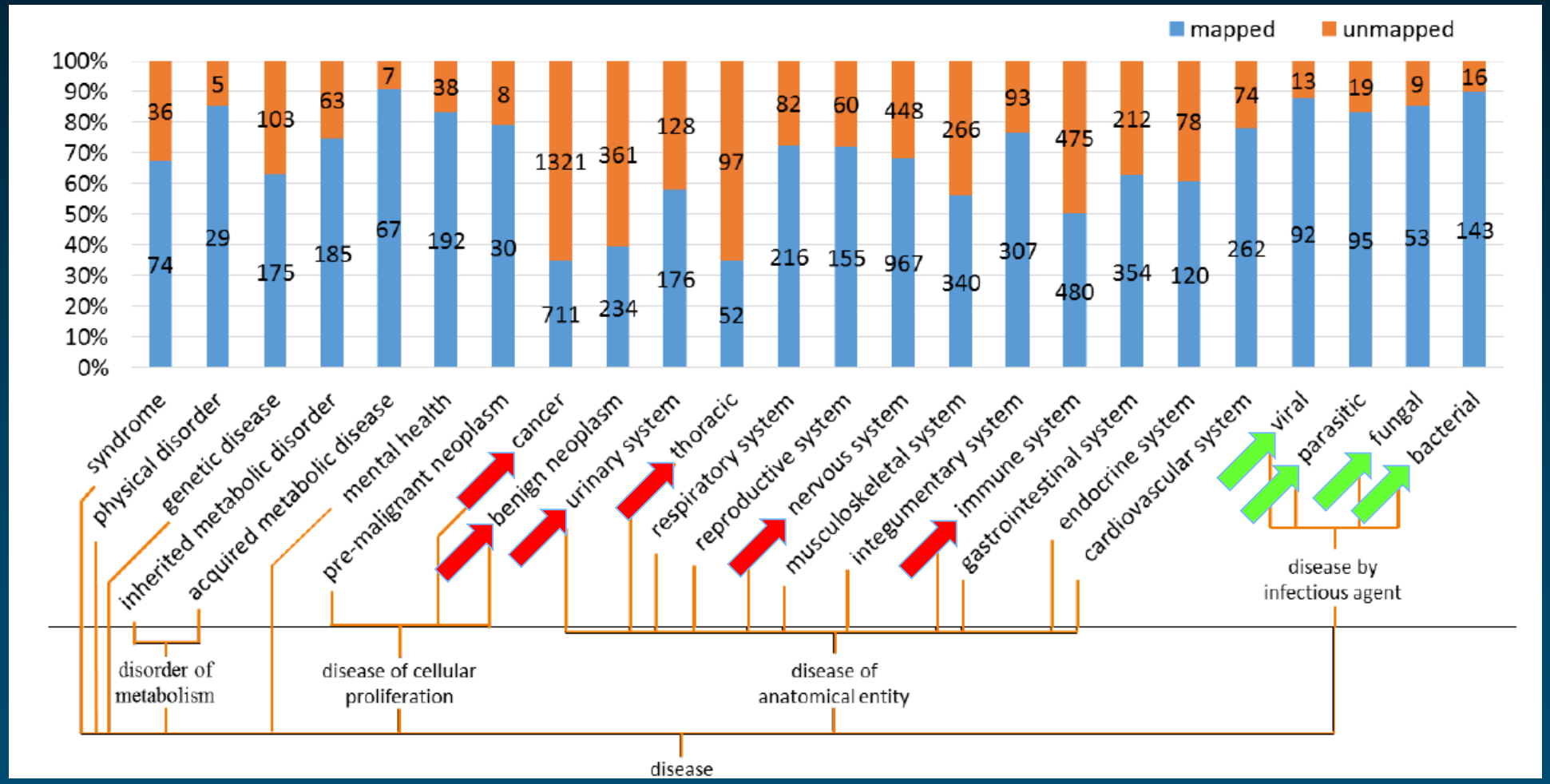
- Semantic constraints on existing DO mappings



3859 (56%) DO concepts with at least one semantically consistent mapping

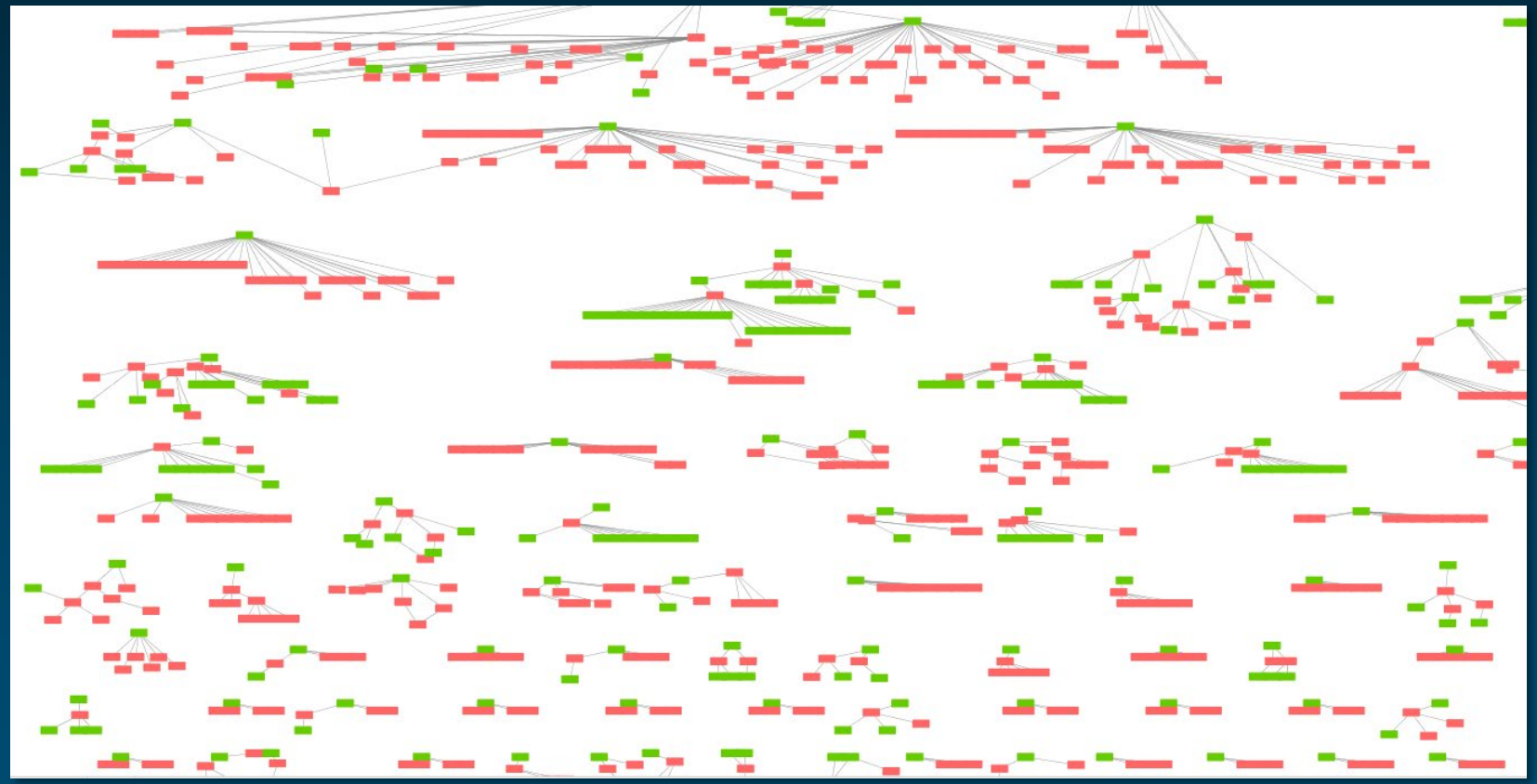
- 619 (9%) additional lexical mappings
- 2453 (35%) DO concepts remain unmapped

Characterizing unmapped DO concepts (semantically)



Characterizing unmapped DO concepts (structurally)

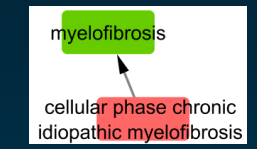
Mapped DO concepts Unmapped DO concepts



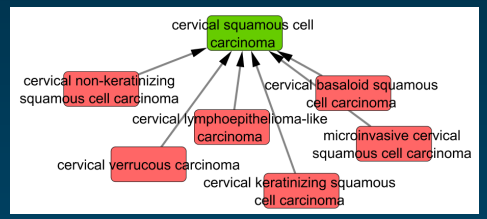
Visualization in Cytoscape

Characterizing unmapped DO concepts (structurally)

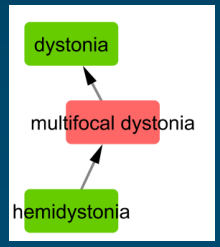
- 401 cases of *isolated unmapped leaf concepts*
  - multiple mucosal neuroma unmapped leaf of neuroma



- 1806 unmapped concepts in *subtrees of unmapped concepts*
  - subtree rooted at chromosomal deletion syndrome contains 35 concepts, including distal 10q deletion syndrome and chromosome 15q11.2 deletion syndrome



- 246 cases of *unmapped intermediary concepts*
  - intermediary "grouper" concepts in DO
  - multifocal dystonia between dystonia (parent) and hemidystonia (child)



Characterizing unmapped DO concepts (differentiae from parent concepts)

Type of differentia	Count	%
specific morphology (e.g. ***follicular*** dendritic cell sarcoma)	831	37.65
morphology and anatomic site	520	23.56
specific subtype X (e.g. spinocerebellar ataxia ***type 1***)	253	11.46
anatomic site (e.g. ***intramuscular*** hemangioma)	147	6.66
morphology and period of onset	61	2.76
period of onset (e.g. ***pediatric*** osteosarcoma)	45	2.04
chromosomal location and anomaly	45	2.04
complex syndrome (e.g. agnathia-otocephaly complex)	42	1.90
other generic subtypes	42	1.90
organism (e.g. ***screw worm*** infectious disease)	30	1.36
others	*191*	*8.65*
Total	2207	

Discussion

- DO has 2453 potentially "new" concepts
  - Adding some semantic differentia
- Pre- vs. post-coordination of concepts
  - Some of the DO concepts could be expressed in SNOMED CT through post-coordination
- Multiple mappings are not good for interoperability
  - Further work needed to formulate rules to resolve such mappings
- Limitation
  - No evaluation of the mappings
    - Inclusion in the "Clinical Findings" hierarchy was the only validation criteria for existing mappings

Conclusion Practical contribution

- When using both DO and SNOMED CT
  - "Better" set of mappings
    - Removed invalid mappings
    - Added missing mappings (lexical match)
- Choosing between DO and SNOMED CT
  - Characterization of specific content in DO
    - Semantic categorization
    - Hierarchical organization

ICBO 2019
August 1, 2019
University at Buffalo

10th International Conference on Biomedical Ontology

Comparing the representation of medicinal products in RxNorm and SNOMED CT

Consequences on interoperability

Jean-Noël Nikiema & Olivier Bodenreider

National Institutes of Health, Bethesda, Maryland, USA

U.S. National Library of Medicine

Motivation

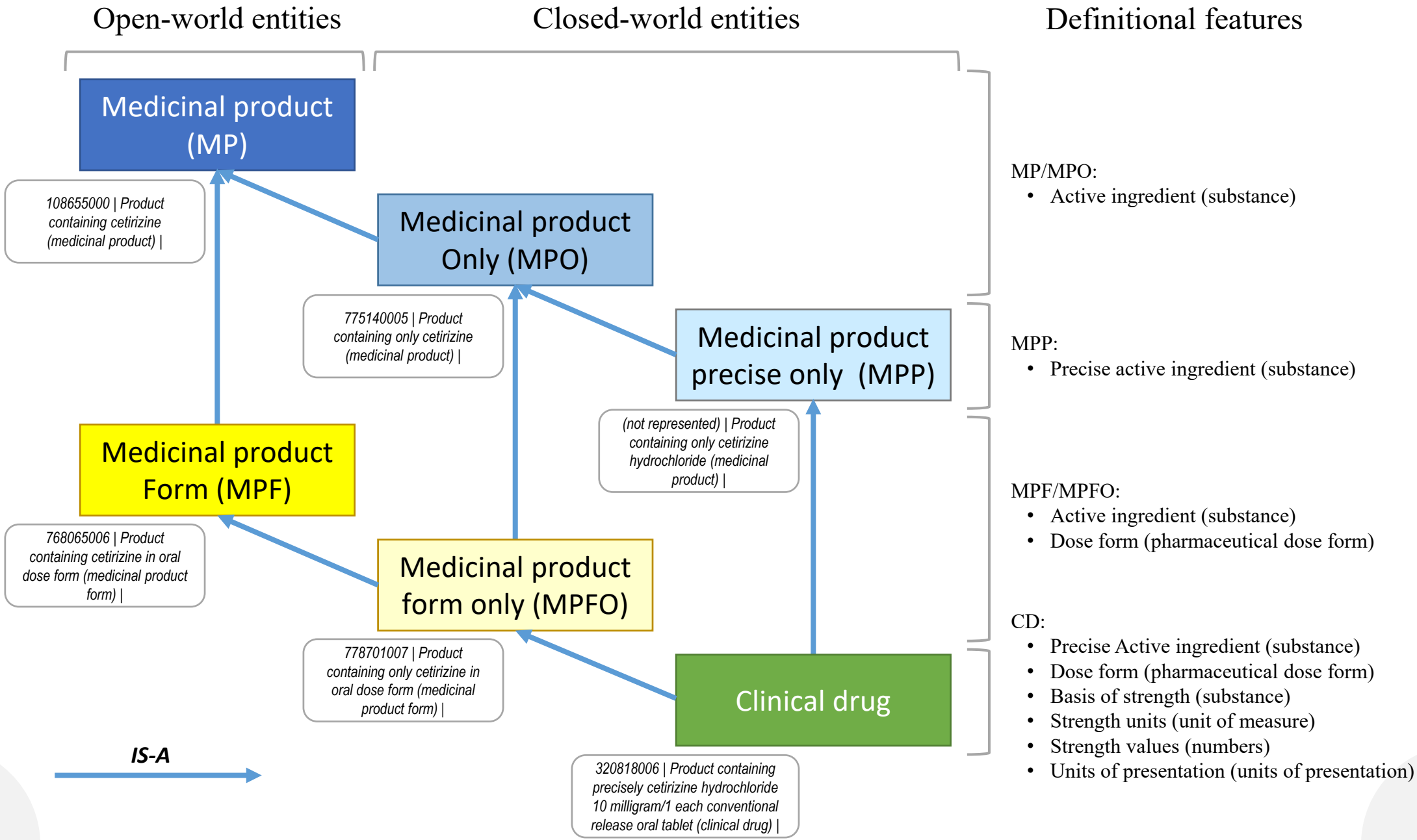
- Different drug terminologies use different models for the representation of medicinal products
- Based on similar definitional features
  - Active ingredient/BoSS atorvastatin
  - Strength 10 mg
  - Dose form oral tablet
- Differences
  - Formalism
  - Compliance with international standards
  - Scope (e.g., country-specific information)
- Are the RxNorm and SNOMED CT drug models interoperable?

Objectives

- To compare the representation of medicinal products in RxNorm and SNOMED CT
  - To analyze their similarities and differences
  - To assess the consequences of these differences on interoperability between the two terminologies

Background: SNOMED CT

- Largest clinical terminology in the world
- Developed by a consortium of over 40 countries
- New model for the representation of medicinal products in 2018
  - Including drug-class membership information
- Integrates requirements from ISO standard IDMP – Identification of Medicinal Products
  - Clinical drugs represented in a closed worldview
  - Dose forms in reference to EDQM – European Directorate for Quality in Medicines
  - Units aligned with UCUM – Unified Code for Units of Measure
- Formalism: description logic
- Scope: generic drugs (excludes branded drugs and packs – country-specific)
- 6 types of entities, with 5 definitional features


Open-world entities
Closed-world entities
Definitional features
Medicinal product (MP)
108655000 | Product containing cetirizine (medicinal product) |
Medicinal product Only (MPO)
775140005 | Product containing only cetirizine (medicinal product) |
Medicinal product precise only (MPP)
(not represented) | Product containing only cetirizine hydrochloride (medicinal product) |
Medicinal product Form (MPF)
768065006 | Product containing cetirizine in oral dose form (medicinal product form) |
Medicinal product form only (MPFO)
778701007 | Product containing only cetirizine in oral dose form (medicinal product form) |
Clinical drug
320818006 | Product containing precisely cetirizine hydrochloride 10 milligram/1 each conventional release oral tablet (clinical drug) |
IS-A
MP/MPO:
• Active ingredient (substance)
MPP:
• Precise active ingredient (substance)
MPF/MPFO:
• Active ingredient (substance)
• Dose form (pharmaceutical dose form)
CD:
• Precise Active ingredient (substance)
• Dose form (pharmaceutical dose form)
• Basis of strength (substance)
• Strength units (unit of measure)
• Strength values (numbers)
• Units of presentation (units of presentation)

Background: RxNorm

- U.S. standard for drug terminology
- Developed by the National Library of Medicine
- Simple model: ingredient + strength + dose form
  - Enriched over time with 2 optional features
    - Quantitative factor — 2 ML Furosemide 10 MG/ML Injection
    - Qualitative distinction — Abuse-Deterrent Oxycodone Hydrochloride 15 MG Oral Tablet
- Formalism: graph representation
- Scope: both generic and branded drugs, including packs
- 4 types of entities (for generic drugs), 5* definitional features

Examples

RxNorm generic drug entities

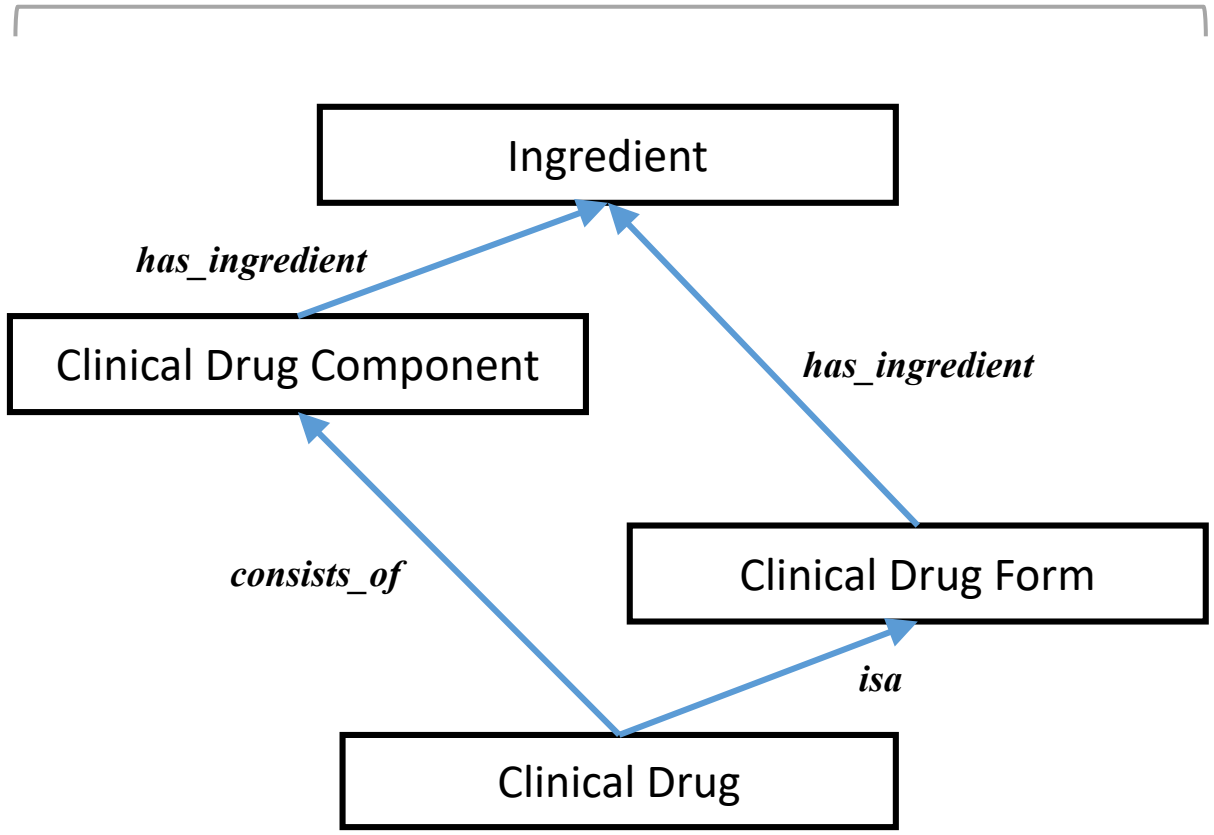
Definitional features

Cetirizine [RxCUI = 20610]

cetirizine hydrochloride 10 MG [RxCUI = 1011480]

Cetirizine Oral Tablet [RxCUI = 371364]

cetirizine hydrochloride 10 MG Oral Tablet [RxCUI = 1014678]



IN:
- Ingredient

SCDC:
- Ingredient
- Strength

SCDF:
- Ingredient
- Dose form

SCD:
- Ingredient/BoSS
- Strength
- Dose form
- Quantity factor (optional)
- Qualitative distinction (optional)

Similarities and differences: Overview

- Major definitional features are common to both models
  - Active ingredient
    - Substance vs. medicinal product; substance modification
  - Strength
    - Concentration strength vs. presentation strength
  - Dose form
    - Dose form vs. unit of presentation
- Specific features in SNOMED CT
  - Explicit closed worldview for clinical drugs
- Specific features in RxNorm
  - Optional qualitative distinction; materialized SCDC (for navigation)

Differences: Ingredients

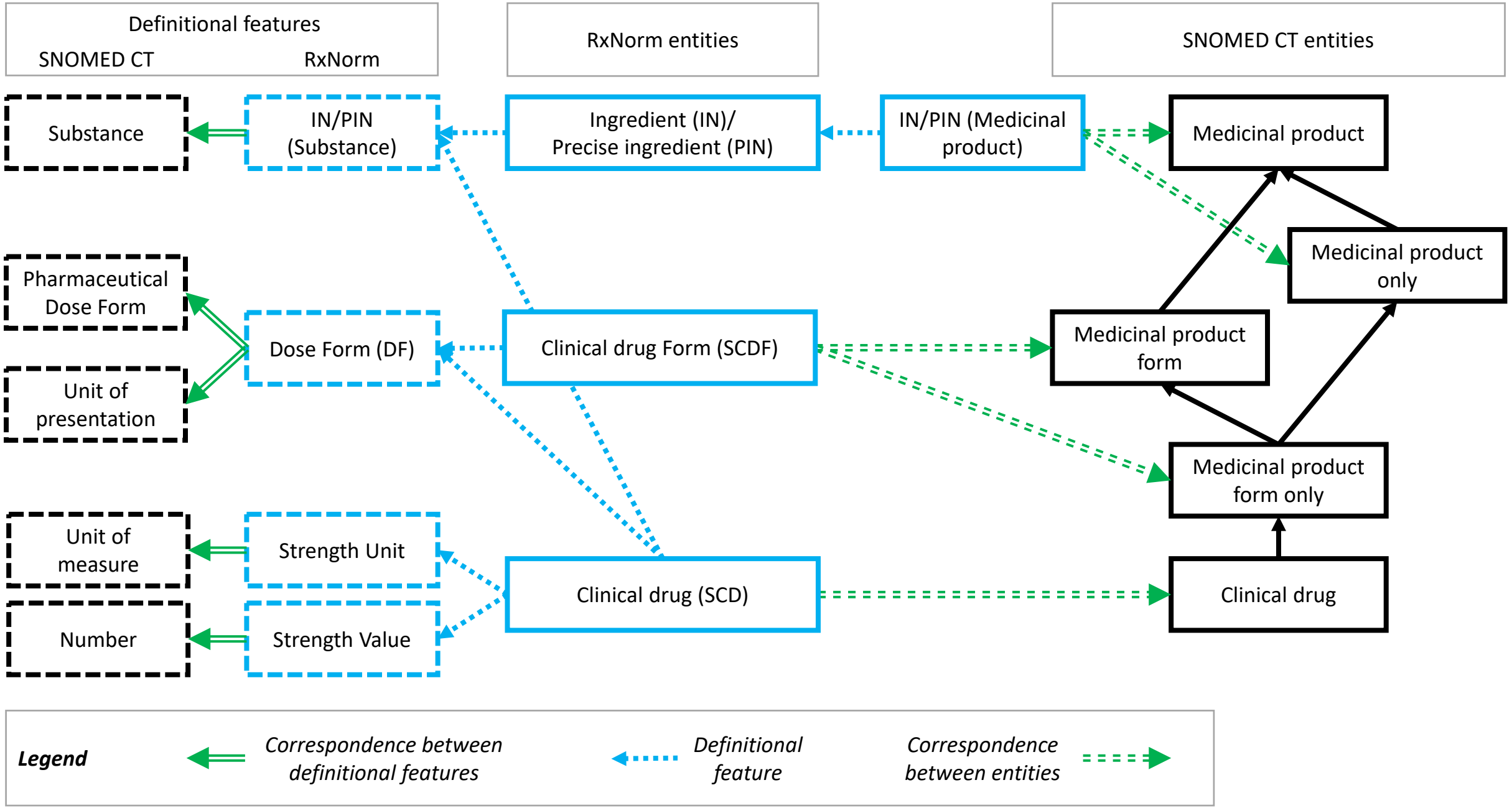
- Substance vs. medicinal product
  - RxNorm: single entity
    - Cetirizine
  - SNOMED CT: distinct entities
    - Medicinal product *has_active_ingredient* Substance
    - Medicinal product: Product containing cetirizine (medicinal product)
    - Medicinal product "only": Product containing only cetirizine (medicinal product)
    - Substance: Cetirizine (substance)
- Substance modification
  - RxNorm: different types of entities (Ingredient vs. Precise ingredient)
    - PIN: Cetirizine hydrochloride *precise_ingredient_of* IN: Cetirizine
  - SNOMED CT: same kind of entity + *modification_of* relation
    - Cetirizine hydrochloride (substance) *modification_of* Cetirizine (substance)

Differences: Concentration vs. presentation strength

- RxNorm
  - Concentration strength (default)
    - 2 ML Furosemide 10 MG/ML Injection
  - Supports presentation strength through “prescribable name”
    - furosemide 20 MG in 2 ML Injection
  - Presentation strength can be computed with QF * concentration strength
    - 2 ML * 10 MG/ML = 20 MG/2 ML
- SNOMED CT (depending on unit of presentation)
  - Concentration strength (only) 10 MG/1 ML
  - Presentation strength (only) 20 MG/2 ML
  - Concentration strength + Presentation strength

Differences: Dose form vs. unit of presentation

- RxNorm
  - Dose form includes unit of presentation (implicitly) Oral Tablet
- SNOMED CT
  - Distinct dose form and unit of presentation
    - Dose form Conventional release oral tablet
    - Unit of presentation Tablet


Definitional features
SNOMED CT
RxNorm
RxNorm entities
SNOMED CT entities
Substance
IN/PIN (Substance)
Ingredient (IN)/ Precise ingredient (PIN)
IN/PIN (Medicinal product)
Medicinal product
Medicinal product only
Pharmaceutical Dose Form
Dose Form (DF)
Clinical drug Form (SCDF)
Medicinal product form
Unit of presentation
Medicinal product form only
Unit of measure
Strength Unit
Clinical drug (SCD)
Clinical drug
Number
Strength Value
Legend
Correspondence between definitional features
Definitional feature
Correspondence between entities

Findings: Similarities and differences

- SNOMED CT
  - More rigorous
  - Better aligned with international standards
  - Differences tend to be made explicit
  - More complex model
- RxNorm
  - Contains implicit knowledge, simplifications and ambiguities
  - Simpler model

Findings: Consequences on interoperability

- Can RxNorm be translated into SNOMED CT?
  - Yes, for the most part
- Specifically
  - Ingredients
    - Trivial disambiguation
  - Strength
    - Different editorial conventions for units (minor)
    - Presentation strength / Concentration strength / Both (depending on unit of presentation)
  - Dose form – requires detailed analysis to identify dose form and unit of presentation

Conclusions

- Similarities and differences between the representation of medicinal products in RxNorm and SNOMED CT
- Both models share major definitional features including ingredient (or substance), strength and dose form
- Subtle differences between the two models
- Translation of RxNorm into SNOMED CT is possible, but not straightforward

Automatic Construction of UMLS Metathesaurus with Deep Learning

Joey Yip
Olivier Bodenreider

NIH U.S. National Library of Medicine

Unified Medical Language System (UMLS) Metathesaurus

- Started in 1986 by the National Library of Medicine (NLM)

Overcome barriers to effective retrieval of machine-readable information

- The variety of ways the same concepts are expressed by different terminologies

(MeSH, MedDRA, RxNORM, ICD-10, SNOMED CT, etc)

- **~ 10** million English medical terms
- From **210** source vocabularies
  - **General**
    - *Anatomy (FMA, Neuronames), drugs (RxNorm, ATC, First DataBank), medical devices (UMD, SPN), clinical terms (SNOMED CT), information sciences (MeSH), administrative terminologies (ICD-9/10)*
  - **Specialized**
    - *Nursing (NIC), psychiatry (DSM, APA), adverse reactions (MedDRA)*
- Grouped into **~ 3.85** million concepts

Used in areas such as patient care, clinical coding, information retrieval, knowledge exploration, and data mining

Integrating Subdomains

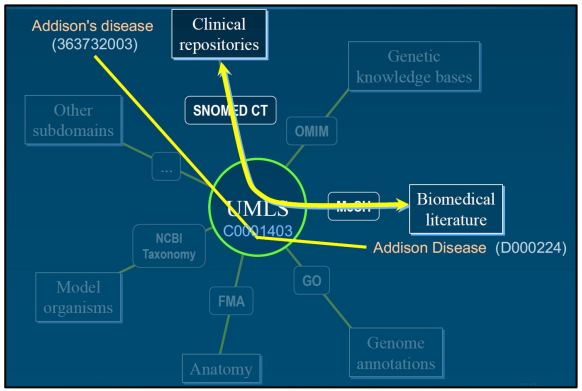
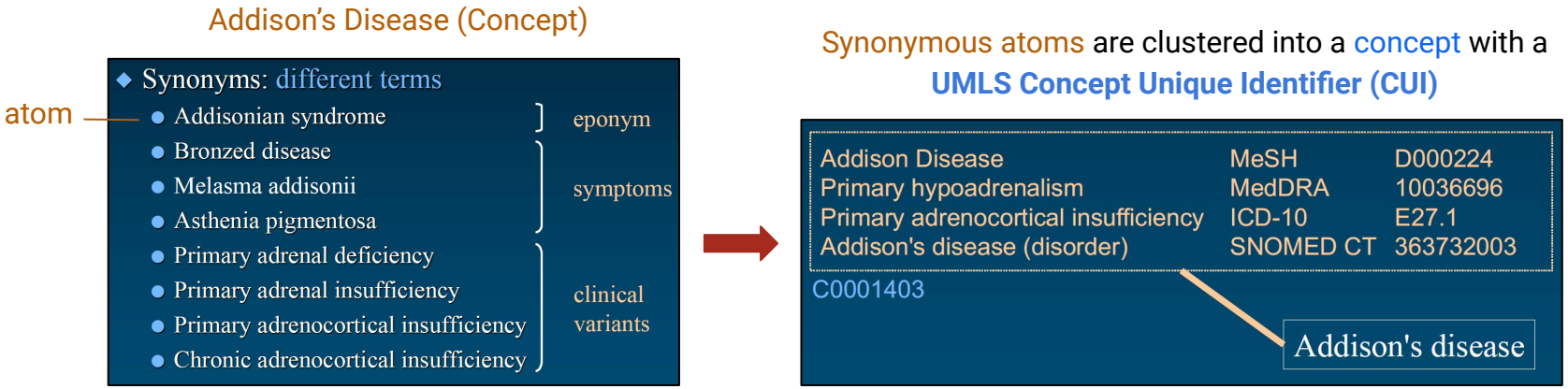


*Image from **Unified Medical Language System Overview** by Olivier Bodenreider*

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.

Unified Medical Language System (UMLS)



*Images from **Unified Medical Language System Overview** by Olivier Bodenreider*

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.

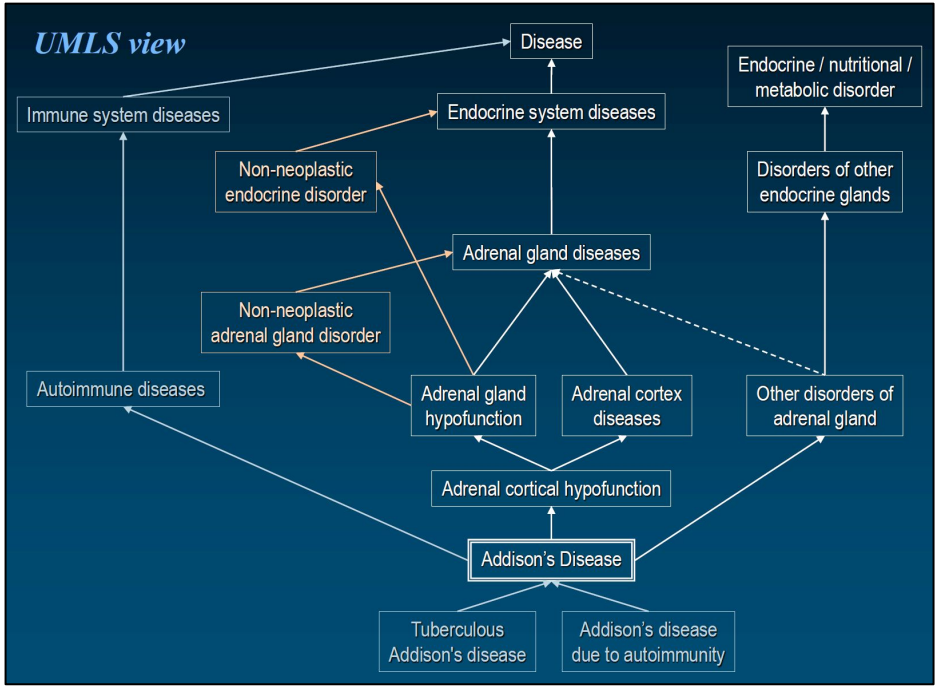
Construction of UMLS Metathesaurus *(Updates bi-annually)*

- **Lexical Knowledge**
  - *Lexical Variant Generator (LVG)*

Adrenal gland diseases
Adrenal disorder
Disorder of adrenal gland
Diseases of the adrenal glands
C0001621

adrenal disease gland

- **Semantic Pre-processing**
- **UMLS Human Editors**



*Images from **Unified Medical Language System Overview** by Olivier Bodenreider*

Motivation

The current approach in adding new resources from identifying lexical variants to manual audits can be both arduous and time-consuming.

*(**~ 10** million English medical terms, **~ 3.85** million concepts)*

Objectives

The project explores the realm of *supervised machine learning approach (Deep Learning)* to

1. Identify **synonymy** and **non-synonymy** among UMLS concepts at the atom level
    - Given two atoms, are they synonymous (same CUI)?
2. Investigate Deep Learning approach could emulate the current building process

Problem Formulation

~~**Approach 1 (Classification task):**~~

- **Training Data:** ~ 10M English language atoms and each with its own CUI assignment
- We can train a classification model to predict which CUI should be assigned to a given "new" atom *(since atoms having the same CUI are synonymous).*
- Input: Atom -> Output label: CUI
- **Challenge:** ~ 3.85M softmax outputs (extreme classification task)

Approach 2 (Similarity task):

- Learn **similarities** between atoms within a CUI and **dissimilarities** between atoms from different CUIs.

A fully-trained model should identify and learn scenarios where

Lung disease and disorder	Two atoms that are ***lexically similar*** in nature but ***are not synonymous***	**Head** disease and disorder
Addison's disease	Two atoms that are ***lexically dissimilar*** but are ***synonymous***	Primary adrenal deficiency

Traditional Neural Network Architecture

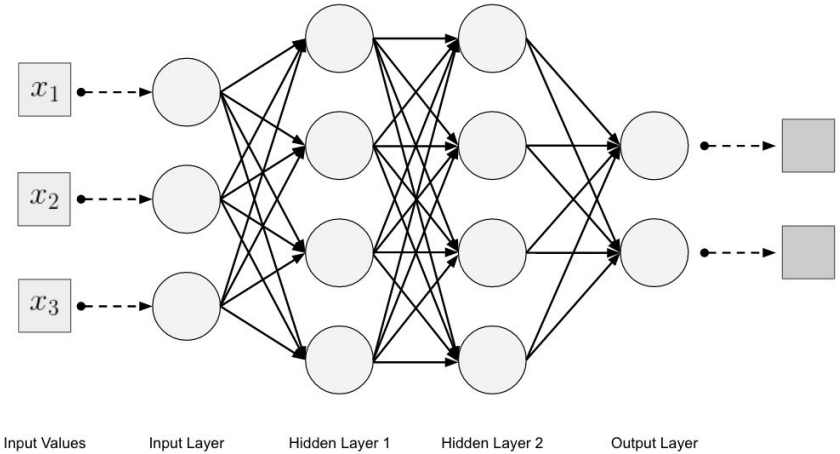


Image source: https://www.oreilly.com/library/view/deep-learning/9781491924570/ch04.html

Feedforward Neural Network (Multilayer Perceptron):
Not suited for Pairwise-similarity task

Siamese (Twin) Neural Network

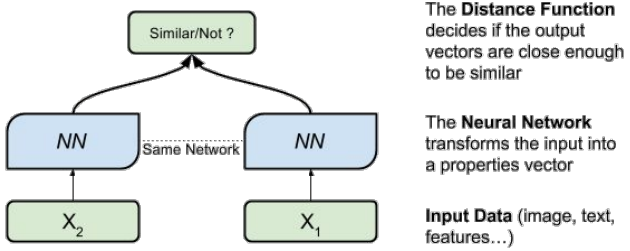


Image source:
https://aws.amazon.com/blogs/machine-learning/combining-deep-learning-networks-gan-and-siamese-to-generate-high-quality-life-like-images/

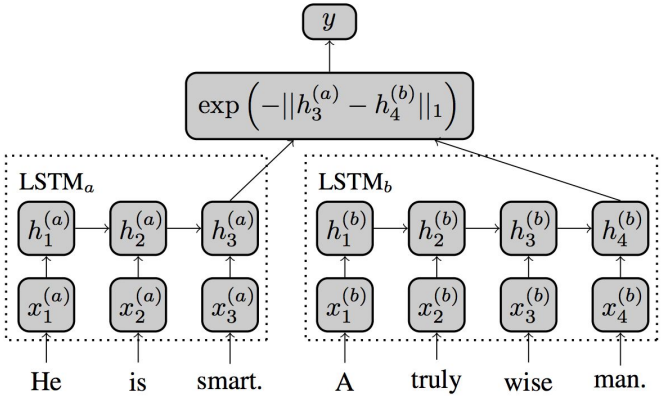
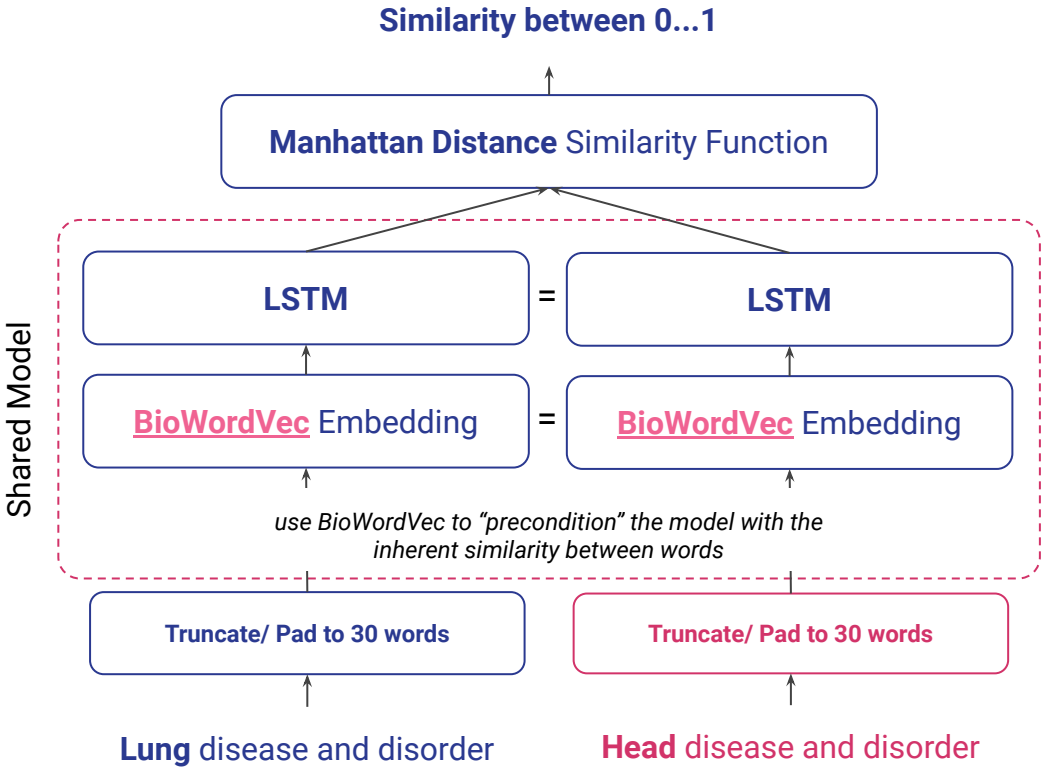
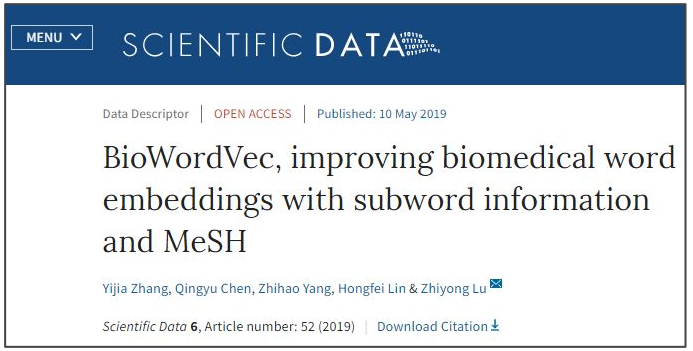


Image source: Siamese Recurrent Architectures for Learning Sentence Similarity

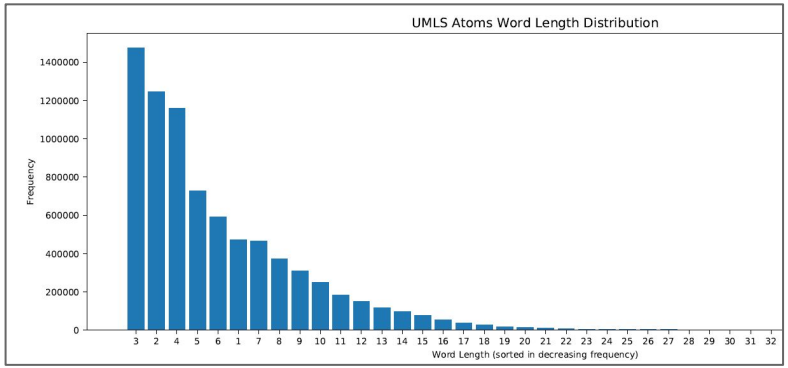
Mueller, J., & Thyagarajan, A. (2016, March). Siamese recurrent architectures for learning sentence similarity. In Thirtieth AAAI Conference on Artificial Intelligence.

Siamese-LSTM





Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific Data. 2019.



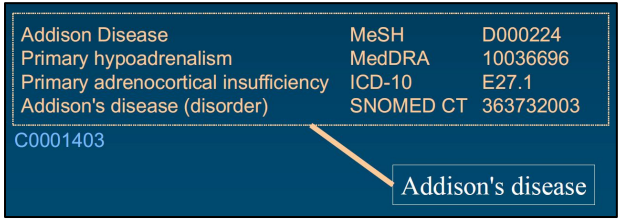
UMLS Atoms Word Length Distribution
(Word length 30 covers 97% of atoms in the UMLS)

Mueller, J., & Thyagarajan, A. (2016, March). Siamese recurrent architectures for learning sentence similarity. In Thirtieth AAAI Conference on Artificial Intelligence.

Dataset (2019-AA UMLS) and Feature Engineering

Positive Pairs (Synonyms)

- (CUI)-asserted synonymy between atoms (~15 million pairs) ✓



Negative Pairs (Non-synonyms)

- Ideally, we want to generate all negative pairs (1 atom against atoms from other non-related CUIs) (~ 10 million atoms * 10 million atoms) ✗

Class Imbalance: Number of *Non-synonyms* > Number of *Synonyms*

Intuition: What we want are interesting negative pairs that are lexically similar but differ in semantics.

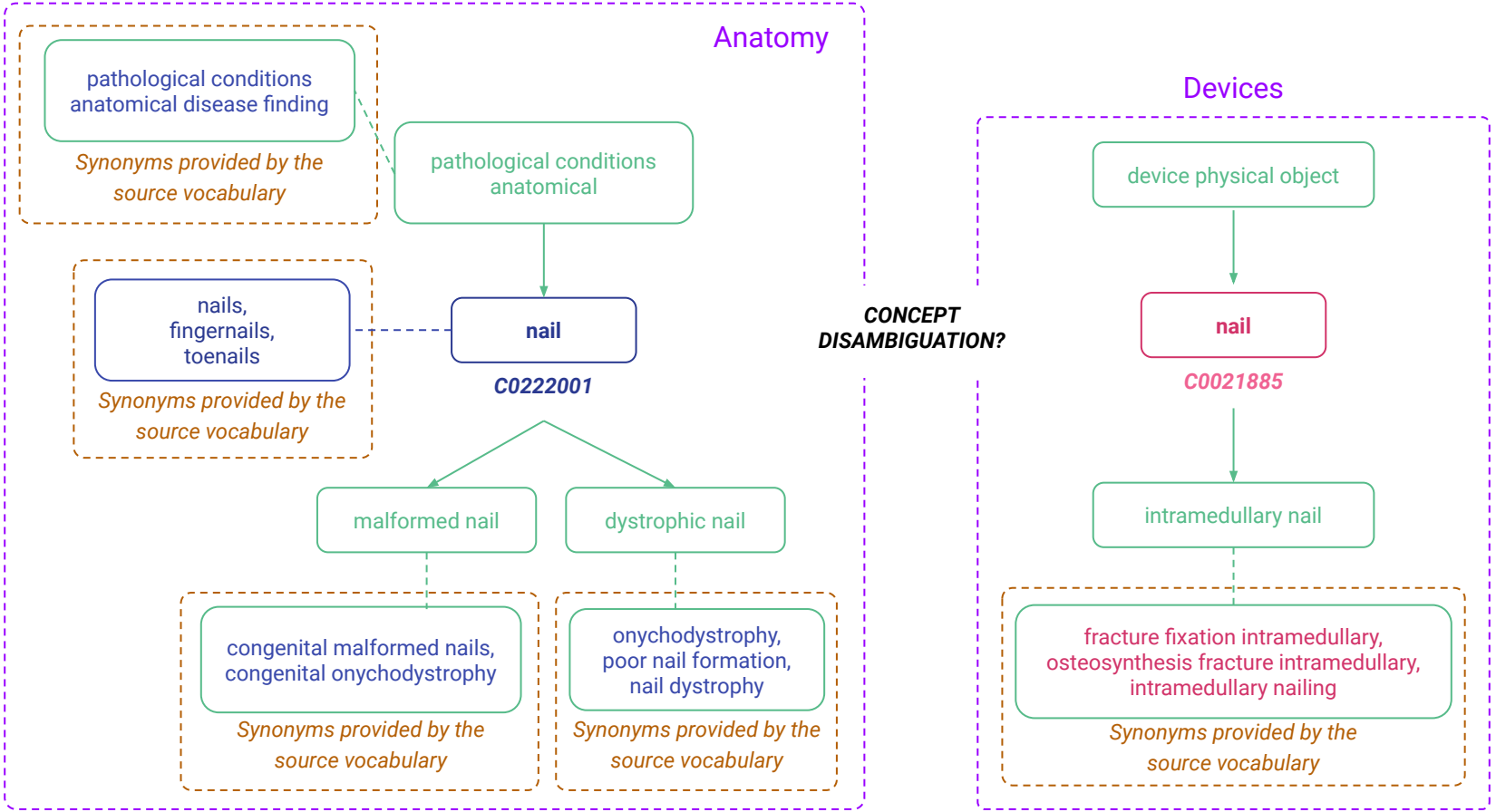
- Heuristic Approach: Use **Jaccard Index** to generate **negative pairs** for atoms with **high Jaccard Similarity** *(Sort and filter top ~15 million pairs)* ✓

C0000473:
Product containing ***para-aminobenzoic*** acid

C0038784:
Product containing ***sulfuric*** acid

Intersection (3)/ Union (5)
= **0.6 Jaccard Index**

Going beyond atoms... Let's Contextualize!



Experimental Setup

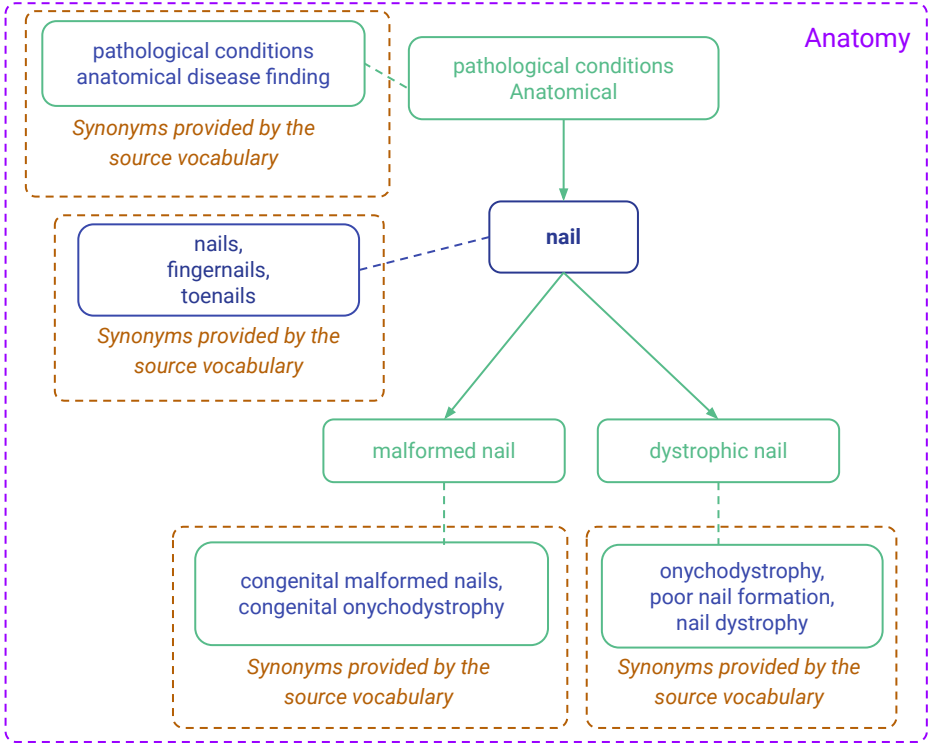
1. "Base" experiment
(Atom lexical features)

2. "Base" (Atom lexical features)
+ Synonyms provided by the source vocabulary

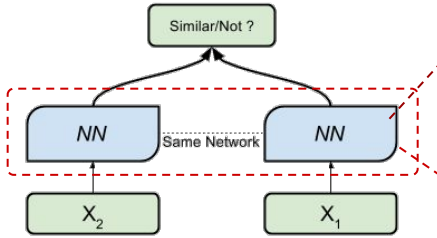
3. "Base" (Atom lexical features)
+ Hierarchical-Context(atom)
+ Semantic Group

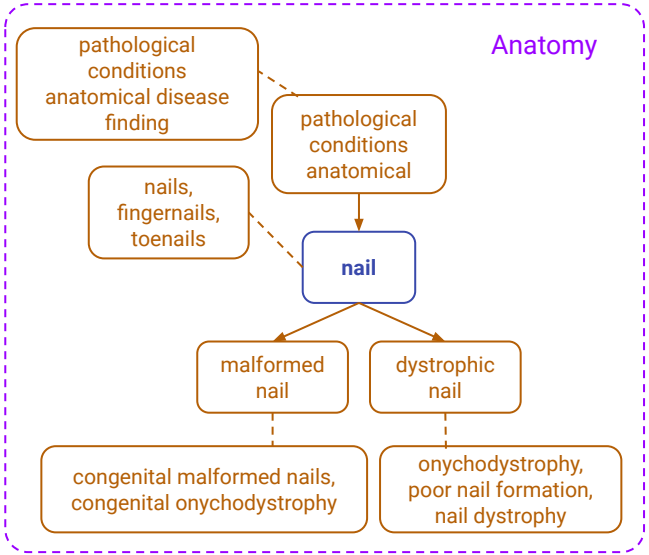
4. "Base" (Atom lexical features)
+ Synonyms provided by the source vocabulary
+ Hierarchical-Context(atom)
+ Semantic Group

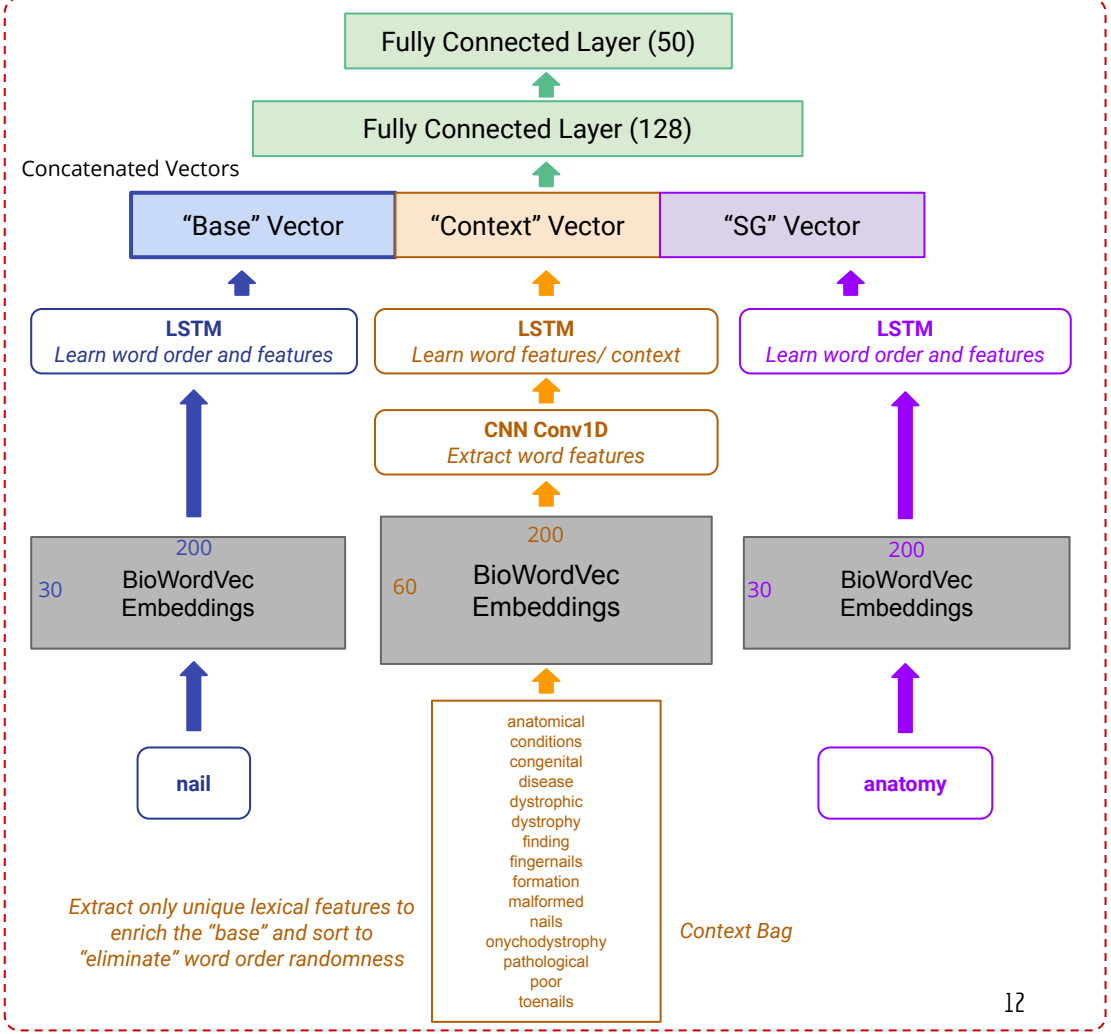
5. "Base" (Atom lexical features)
+ Synonyms provided by the source vocabulary
+ Hierarchical-Context(atom)
+ Synonyms of the Hierarchical-Context(atom)
+ Semantic Group



Architecture

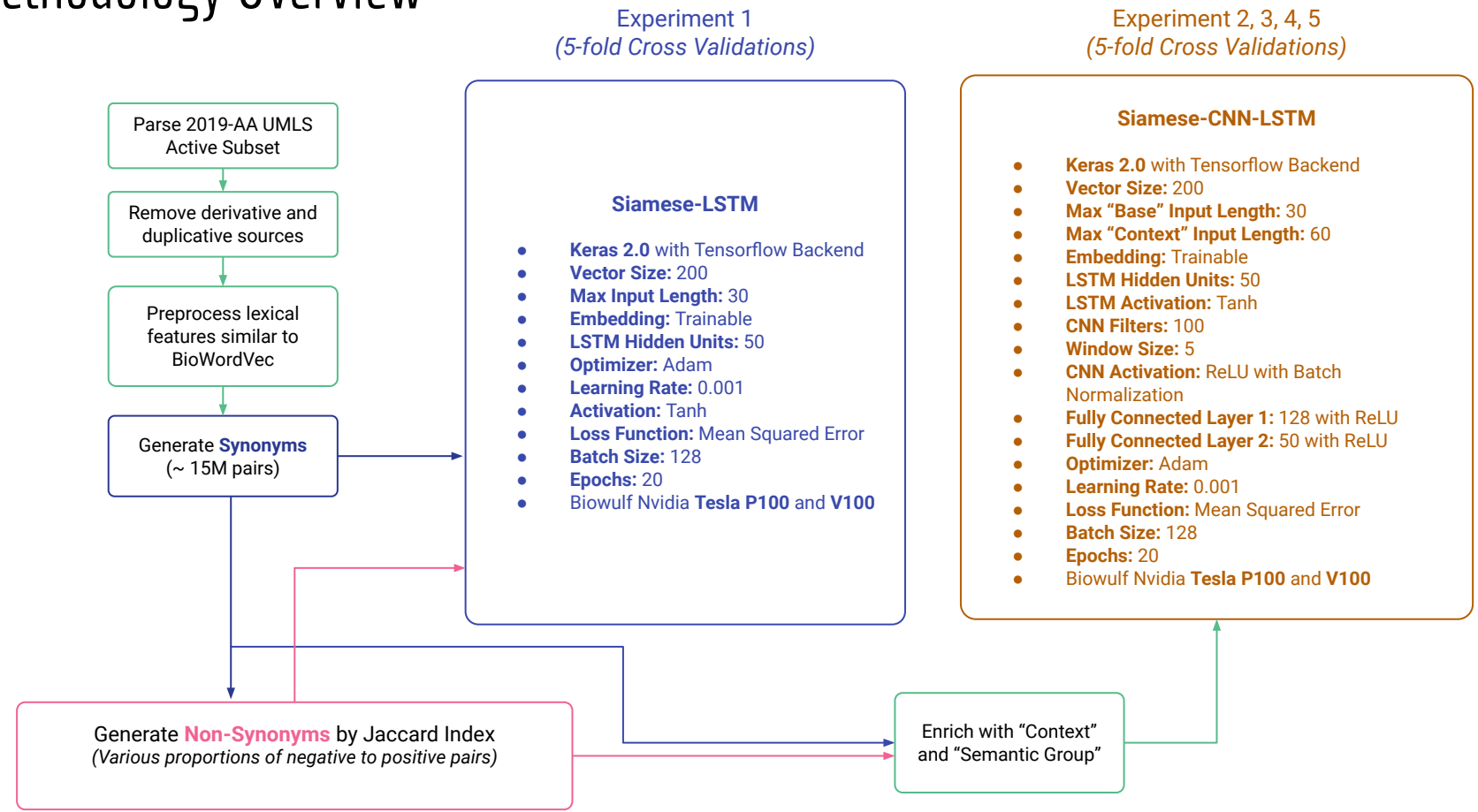






Pontes, E. L., Huet, S., Linhares, A. C., & Torres-Moreno, J. M. (2018). Predicting the Semantic Textual Similarity with Siamese CNN and LSTM. arXiv preprint arXiv:1810.10641.

Methodology Overview



Results & Evaluations *(based on optimal runs)*

Model/ Performance Metrics	Base	Base + Source Synonymy	Base + Hier. Context + Semantic Group	Base + Source Synonymy + Hier. Context + Semantic Group	Base + Source Synonymy + Hier. Context + Hier. Source Synonymy + Semantic Group
Accuracy	0.9333	0.8720	0.9486	0.9520	**0.9541**
Precision	0.7828	**0.8654**	0.7643	0.8296	0.8009
Recall	0.7379	0.8874	0.8381	**0.9038**	0.8978
F1-Score	0.7597	**0.8763**	0.7995	0.8428	0.8466
Matthew CC	0.7214	0.7441	0.7712	**0.8173**	0.8215
Specificity	0.9659	0.8560	**0.9640**	0.9601	0.9633
Sensitivity	0.7379	0.8874	0.8381	**0.9038**	0.8978
False Positive Rate	**0.0341**	0.1440	0.0360	0.0399	0.0367

Observations:

- Source synonymy is responsible for achieving high precision and overall F-1 score.
- Adding hierarchical context trades precision for higher recall.
- Adding source synonymy, hierarchical context, and semantic group give an overall boost to accuracy and recall.
- However, adding source synonymy of hierarchical context did not yield any noticeable improvement.

Examples of True Positives and True Negatives Correctly Identified

True Positives (Synonyms) Correctly Identified	
nail clipper	cutters nail
injury of salivary gland	salivary gland injury
avulsion	fracture sprain
True Negatives (Non-synonyms) Correctly Identified	
fingernail	infection of fingernail
product containing only **iron** medicinal product	product containing only **levorphanol** medicinal product
medical and surgical gastrointestinal system insertion **ileum** via natural or artificial opening endoscopic **infusion** device	medical and surgical gastrointestinal system revision **stomach** via natural or artificial opening endoscopic **other** device

Examples of False Positives Identified and False Negatives Not Identified

False Positives (Non-synonyms) Identified	
finding of **wrist** joint	finding of **knee** joint
malignant neoplasm of upper limb	malignant neoplasm of **muscle of** upper limb
skin **wound** of axillary fold	skin **cyst** of axillary fold
False Negatives (Synonyms) Not Identified	
hla antigen	human leukocyte antigen
pyelotomy	incision of renal pelvis treatment
routine cervical smear	screening for malignant neoplasm of cervix

Conclusion & Future Work

- Deep learning approach provides good performance in identifying synonymy among atoms.
- Adding source synonymy yields better precision and overall F-1 score.
- Adding hierarchical context trades precision for higher recall.
- Adding source synonymy, hierarchical context, and semantic group give an overall boost to accuracy and recall.

- **Limitation:** This approach does not address the *inter-concept* and *semantic type categorizations* *(other components in the UMLS Metathesaurus)*.

- **Future work:** How can the models be used in conjunction to complement the current lexical processing and human editors.