

Biomedical Informatics Center  
George Washington University  
October 7, 2019



# Using lexical and structural features for quality assurance of biomedical ontologies

## *Application to SNOMED CT*



*Olivier Bodenreider, MD, PhD*

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



U.S. National Library of Medicine



# Outline

- ◆ Motivation
- ◆ SNOMED CT
- ◆ Terminology QA approaches
  - Structural
  - Lexical
  - Hybrid (structural + lexical)



# Motivation

## Research and applications

# Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications

Alan L Rector,<sup>1</sup> Sam Brandt,<sup>2</sup> Thomas Schneider<sup>1</sup>

► Additional appendices are published online only. To view these files please visit the journal online ([www.jamia.org](http://www.jamia.org)).

<sup>1</sup>School of Computer Science, University of Manchester, Manchester, UK

<sup>2</sup>Siemens Health Services, Malvern, Pennsylvania, USA

## Correspondence to

Alan L Rector, School of Computer Science, University of Manchester, Manchester M13 9PL, UK;  
[rector@cs.manchester.ac.uk](mailto:rector@cs.manchester.ac.uk)

Received 13 December 2010  
Accepted 30 December 2010  
Published Online First  
21 April 2011

## ABSTRACT

**Objectives** (a) To determine the extent and range of errors and issues in the Systematised Nomenclature of Medicine — Clinical Terms (SNOMED CT) hierarchies as they affect two practical projects. (b) To determine the origin of issues raised and propose methods to address them.

**Methods** The hierarchies for concepts in the Core Problem List Subset published by the Unified Medical Language System were examined for their appropriateness in two applications. Anomalies were traced to their source to determine whether they were simple local errors, systematic inferences propagated by SNOMED's classification process, or the result of problems with SNOMED's schemas. Conclusions were confirmed by showing that altering the root cause and reclassifying had the intended effects, and not others.

**Main results** Major problems were encountered, involving concepts central to medicine including myocardial

codes. When doctors apply SNOMED codes to a patient, they are stating that those codes and all their ancestors in the hierarchy apply to that patient. When researchers use codes in queries, they are querying for those codes and all of their descendants. When software interprets postcoordinated expressions, it depends on the hierarchies to give those expressions their correct meaning.

This paper reports attempts to use the SNOMED hierarchies in two practical applications:

- as a contributor to the 'ontological component' of the eleventh revision of the International Classification of Diseases (ICD-11);
- as part of the documentation tools for a commercial clinical information system.

By contrast with most previous studies, we are concerned here only with inferences that are incorrect or misleading clinically. We are not concerned with other types of SNOMED anomalies



# Motivation

- ◆ Biomedical terminologies and ontologies are enabling resources for clinical decision support systems and data integration systems for translational research and health analytics
- ◆ Their quality has a direct impact on healthcare and biomedical research
- ◆ Quality assurance (QA) of biomedical terminologies remains an active field of research



# SNOMED Clinical Terms

**SNOMED**  
International

Leading healthcare  
terminology, worldwide

# SNOMED CT Characteristics

- ◆ Developed by SNOMED International
  - Consortium of over 40 member countries
- ◆ Largest clinical terminology in the world
  - ~350,000 active concepts
  - ~1 million terms (“descriptions”)
- ◆ Major organizing principles
  - Logical definitions (incomplete: many primitives)
  - Built using description logics ( $\mathcal{EL}^{++}$ )



# SNOMED CT Example

## Parents

- ▶ ☰ Operation on appendix (procedure)
- ▶ ☰ Partial excision of large intestine (procedure)

## ☰ Appendectomy (procedure) ☆ ↗

SCTID: 80146002

80146002 | Appendectomy (procedure) |

- Appendectomy
- Excision of appendix
- Appendicectomy
- Appendectomy (procedure)

Procedure site - Direct → Appendix structure  
Method → Excision - action

## Children (8)

- ☰ Appendectomy with drainage (procedure)
- ▶ ☰ Emergency appendectomy (procedure)
- ● Excision of appendiceal stump (procedure)
- ● Excision of ruptured appendix by open approach (procedure)
- ● Incidental appendectomy (procedure)
- ● Interval appendectomy (procedure)
- ▶ ☰ Laparoscopic appendectomy (procedure)
- ☰ Non-emergency appendectomy (procedure)



# SNOMED CT Challenges

## ◆ Legacy

- Many primitive concepts
- Not amenable to automatic DL classification

## ◆ Maintenance

- Human editors
- Error prone

## ◆ Quality assurance

- Difficult due to its size
- Ontology templates
  - Difficult to apply retrospectively



# Quality assurance approaches

# Quality assurance approaches

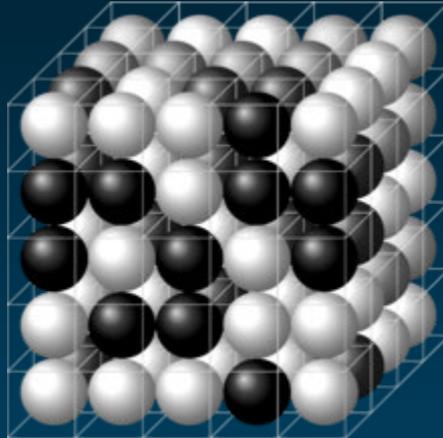
- ◆ Three types of QA approaches applied to SNOMED CT by researchers
  - Lexical
    - Based on the properties of terms, such as compositionality
  - Structural
    - Based on the organizational structure of concepts
  - Semantic
    - Rely on the logical definitions of concepts in description logic-based terminologies
- ◆ Hybrid approaches (structural + lexical)



# Quality assurance approaches

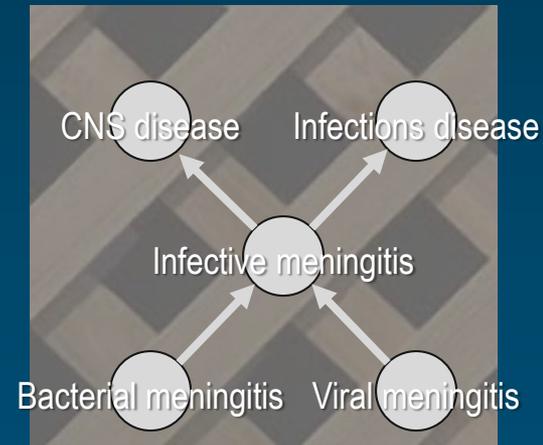
*Structural approaches*

# Lattices



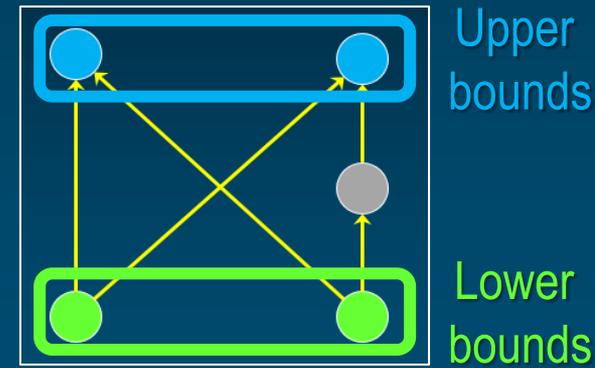
## ◆ Lattice

- Specific type of directed acyclic graph (DAG)
- Any two nodes have a unique maximal common descendant, as well as a unique minimal common ancestor

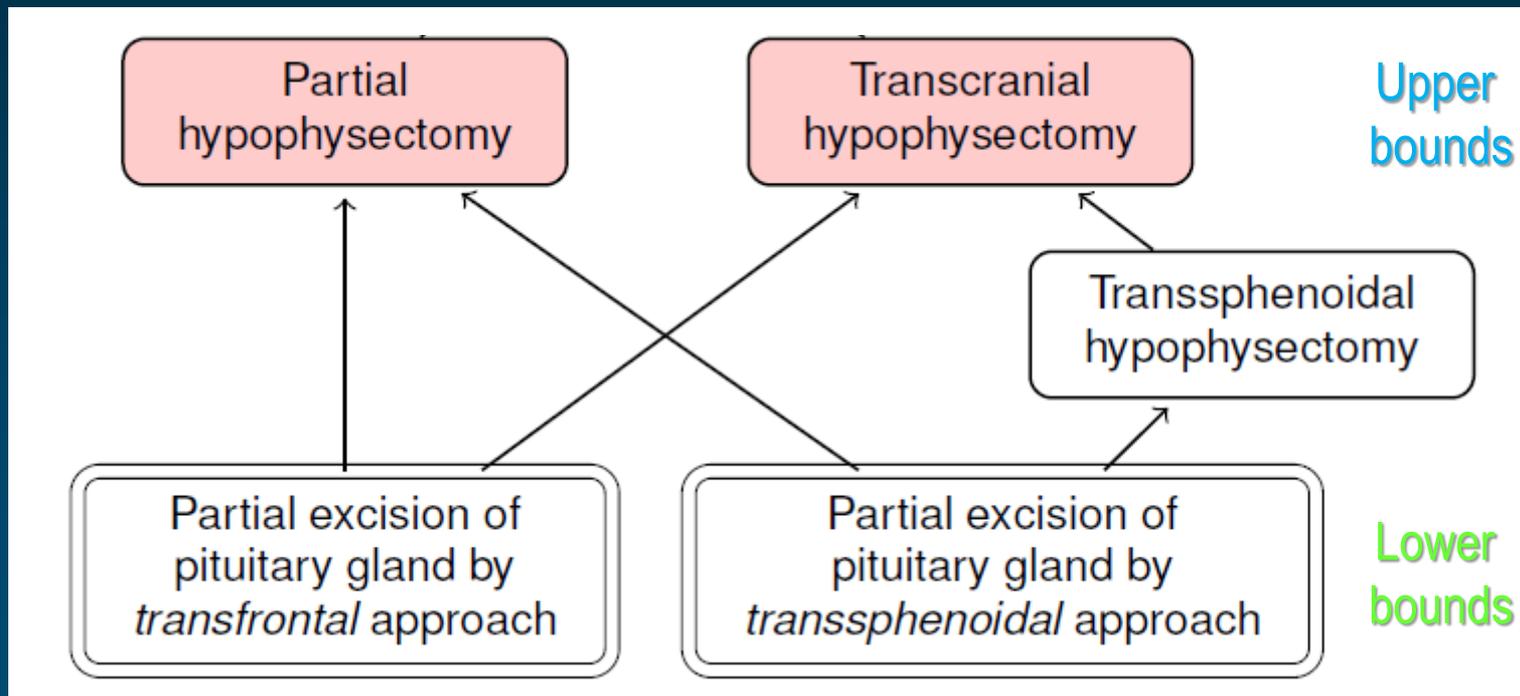


# Lattice-based structural QA

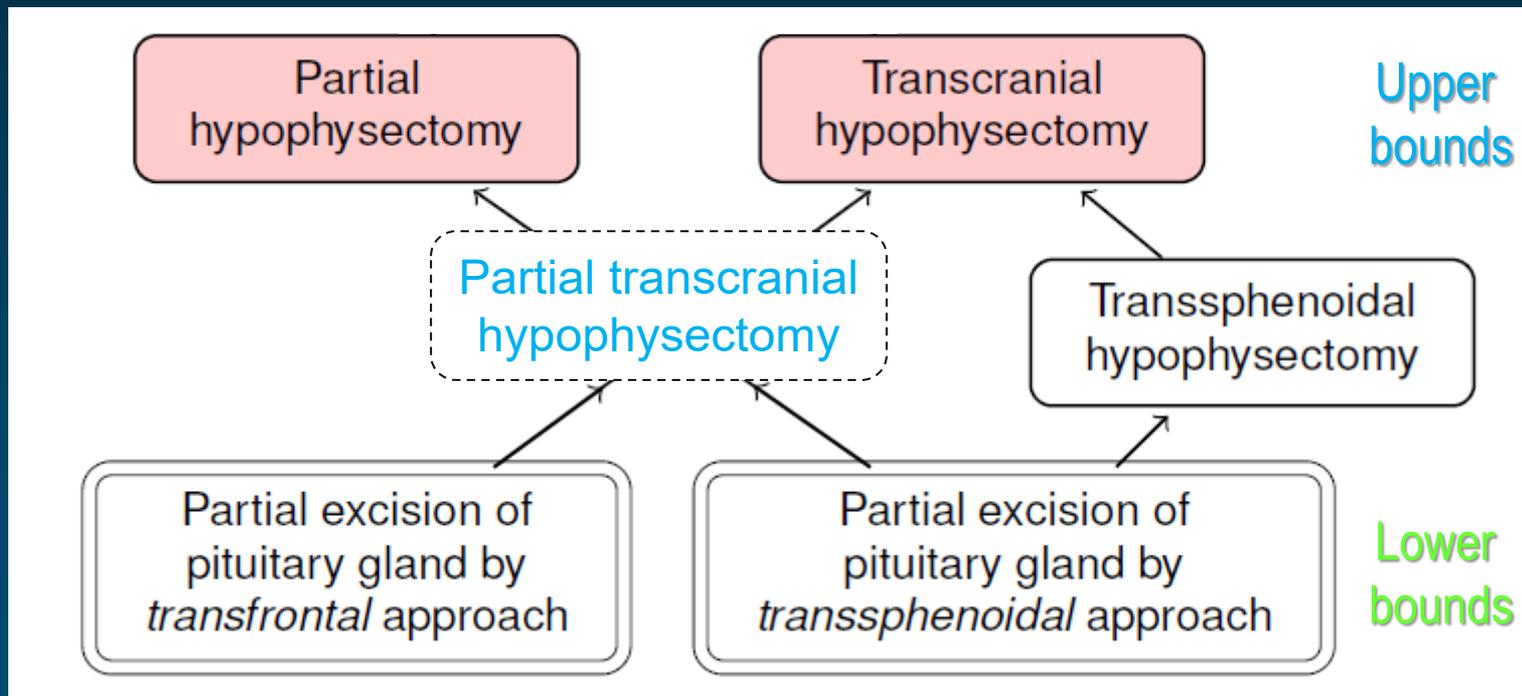
- ◆ Non-lattice (approximation)
  - When two concepts have more than one ancestor in common
- ◆ Non-lattice subgraphs are often indicative of a problem in ontology construction
  - Missing hierarchical relation
  - Missing intermediary concept



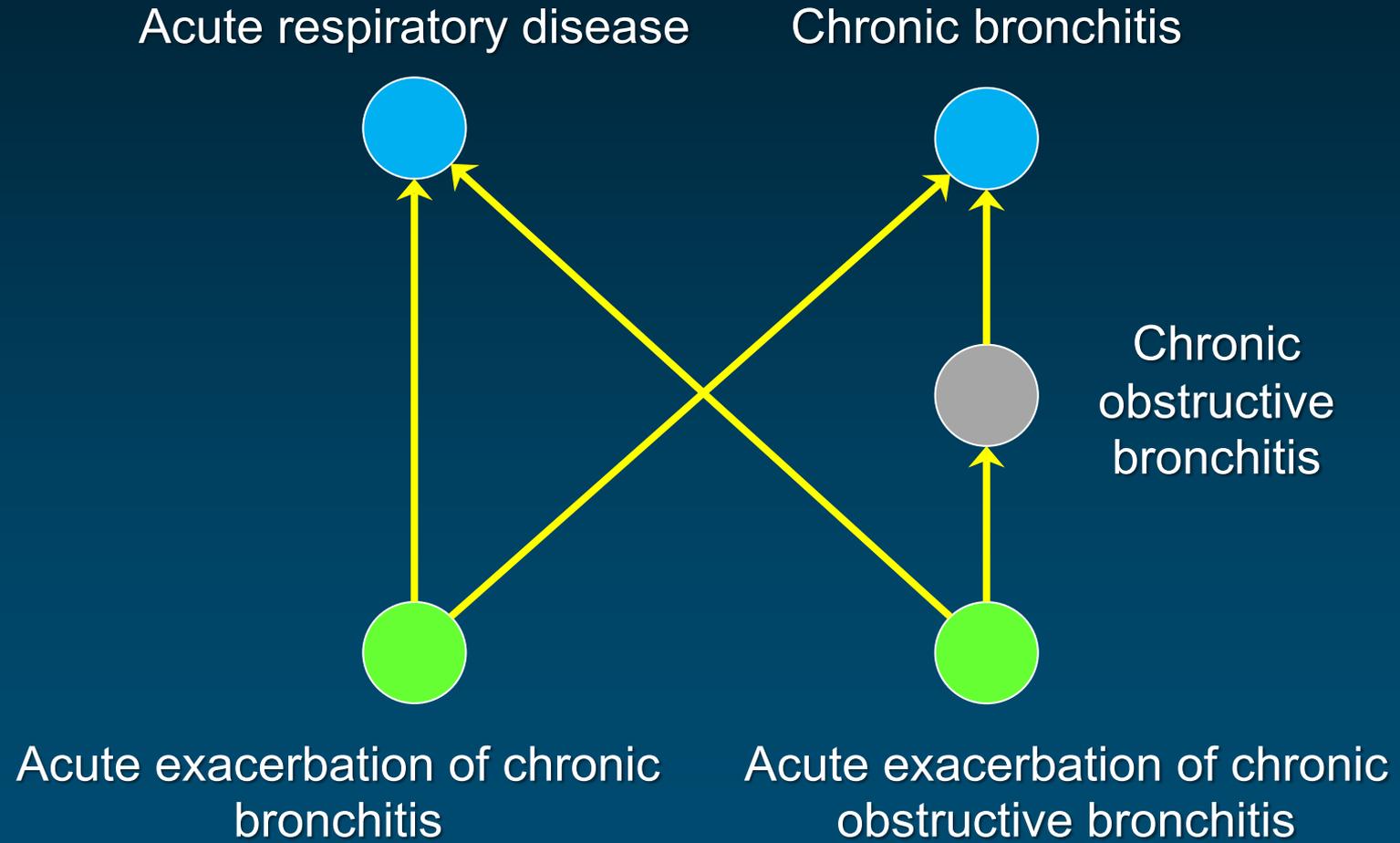
# Example of non-lattice subgraph



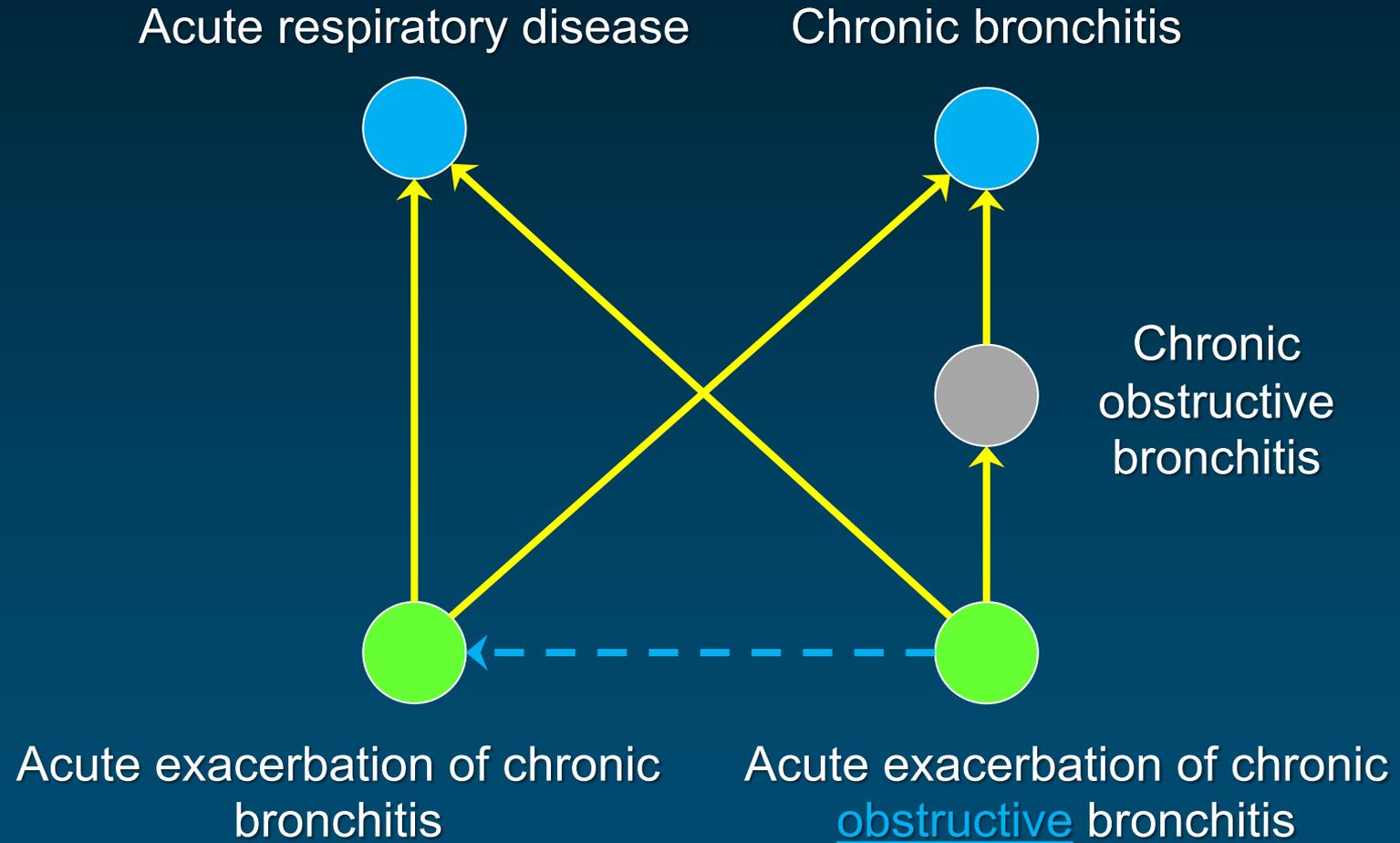
# Missing intermediary concept



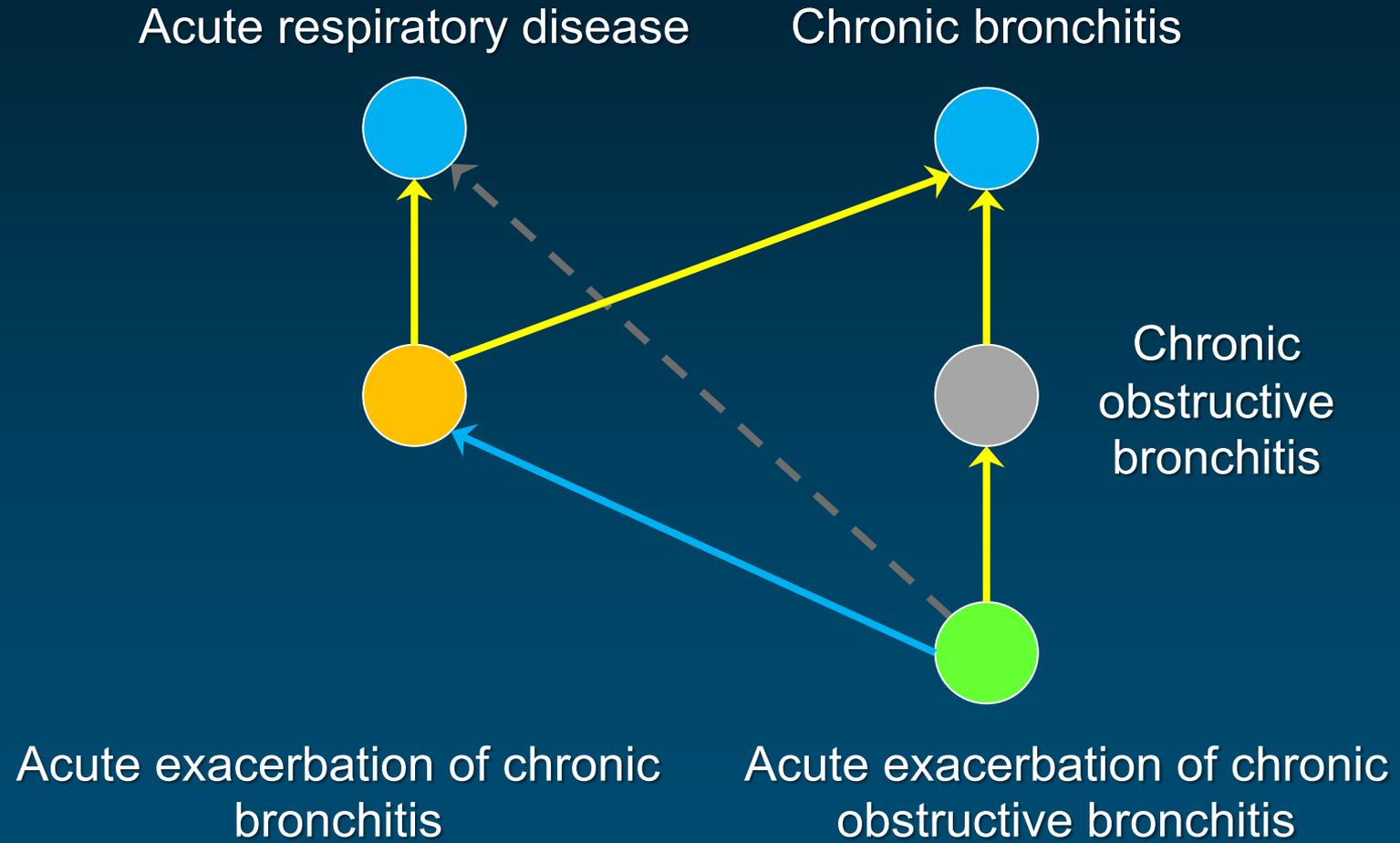
# Non-lattice subgraph in SNOMED CT



# Missing hierarchical relation



# ~~Non-lattice~~ subgraph in SNOMED CT



# Limitations of the structural approach

## ◆ Technically

- Computationally intensive (initially)

## ◆ Practically

- Limited precision
  - Not all non-lattice subgraphs are indicative of an error
- Editorial guidelines in SNOMED CT
  - Avoid systematic pre-coordination
- Trade-off between
  - “Purity” of lattice representation
  - Parsimony

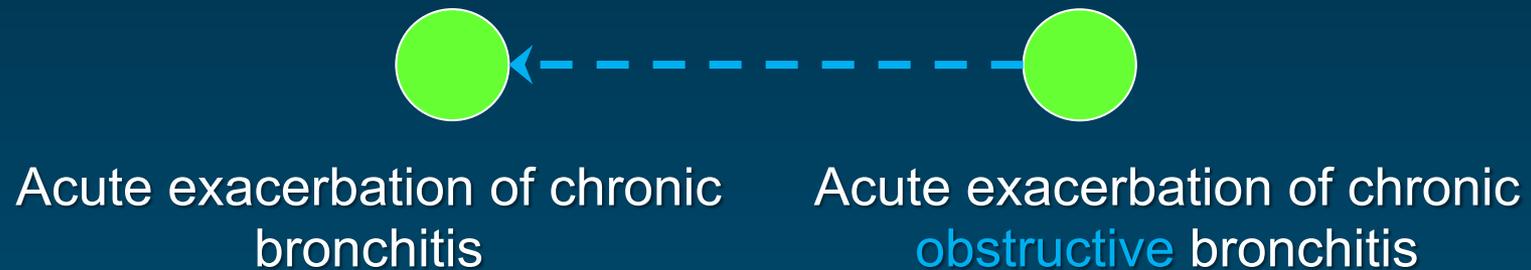


# Quality assurance approaches

*Lexical approaches*

# QA based on lexical patterns

- ◆ Lexical differences among terms are often indicative of semantic relations among them
- ◆ Term compositionality



- ◆ Simple implementation through bags of words

# Suggested missing hierarchical relations

Child name	Parent name
Alveolar bone graft <u>to mandible</u>	Alveolar bone graft
<u>Basal cell carcinoma of skin</u> of lip	Carcinoma of lip
Carcinoma <u>in situ of</u> palate	Palate carcinoma
Chronic <u>bacterial</u> otitis externa	Chronic otitis externa
<u>Congenital</u> vascular anomaly of eyelid	Vascular anomaly of eyelid
<u>Electrocoagulation</u> of retina <u>for</u> repair <u>of</u> tear	Repair of retina
<u>Hallucinogen intoxication delirium</u>	Hallucinogen intoxication
Infection of <u>preauricular sinus</u>	Preauricular sinus
<u>Pituitary stalk compression</u> hyperprolactinemia	Pituitary stalk compression
<u>Suture of tongue to lip</u> for micrognathia	Suture of lip



# Limitations of the lexical approach

- ◆ Limited precision
  - Many false positives when using a simple bag-of-words approach
- ◆ Limited recall
  - Many false negatives when using only the preferred terms
- ◆ Possible mitigation strategies
  - Add lexico-syntactic constraints to increase precision
  - Also use synonyms to increase recall
  - Overall: gain in recall does not compensate loss in precision



# Quality assurance approaches

*Combining structural and lexical approaches*

---

## Research and Applications

# Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT

Licong Cui,<sup>1,2</sup> Wei Zhu,<sup>2</sup> Shiqiang Tao,<sup>2,3</sup> James T Case,<sup>4</sup> Olivier Bodenreider,<sup>4</sup> and Guo-Qiang Zhang<sup>2,3</sup>

<sup>1</sup>Department of Computer Science, University of Kentucky, Lexington, KY, USA, <sup>2</sup>Institute for Biomedical Informatics, University of Kentucky,

<sup>3</sup>Division of Biomedical Informatics, College of Medicine, University of Kentucky and <sup>4</sup>National Library of Medicine, Bethesda, MD, USA

Corresponding Author: Licong Cui, Guo-Qiang Zhang, 301 Rose Street, 233 James F. Hardyman Building, Lexington, KY, 40506, USA. E-mail: licong.cui@uky.edu, gq.zhang@uky.edu. Phone: 859-257-3062, Fax: 859-323-3740.

Received 27 June 2016; Revised 17 November 2016; Accepted 3 December 2016

## ABSTRACT

**Objective:** Quality assurance of large ontological systems such as SNOMED CT is an indispensable part of the terminology management lifecycle. We introduce a hybrid structural-lexical method for scalable and systematic discovery of missing hierarchical relations and concepts in SNOMED CT.

**Material and Methods:** All non-lattice subgraphs (the structural part) in SNOMED CT are exhaustively extracted

# Objectives

- ◆ To combine lexical and structural QA approaches to automatically and precisely identifying missing hierarchical relations and missing concepts in SNOMED CT
- ◆ To suggest remediation for such inconsistencies
- ◆ Materials: September 2015 version of SNOMED CT (U.S. edition)



# Overview of the methods

- ◆ Identifying non-lattice pairs and subgraphs
- ◆ Analyzing non-lattice subgraphs with lexical patterns
  - Containment
  - Intersection
  - Union
  - Intersection-Union
- ◆ Evaluation



# Identifying non-lattice pairs and subgraphs

- ◆ Hadoop-based technique
  - 30 hours to analyze all pairs of SNOMED CT concepts
- ◆ Aggregation of non-lattice pairs with the same shared ancestors into non-lattice subgraphs
  - Smaller subgraphs contained in larger subgraphs
- ◆ Select small non-lattice subgraphs (Size 4-6)
  - Cognitively manageable
- ◆ 171,011 non-lattice subgraphs
  - 70,250 small non-lattice subgraphs
    - 2046 exhibit one of the 4 lexical patterns



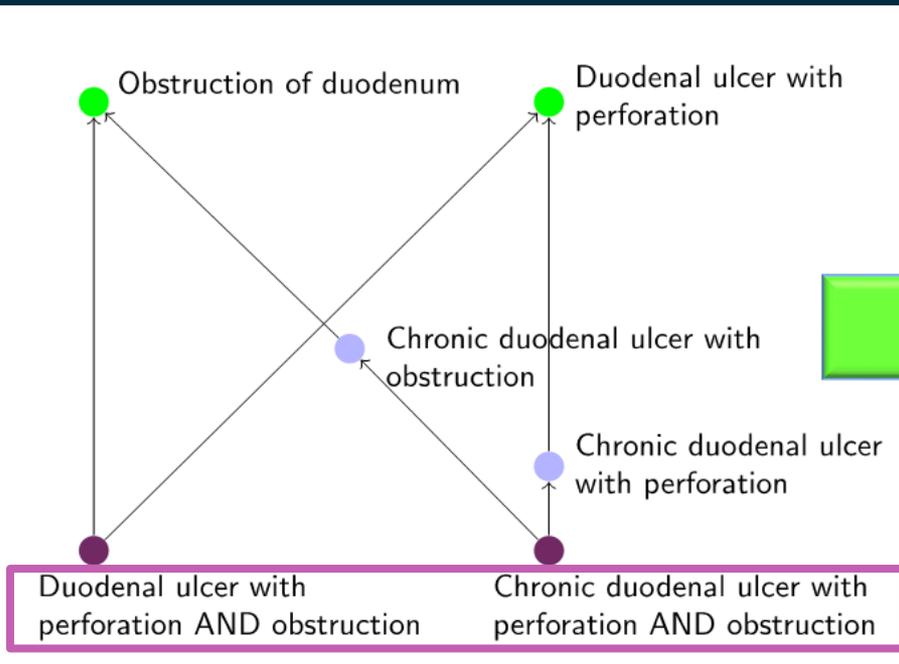
# Lexical patterns (1) Containment

- ◆ The set of words for one concept in the upper (resp. lower) bounds is contained in the set of words for another concept in the upper (resp. lower) bounds
- ◆ Suggests a *missing hierarchical relation* between concepts in the upper (resp. lower) bounds
- ◆ 736 small non-lattice subgraphs with this pattern

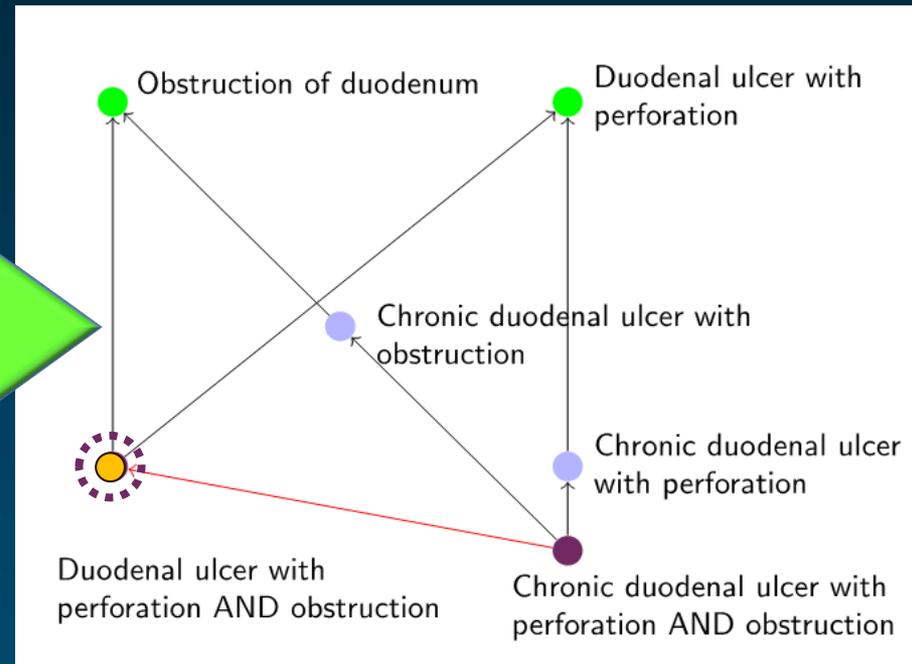


# Lexical patterns (1) Containment

Non-lattice subgraph



Suggested remediation



Duodenal ulcer with perforation AND obstruction

Chronic duodenal ulcer with perforation AND obstruction

# Lexical patterns (2) Intersection

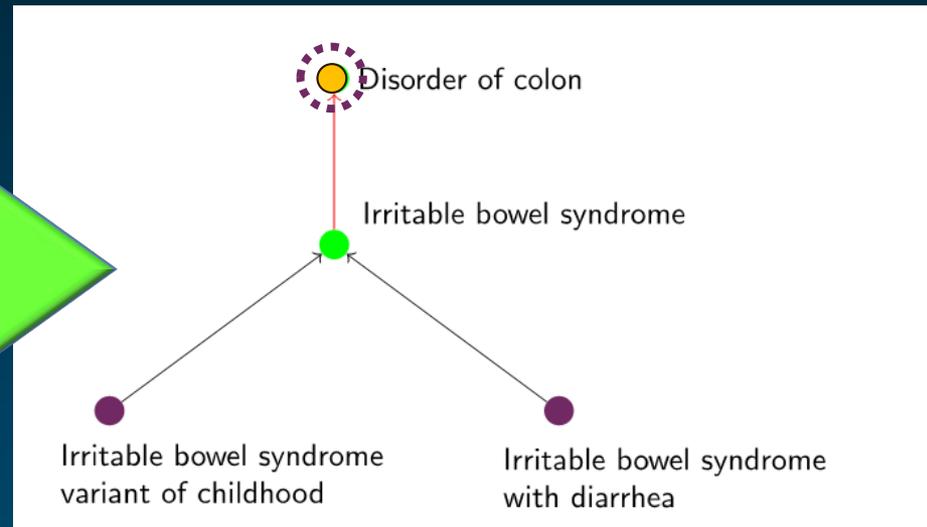
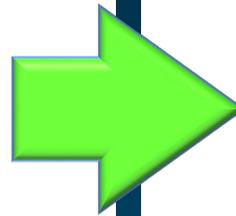
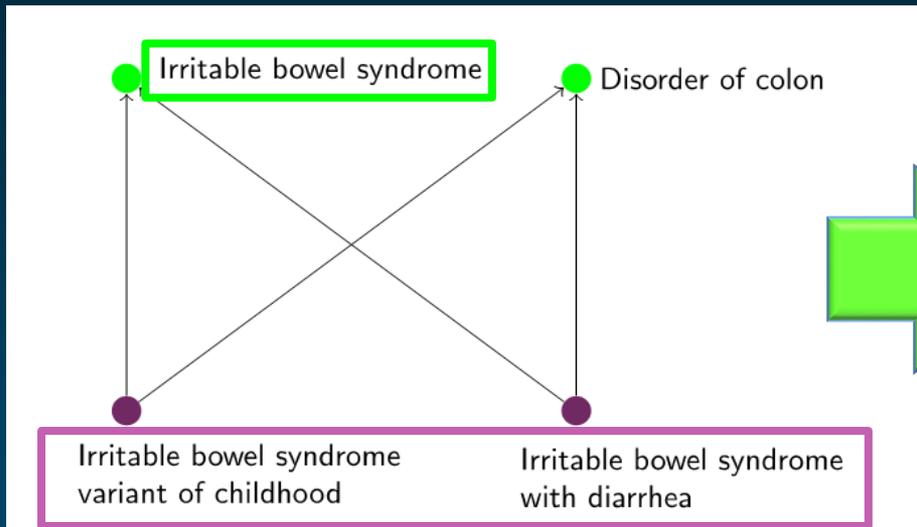
- ◆ The intersection of sets of words for concepts in the lower bounds is equal to the set of words for some concept in the upper bounds
- ◆ Suggests a *missing hierarchical relation* between concepts in the upper bounds
- ◆ 1085 small non-lattice subgraphs with this pattern



# Lexical patterns (2) Intersection

Non-lattice subgraph

Suggested remediation



Irritable bowel syndrome



Irritable bowel syndrome  
variant of childhood



Irritable bowel syndrome  
with diarrhea

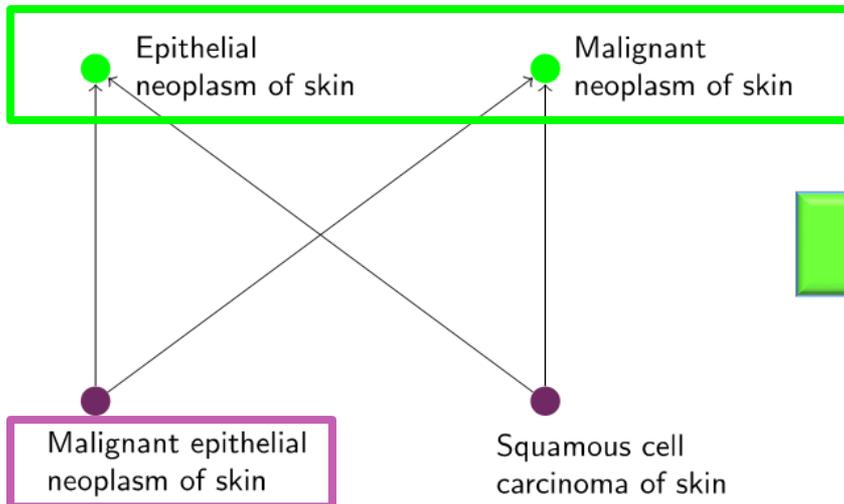
# Lexical patterns (3) Union

- ◆ The union of the sets of words for concepts in the upper bounds is equal to the set of words for some concept in the lower bounds
- ◆ Suggests a *missing hierarchical relation* between concepts in the lower bounds
- ◆ 164 small non-lattice subgraphs with this pattern

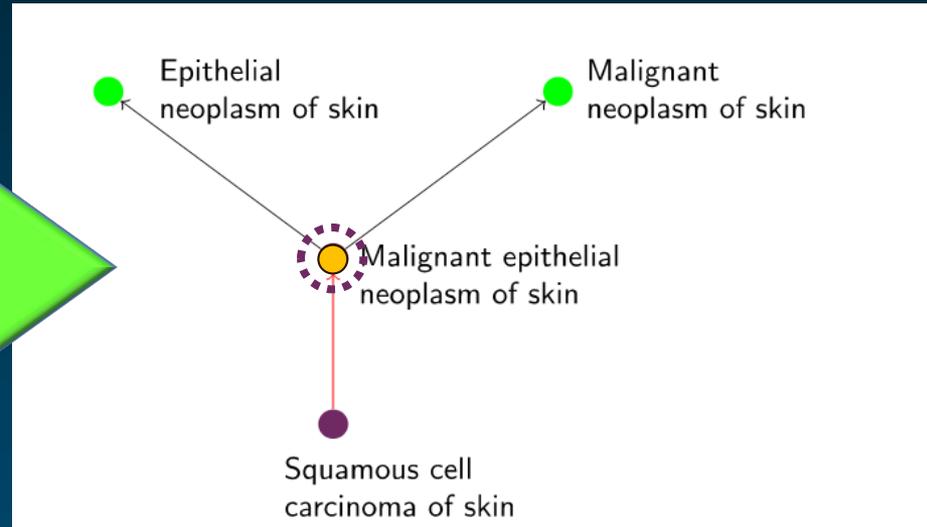


# Lexical patterns (3) Union

Non-lattice subgraph



Suggested remediation



*Epithelial* neoplasm of skin  $\cup$  *Malignant* neoplasm of skin

=

*Malignant epithelial* neoplasm of skin

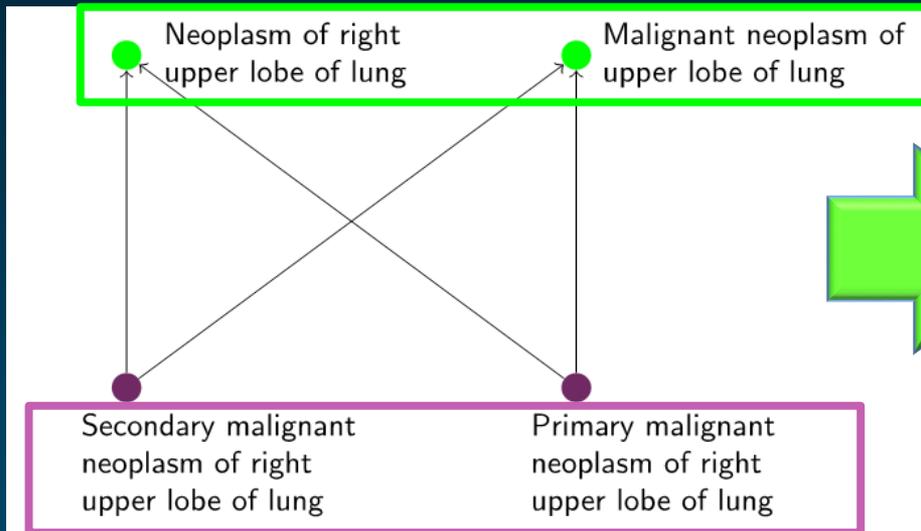
## Lexical patterns (4) Union-Intersection

- ◆ The union of the sets of words for concepts in the upper bounds is equal to the intersection of sets of words for concepts in the lower bounds
- ◆ Suggests a *missing intermediary concept* between the upper bounds and the lower bounds
- ◆ 61 small non-lattice subgraphs with this pattern

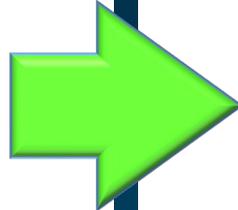
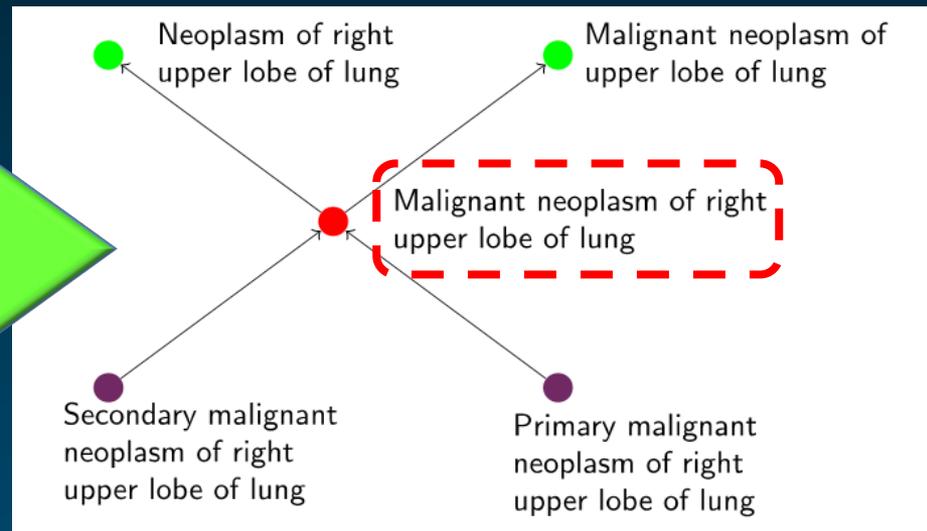


# Lexical patterns (4) Union-Intersection

Non-lattice subgraph



Suggested remediation



Neoplasm of *right* upper lobe of lung



*Malignant* neoplasm of upper lobe of lung



*Secondary malignant* neoplasm of right upper lobe of lung



*Primary malignant* neoplasm of right upper lobe of lung

# Evaluation

- ◆ 59 subgraphs independently reviewed by 2 experts after triaging
  - Differences resolved by discussion
- ◆ All contained errors – 61 errors
  - Missing hierarchical relation: 59
  - Missing intermediary concept: 2
- ◆ Lexical patterns
  - Containment: 34; Intersection: 14; Union: 8; U/I: 3
- ◆ Suggested remediation
  - Accepted for 53 subgraphs
  - Rejected for 6 subgraphs (deeper modeling issues)



# Significance

- ◆ Most terminology QA techniques merely identify potential errors
- ◆ Our approach
  - Identified unreported errors
    - Confirmed by experts
  - Suggested appropriate remediation in many cases
- ◆ Should greatly facilitate error correction by the developers of SNOMED CT
- ◆ Scalable and applicable to other terminologies



# Limitations and future work

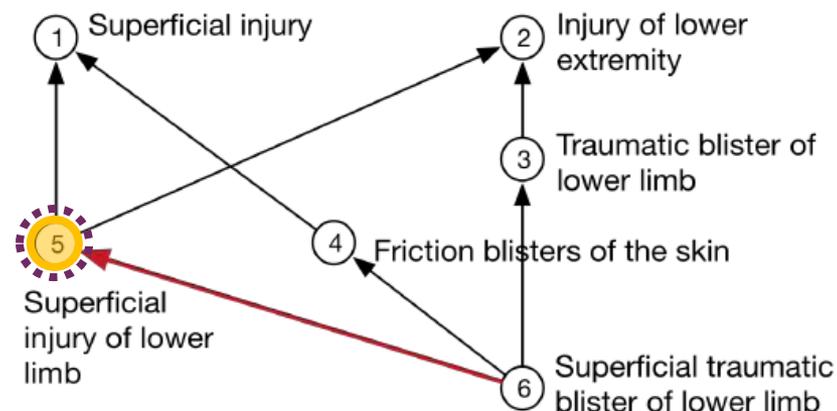
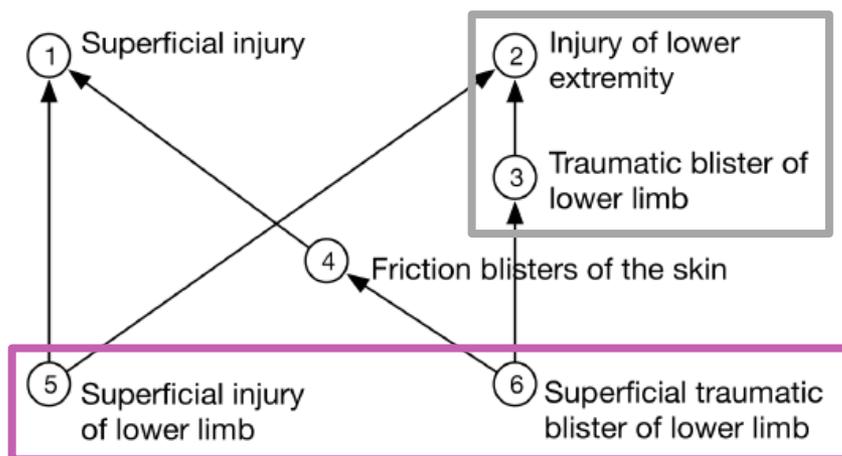
- ◆ Suggested remediation (e.g., to add missing hierarchical relations) is based on the inferred concept hierarchy of SNOMED CT
  - Does not address the root cause (e.g., incomplete/inaccurate logical definition)
  - Root cause needs to be addressed by the SNOMED CT editors
- ◆ Only 4 lexical patterns considered
  - Could be refined with additional patterns



# Follow-up investigation

## ◆ Additional lexical patterns

- Enrich bags of words with
  - Words from ancestors in the non-lattice subgraph
  - Pairs of hypernyms harvested from the subgraph



olivier.bodenreider@nih.gov  
<https://mor.nlm.nih.gov/>



U.S. National Library of Medicine