Categorizing the Relationships between Structurally Congruent Concepts from Pairs of Terminologies for Semantic Harmonization

Zhe He, PhD[1,2]

James Geller, PhD[2], Gai Elhanan, MD[3]

[1]*Department of Biomedical Informatics, Columbia University*

[2]*Department of Computer Science, New Jersey Institute of Technology*

[3]*Halfpenny Technologies, Inc*

2014 AMIA Summit on Translational Bioinformatics
April 8, 2014

Disclosure

- Zhe He discloses that he has no relationships with commercial interests.
- James Geller discloses that he has no relationships with commercial interests.
- Gai Elhanan discloses that he has no relationships with commercial interests.

Learning Objective

- Use structural method to find potential concepts for enriching the conceptual content of a biomedical terminology

Overview

- Motivation
  - Exploring structural method for semantic harmonization
- Background
  - Importance of the conceptual content of SNOMED CT
- Methods
  - Structural matching of pairs of terminologies in the UMLS
- Results
  - Reusable knowledge can be derived by structural matching, including discovery of possible synonym
- Limitations and Future Work
- Conclusions

Motivation

- Need of well-developed and well-maintained terminologies
- NLP tools that process clinical text need a terminology with fruitful concepts and synonyms.
- Complex clinical research texts require combined use of multiple terminologies (Weng *et al.* 2010)
- Terminologies need harmonization to achieve semantic interoperability (Bittner *et al.* 2005)

Semantic Harmonization Between Different Terminologies

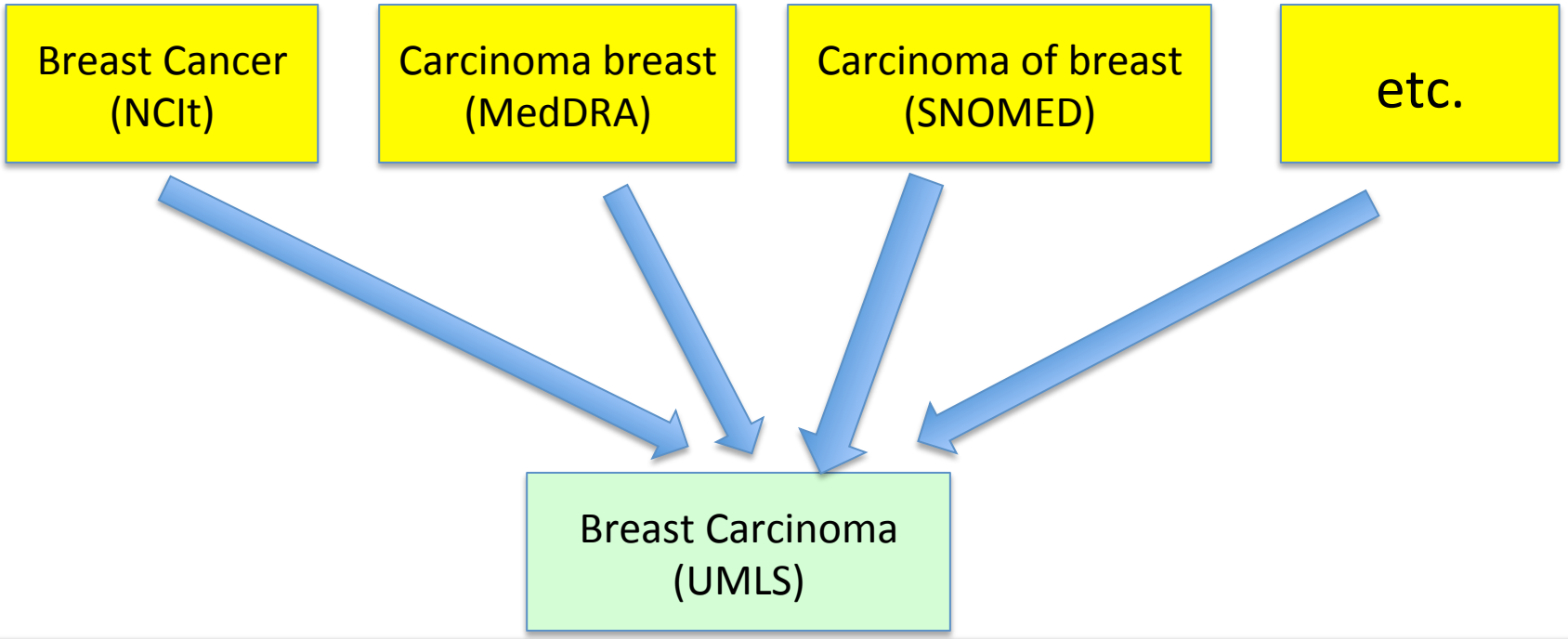
-
  (Weng *et al.* 2010)
- Harmonized existing time ontologies for annotating temporal relation in clinical narratives (Tao *et al.* 2011)
- Semantic harmonization efforts have recently been extended for various terminologies
  - SNOMED CT and LOINC (AMIA 2013 Informatics Year in Review)
  - SNOMED CT and ICD 11 (Rodrigues *et al.* 2013)

Importance of Conceptual Content of SNOMED CT

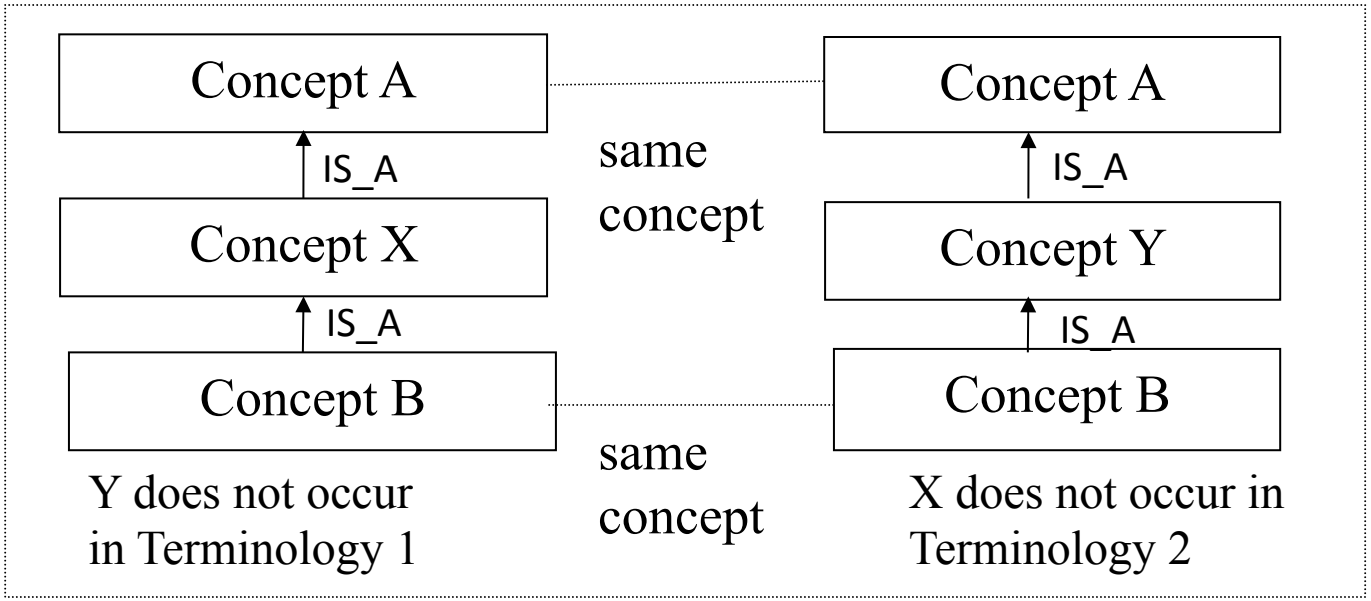
- SNOMED CT is going to be a major terminology for EHR encoding of diagnoses and problem lists by 2015
- SNOMED CT has many problems!
- Top two mentioned deficiencies of SNOMED CT (Elhanan 2011):
  - Missing concepts – 23%
  - Missing synonyms – 17%
- Users will expect SNOMED CT to have correct synonyms and sufficient concepts to be used in EHR

Leveraging Common Structure of Pairs of Terminologies in the UMLS

- UMLS (Unified Medical Language System):
  - More than 170 source terminologies, 8.9 million terms, 2012AB release



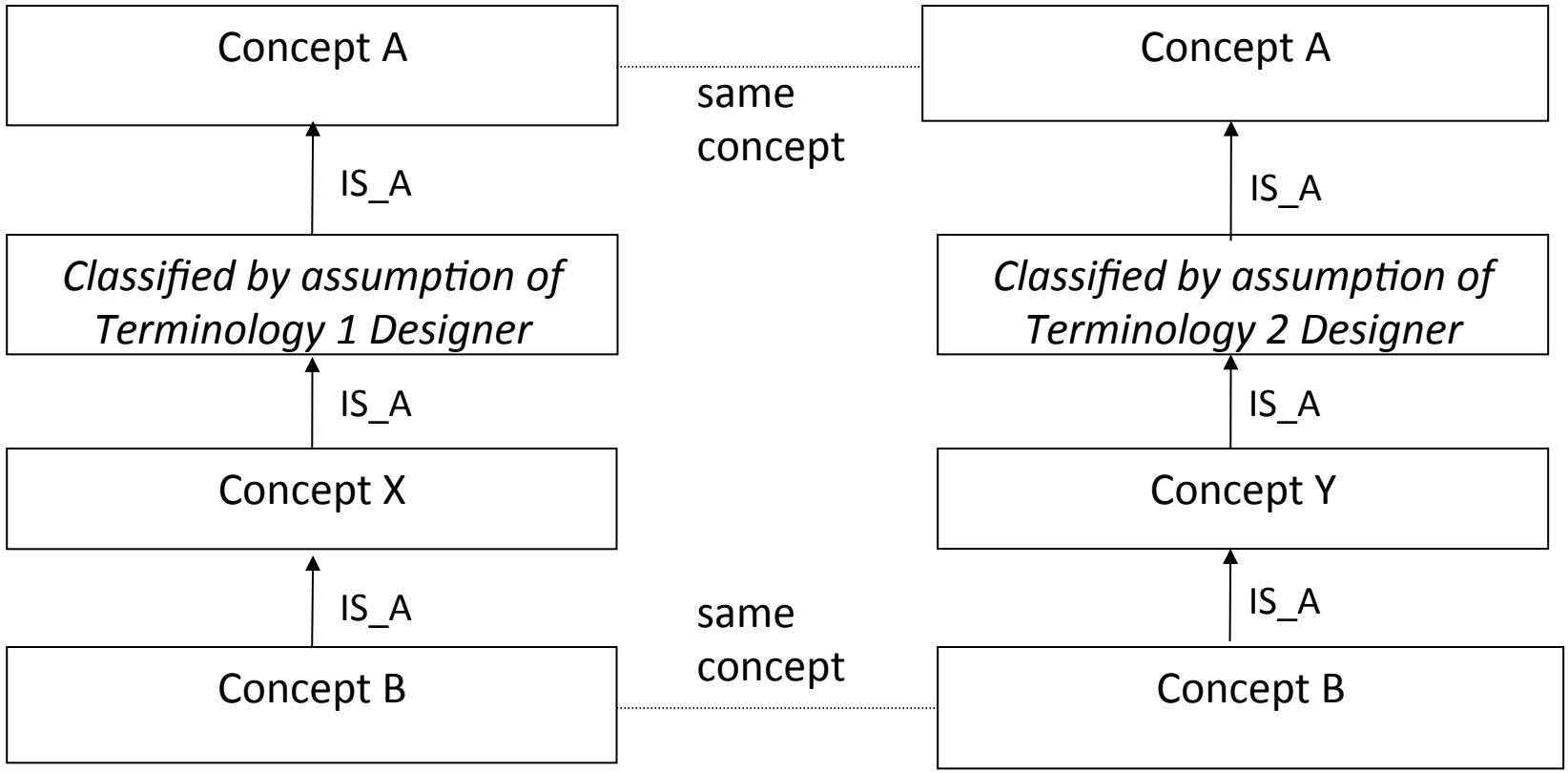
Structurally Congruent Concepts



Cycles were eliminated during the processing

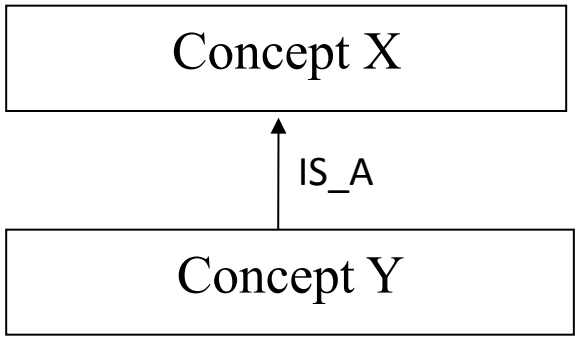
1) X and Y are alternative classifications
2) X can be a parent of Y
3) Y can be a parent of X
4) X and Y are synonymous
5) Structural errors in Terminology 1
6) Structural errors in Terminology 2

Alternative Classification

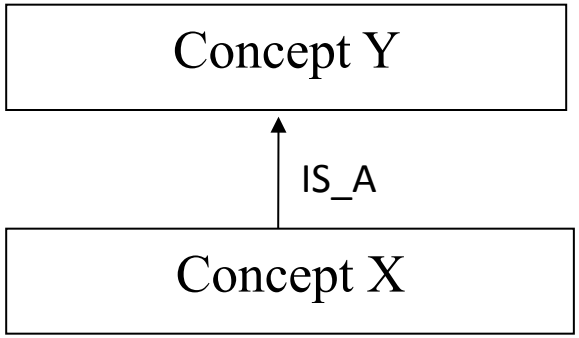


Parent-Child Relationship

X can be a parent of Y



Y can be a parent of X



Synonymous

X and Y are synonymous but have not been identified by the UMLS

Concept X (Synonym: Y)

META Terminologies in this Study

- Metathesaurus terminologies with "PAR" relationship and "inverse_isa" relationship attribute were chosen
- SCTUSX and UWDA were excluded
- Reference Terminologies (Terminology 1):
  - MEDCIN (MEDCIN)
  - National Cancer Institute Thesaurus (NCIt)
  - Gene Ontology (GO)
  - Medical Entities Dictionary (CPM)
  - UMDNS: product category thesaurus (UMD)
  - Foundational Model of Anatomy Ontology (FMA)
- Terminology 2: SNOMED CT

Evaluation: Pairs of Congruent Concepts Found

Reference Terminology	Size of Terminology	# of Pairs of Congruent Concepts	Sample Size
MEDCIN	279529	655	70
NCI	95523	582	70
FMA	82062	116	70
UMD	15956	18	18
GO	61925	6	6
CPM	3078	7	7
Total	--	1384	**241**

Review Results for Pairs of Congruent Concepts

Co-author GE reviewed the sample

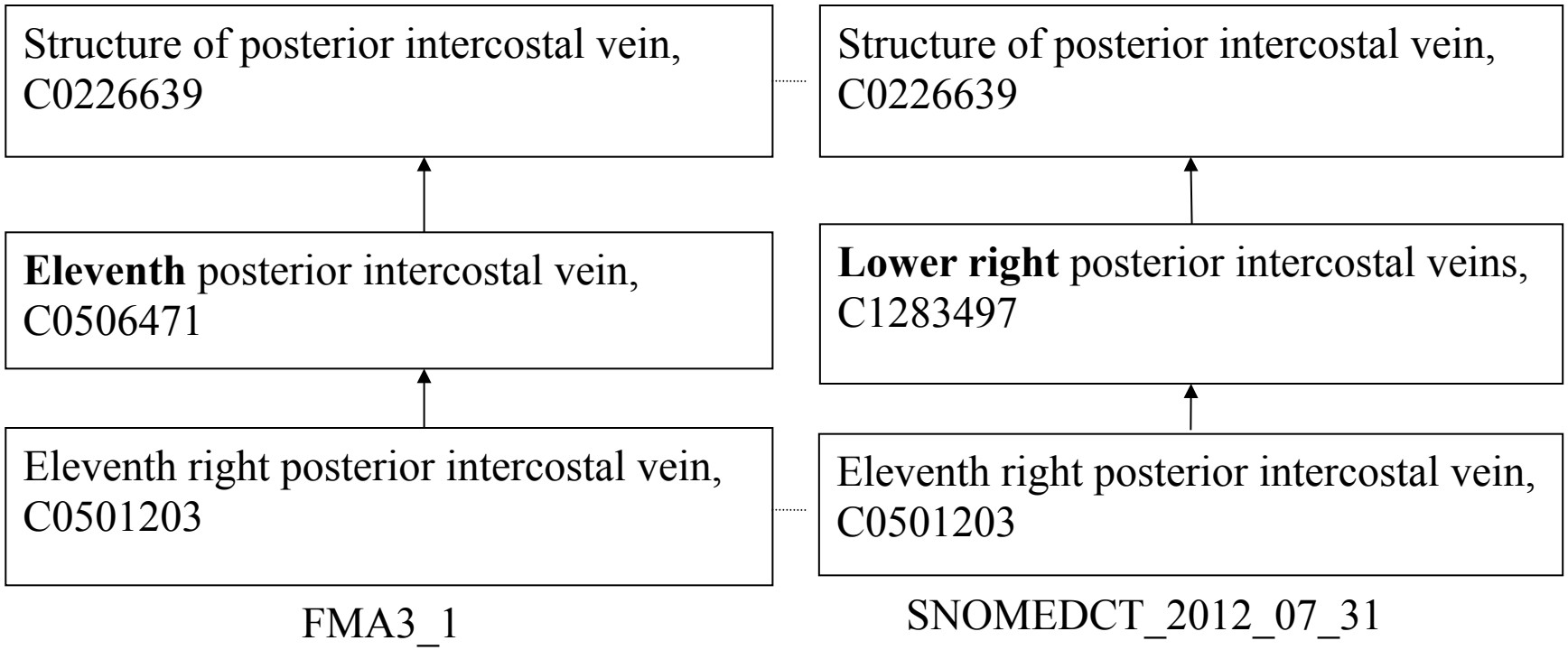
Reference Terminology	Sample Size	Alternative Classification	Y → X	X → Y
MEDCIN	70	44	10	7
NCI	70	38	12	6
GO	6	2	--	4
CPM	7	5	--	--
UMD	18	9	1	--
FMA	70	45	13	4
Total	241	143	36	21
Percentage	100%	**59.3%**	**14.9%**	**8.7%**

23.6%

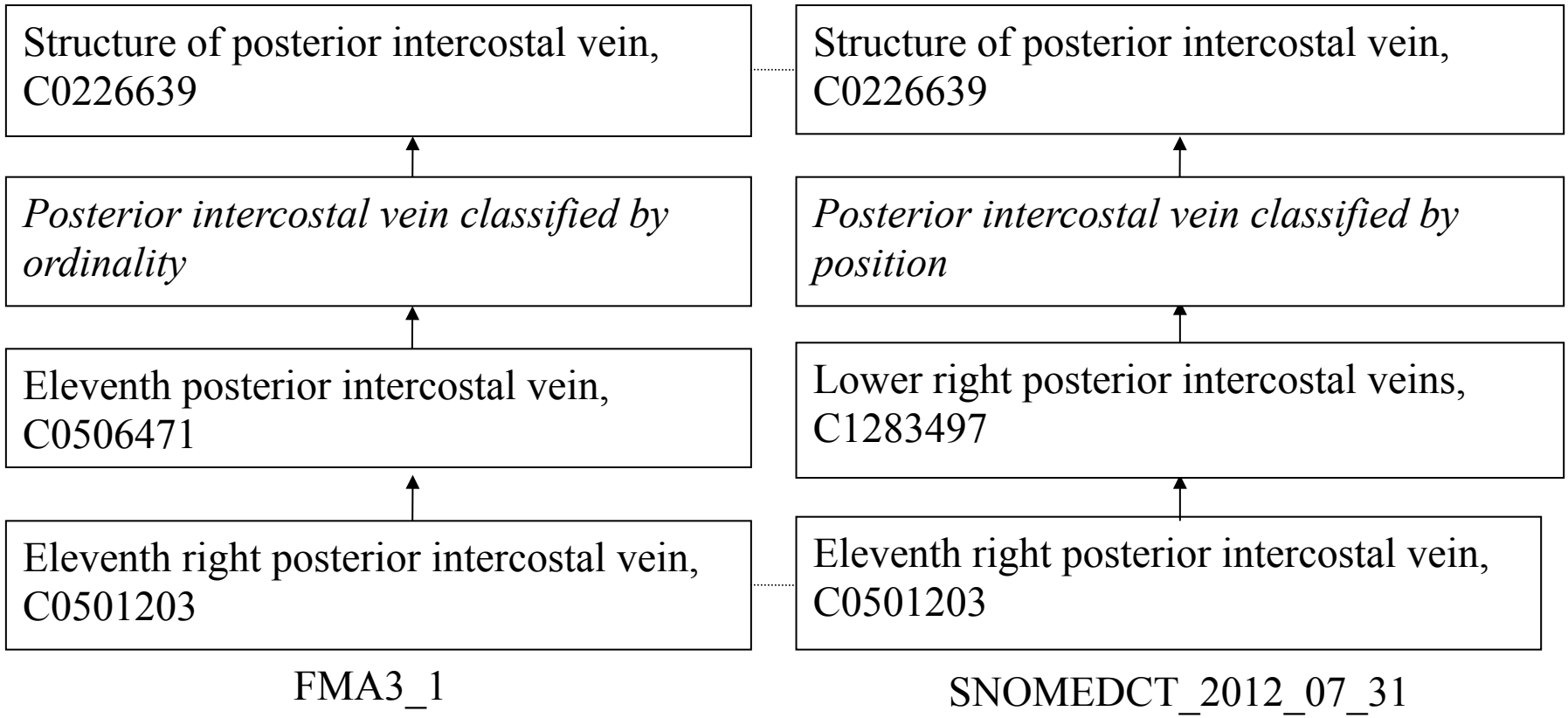
Review Results for Pairs of Congruent Concepts

Reference Terminology	Sample Size	Error in Termi-nology 1	Error in Termi-nology 2	Synonym
MEDCIN	70	--	1	8
NCI	70	--	3	11
GO	6	--	--	--
CPM	7	--	--	2
UMD	18	--	--	8
FMA	70	2	--	6
Total	181	2	4	35
Percentage	100%	0.8%	1.7%	**14.5%**

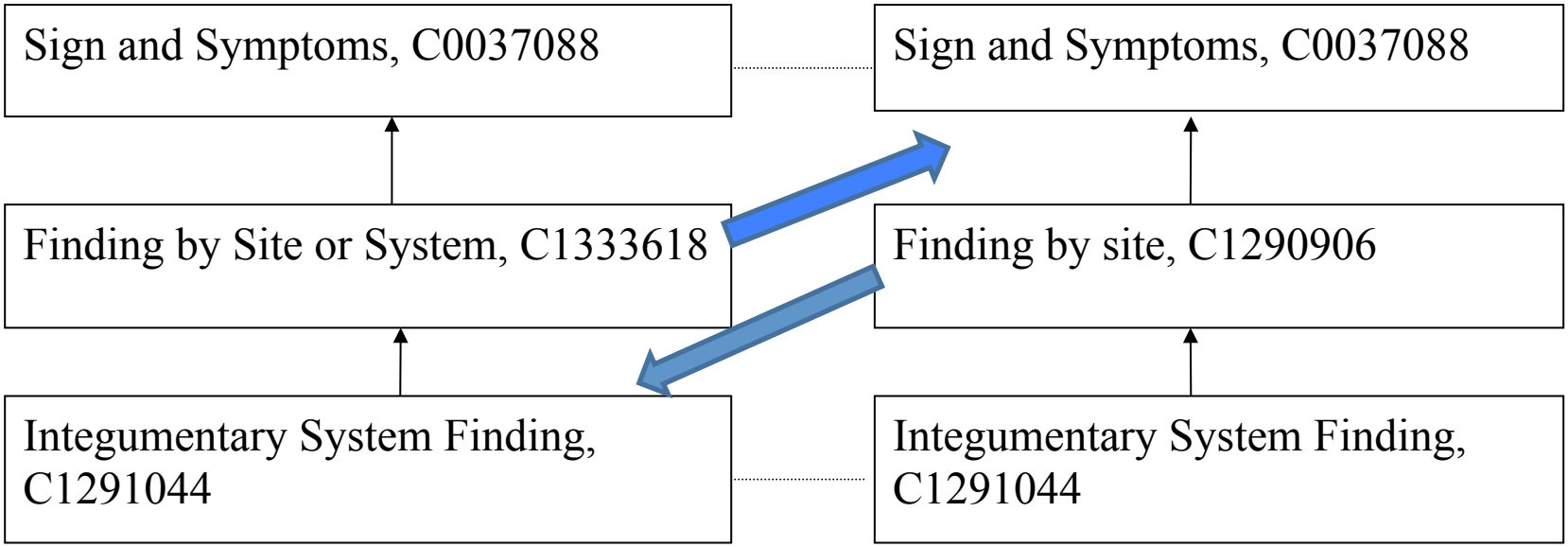
Example: Alternative Classification



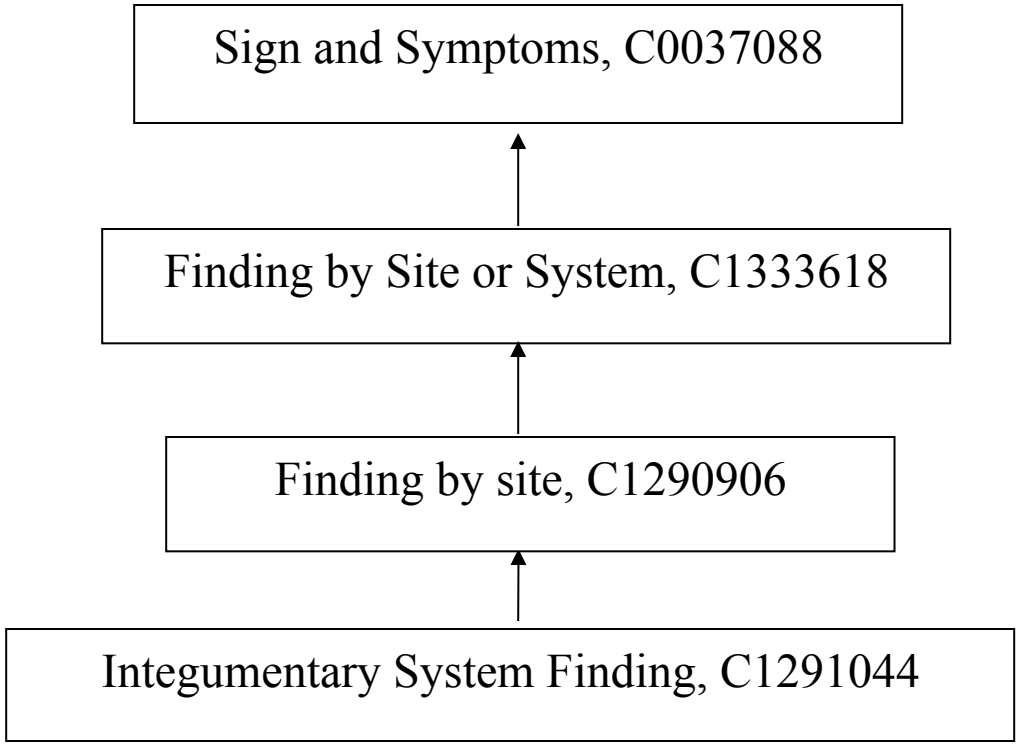
Making Explicit an Implicit Assumption of the Two Original Terminology Designers



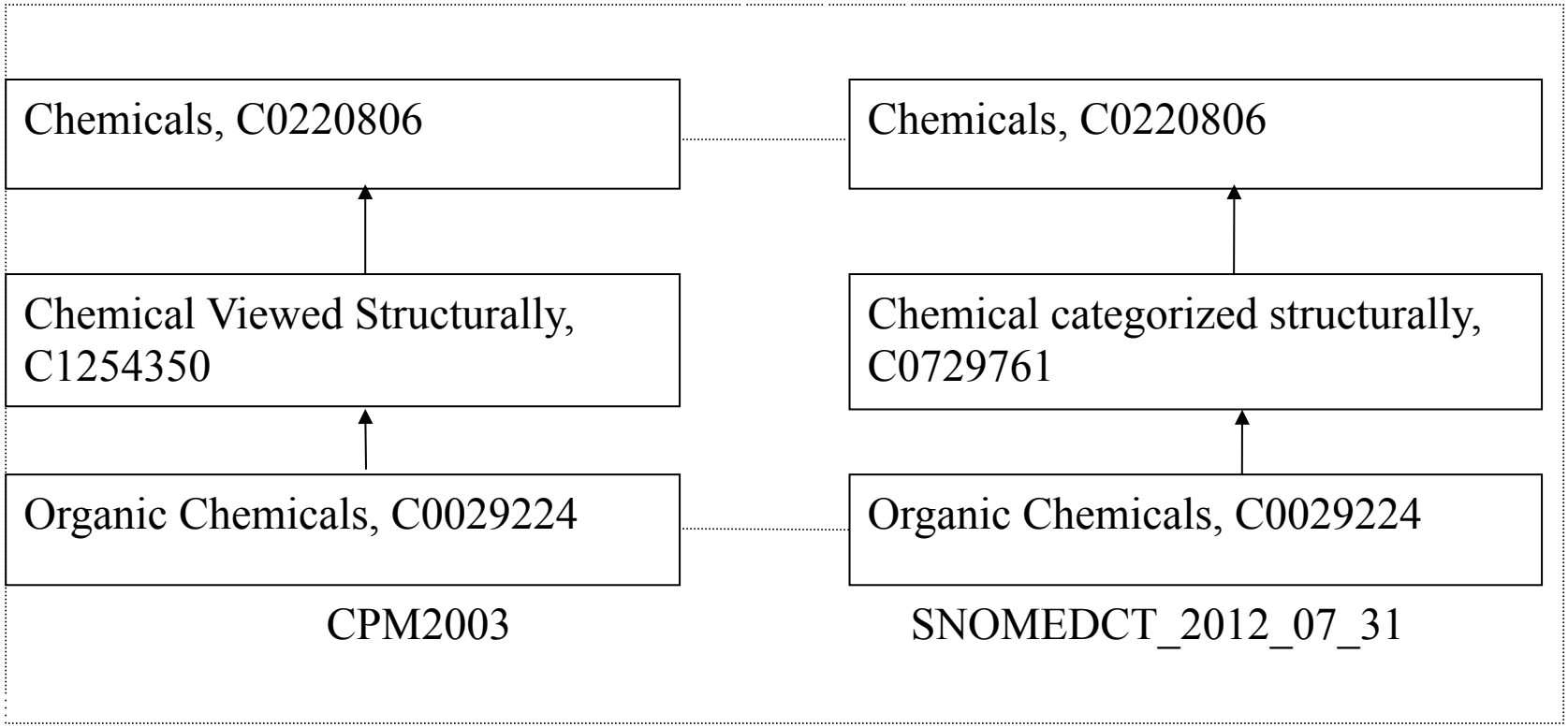
Example: Parent of the Other



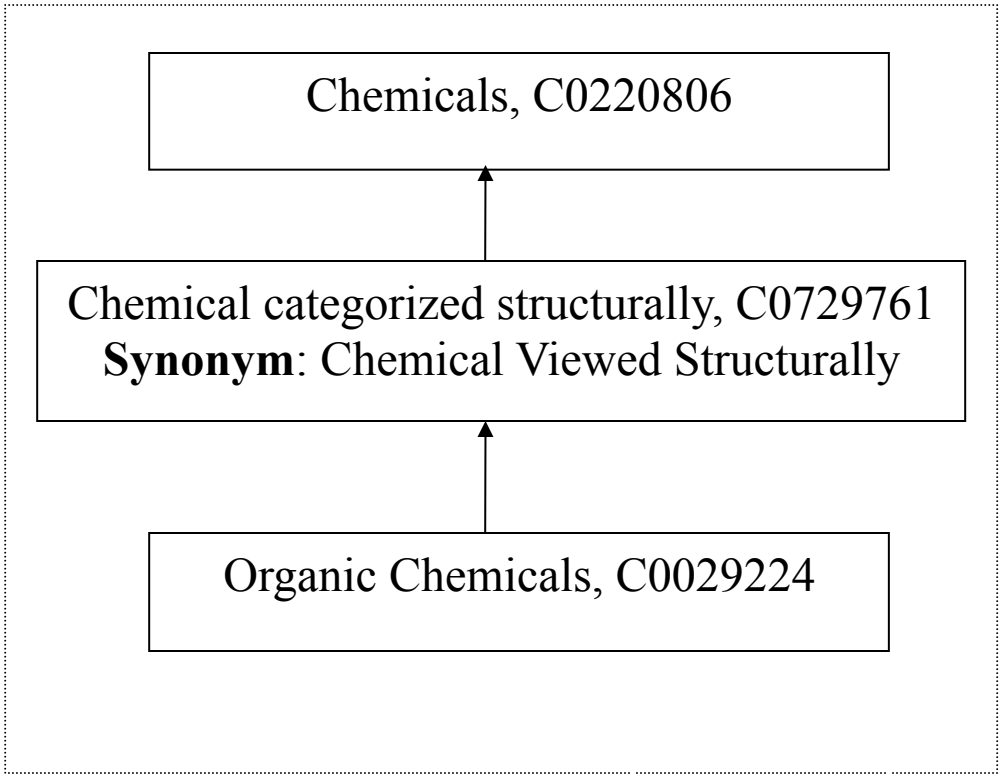
Suggestion: Concept Import



Example: Synonym of the Other



Suggestion: Merge Two Synonymous Concepts



Limitations

- Harmonization cannot be done without the consent of terminology curators
- All the terminologies are in UMLS Rich Release format

Future Work

- More complex configurations: more intermediate concepts
- Algorithm to identify the relationships between intermediate concepts in complex configurations
- Pairs of any two terminologies

Conclusions

- Six reference terminologies of the UMLS vs. SNOMED
- In a sample of 241 congruency pairs
  - 143 out of 241 (59.3%) concept pairs: alternative classification.
  - 47 out of 241 (23.6%) concept pairs: parent-child relationships.
  - 35 (14.5%) new synonyms
  - Three pairs of concepts indicated errors
- Take home message:
  - A semi-automated way based on common structure of the UMLS may complement existing human-expert centered methods to find potential concepts for import and export to a terminology.

Acknowledgment

We would like to thank:

Dr. Yehoshua Perl and Dr. Chunhua Weng

for sharing their insights and giving feedback for this work

References (1)

Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform. 2010;43(3):451-67.

Bittner T, Donnelly M, S. W. Ontology and semantic interoperability. In: Prosperi D, Zlatanova S, editors. Large-scale 3D data integration: Problems and challenges: CRCpress (Tailor & Francis); 2005: 139-60.

Tao C, Solbrig HR, Chute CG. CNTRO 2.0: A Harmonized Semantic Web Ontology for Temporal Relation Inferencing in Clinical Narratives. AMIA Summits Transl Sci Proc. 2011;2011:64-8.

Rodrigues JM, Schulz S, Rector A, Spackman KA, Üstün B, Chute CG, Della Mea V, Millar J, Persson KB. Sharing Ontology between ICD 11 and SNOMED CT will enable Seamless Re-use and Semantic Interoperability. Stud Health Technol Inform. 2013;192:343-6

Elhanan G, Perl Y, Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. J Am Med Inform Assoc. 2011; Suppl 1:i33-44

References (2)

Kumar A, Smith B, Novotny DD. Biomedical informatics and granularity. Comp Funct Genomics. 2004;5(6-7):501-8.

Weng C, Gennari JH, Fridsma DB, User-centered Semantic Harmonization: A Case Study, J Biomed Inform, 2007, 40(3):353-64

Schulz S, Boeker M, Stenzhorn H. How Granularity Issues Concern Biomedical Ontology Integration. In Proceedings of the International Congress of the European Federation for Medical Informatics (MIE 2008). Gothenburg, Sweden; 2008: 863-68.

Rector A, Rogers J, Bittner T. Granularity, scale and collectivity: when size does and does not matter. J Biomed Inform. 2006 Jun;39(3):333-49.

Mougin F, Bodenreider O. Approaches to eliminating cycles in the UMLS Metathesaurus: naive vs. formal. AMIA Annu Symp Proc. 2005:550-4.

Thanks!

Zhe He, PhD
zh2132@columbia.edu