

## **Mining Non-Lattice Subgraphs for Detecting Missing Hierarchical Relations and Concepts in SNOMED CT**

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

<sup>1</sup>Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA

<sup>2</sup>Division of Biomedical Informatics, College of Medicine, University of Kentucky, Lexington, KY 40536, USA

<sup>3</sup>Institute of Biomedical Informatics, University of Kentucky, Lexington, KY 40536, USA

<sup>4</sup>National Library of Medicine, Bethesda, MD 20892, USA

*Key words:* SNOMED CT, Ontology, Quality Assurance, Non-Lattice Subgraph

Word Count: 4268

\*Corresponding authors: Guo-Qiang Zhang, Licong Cui, 725 Rose Street, Multidisciplinary Science Building 230, Lexington, KY, 40536, USA. [gq.zhang@uky.edu](mailto:gq.zhang@uky.edu), [licong.cui@uky.edu](mailto:licong.cui@uky.edu).

Telephone: 859-323-3162, Fax: 859-323-0069.

## 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 using a scalable MapReduce algorithm. Four lexical patterns (the lexical part) are identified among the extracted non-lattice subgraphs. Non-lattice subgraphs exhibiting such lexical patterns are often indicative of missing hierarchical relations or concepts. Each lexical pattern is associated with a potential specific type of error.

**Results:** Applying the structural-lexical method to SNOMED CT (September 2015 U.S. edition), we found 6,801 non-lattice subgraphs that matched these lexical patterns, of which 2,046 were amenable to visual inspection. We evaluated a random sample of 100 small subgraphs, of which 59 were reviewed in detail by domain experts. All the subgraphs reviewed contained errors confirmed by the experts. The most frequent type of error was missing *is-a* relations due to incomplete or inconsistent modeling of the concepts.

**Conclusions:** Our hybrid structural-lexical method is innovative and proved effective not only in detecting errors in SNOMED CT, but also in suggesting remediation for these errors.

## **OBJECTIVES**

The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) is the most comprehensive clinical healthcare terminology worldwide and its use is mandated in the U.S. as part of the Meaningful Use incentive program. Quality assurance is an indispensable part of the lifecycle management of biomedical terminologies, including SNOMED CT.[1] However, quality assurance of such a large terminology system is difficult due to its sheer size and complex structure. Effective, automated approaches for improving the quality of SNOMED CT are needed to overcome the limitations of manual work.

In this paper, we introduce a novel approach to systematically identify inconsistencies, such as missing hierarchical relations and concepts in SNOMED CT, based on the structural properties of non-lattice subgraphs and the lexical properties of concepts involved in these subgraphs. A random subset of subgraphs automatically generated using this approach was reviewed by domain experts to confirm the uncovered inconsistencies.

## **BACKGROUND AND SIGNIFICANCE**

### **SNOMED CT**

Biomedical ontologies play an important role in healthcare information management, biomedical information extraction, and data integration.[2] SNOMED CT, managed by the International Health Terminology Standards Development Organisation (IHTSDO), is the largest clinical terminology worldwide. SNOMED CT supports the development of high-quality EHRs and facilitates information retrieval, semantic interoperability, and clinical decision support and quality measures.[3,4] Under the Health Information Technology for Economic and Clinical

Health (HITECH) Act,[5] SNOMED CT has been required in the U.S. for encoding relevant clinical information in certified Electronic Health Record (EHR) systems.[6]

SNOMED CT contains over 300,000 concepts organized in 19 top-level hierarchies including *Clinical finding, Procedure, Body structure, and Substance*. Each concept in SNOMED CT represents a unique clinical meaning and is assigned a unique identifier, as well as a unique fully specified name. There are over 1,360,000 relations among these concepts, relating concepts using *subtype relationships* (a.k.a. *is-a*) and attribute relationships (e.g., *associated morphology, causative agent, finding site*).[7]

### **Quality assurance of SNOMED CT**

Given its size and complexity, it is unavoidable that errors are introduced in SNOMED CT as part of its development, update, and maintenance lifecycle. It is impractical for domain experts to systematically detect errors and inconsistencies purely based on manual review. Automatic and effective approaches to quality assurance are highly desirable, moving domain experts' role towards review and confirmation of automatically uncovered error candidates, and, ideally, correction of these errors in subsequent versions.

Researchers have proposed lexical, structural, and semantic methods for auditing and quality improvement of biomedical terminologies.[8-10] Bodenreider et al. evaluated the consistency of SNOMED using lexical methods.[11] Agrawal and Elhanan proposed a lexical method to detect inconsistencies in the formal definitions in SNOMED CT.[12] Jiang and Chute audited the semantic completeness of SNOMED CT using Formal Concept Analysis (FCA), a structural method, and identified missing concepts.[13] Rector and Iannone audited the use of common qualifiers in SNOMED CT definitions by combining lexical and semantic techniques.[14] Wang et al. proposed structural methodologies based on abstraction networks to detect erroneous

concepts in SNOMED CT.[15-17] Ochs et al. presented subject-based and “tribal-based” abstraction network methods to audit SNOMED CT.[18, 19] Zhang and Bodenreider proposed a lattice-based approach to structurally and exhaustively audit SNOMED CT.[20, 21]

### **Lattice-based structural auditing of SNOMED CT**

A lattice is a specific type of directed acyclic graph (DAG) such that any two nodes have a unique maximal common descendant, as well as a unique minimal common ancestor. A lattice is a desirable structural property for a well-formed ontology.[20-22] The philosophical and mathematical reason for this can be elucidated using Formal Concept Analysis, a theory for the formalization of concepts and concept hierarchies (or ontologies) from a collection of objects and their attributes. Each concept represents the set of objects (called extension) that share the same attributes (called intension) of the concept. The concept hierarchy derived always forms a complete lattice.[23,24]

Concepts in SNOMED CT can have multiple parents and are structured as a rooted DAG with respect to the *is-a* taxonomic relationship. However, the SNOMED CT concept hierarchy does not form a lattice.[20] This suggests that investigating concept pairs violating the lattice property (or non-lattice pairs) provides a mechanism for identifying potentially problematic fragments in SNOMED CT, regardless of the type of error involved (e.g., missing intermediary concept, missing hierarchical relation). For example, in Figure 1A the concept pairs *Irritable bowel syndrome variant of childhood* and *Irritable bowel syndrome with diarrhea* have both *Irritable bowel syndrome* and *Disorder of colon* as shared parents. A hierarchical structure in which two concepts have multiple shared parents is a special case of non-lattice fragment. Moreover, *Irritable bowel syndrome* is not classified as a *Disorder of colon*, as it should. This is a typical example of missing *is-a* relation causing a non-lattice fragment. If *Irritable bowel*

*syndrome* was placed as a child of *Disorder of colon, Irritable bowel syndrome variant of childhood* and *Irritable bowel syndrome with diarrhea* would only have *Irritable bowel syndrome* as a single shared parent, and the hierarchical structure would become a lattice (see Figure 1B). This example illustrates our approach for lattice-based ontology quality improvement. A non-lattice fragment represents a possible error, typically a missing hierarchical relation or missing intermediary concept. After correcting the error, the hierarchical structure acquires the properties of a lattice. In this example, the shared ancestors were direct parents. More generally, however, non-lattice fragments may involve shared ancestors beyond direct parents, making their identification a non-trivial, computationally intensive task.

The lattice-based approach [20] to auditing SNOMED CT aims to systematically detect all non-lattice pairs for further analysis. In early experiments using an RDF triple store and SPARQL queries, it took nearly three months to compute all the non-lattice pairs in the July 2009 version of SNOMED CT using a high-end desktop machine.[21] In more recent work,[25,26] it took less than 3 hours using MapReduce parallel processing framework in a 30-node Hadoop cloud.

### **Specific contribution**

The specific contribution of this work is to combine structural and lexical information for the identification of missing hierarchical relations or missing intermediary concepts in SNOMED CT. We extend our earlier work on non-lattice subgraphs by incorporating lexical patterns to precisely identify error types in SNOMED CT, along with suggestions for remediation.

Compared to other methods developed for quality assurance in SNOMED CT, the main difference in our approach is that other methods only identify potential errors, while we also provide remediation for the errors identified.

## MATERIALS AND METHODS

Our approach to identifying potential errors in SNOMED CT based on structural and lexical information can be summarized as follows. We identify non-lattice pairs in SNOMED CT and generate the corresponding non-lattice subgraphs. We identify lexical patterns indicative of missing concepts or hierarchical relations, which we apply to the non-lattice subgraphs. Finally, experts evaluate a sample of the potential errors detected, as well as the proposed remediation. We used the distribution files of the September 2015 version of SNOMED CT (U.S. edition).

### Identifying non-lattice pairs

A non-lattice pair is a concept pair having more than one maximal shared common descendant. A non-lattice pair determines a graph fragment consisting of the concepts between any member of the non-lattice pair and any member of the maximal shared common descendants.

It is possible that multiple non-lattice pairs have identical maximal common descendants.

For example, in Figure 2, three non-lattice pairs: *Neoplasm of pancreas* and *Mass of pancreas* ( $p_1$ ), *Neoplasm of pancreas* and *Neoplasm of digestive organ* ( $p_2$ ), and *Mass of pancreas* and *Neoplasm of digestive organ* ( $p_3$ ), share the same maximal common descendants *Benign neoplasm of pancreas* and *Tumor of exocrine pancreas*. It would not be economical to analyze each of the three non-lattice pairs separately. Moreover, simple aggregation of all non-lattice pairs with the same maximal common descendants may include concepts with ancestor-descendant relationship, which may again result in redundant analysis.

### Identifying non-lattice subgraphs

To avoid such redundant subgraphs, we introduce the notion of non-lattice subgraphs to only include the minimal concepts sharing the same maximal common descendants. Here a **non-**

**lattice subgraph** is determined by a given non-lattice pair  $p = (c_1, c_2)$  and its maximal common descendants  $mcd(p)$ , and can be obtained by

- reversely computing the minimal common ancestors of the maximal common descendants, denoted by  $mca(mcd(p))$ .
- aggregating all the concepts and edges between (including) any concept in  $mca(mcd(p))$  and any of the maximal common descendants  $mcd(p)$ .

We call  $mca(mcd(p))$  and  $mcd(p)$  the **upper bounds** and **lower bounds** of the non-lattice subgraph, respectively. For the three non-lattice pairs  $p_1, p_2$ , and  $p_3$  in Figure 2, they derive the same non-lattice subgraph shown in Figure 2. The **size** of a non-lattice subgraph is the number of concepts it contains. Thus the subgraph in Figure 2 is of size 6.

In previous work, we computed the maximal common descendants for each candidate pair of concepts using a MapReduce pipeline in order to generate an exhaustive list of non-lattice pairs.[25,26] Concept pairs with more than one maximal shared common descendant were identified as non-lattice pairs. To determine the non-lattice subgraphs suitable for error pattern mining, we used all non-lattice pairs as seeds and generated non-lattice subgraphs by modifying the MapReduce pipeline to compute  $mca(mcd(p))$  for each candidate pair  $p = (c_1, c_2)$ .

### **Identifying lexical patterns indicative of missing concepts and relations**

Because it is impractical to manually review large numbers of non-lattice subgraphs, we introduce an automatic approach leveraging additional lexical information (concept names) to identify lexical patterns in non-lattice subgraphs indicative of certain types of errors. We consider the fully specified name of a concept  $c$  as a set (bag) of words in lower case  $\{c\}$ . For instance, the fully specified name of the concept ID 235838003, ( $c$ ), is *Irritable bowel syndrome variant of childhood* (see Figure 1), and its set of words,  $\{c\}$ , is  $\{irritable, bowel, syndrome,$

*variant, of, childhood*}. Utilizing the information of sets of words for concepts in the upper and lower bounds, we define four such patterns indicative of a situation where hierarchical relations or intermediary concepts may be missing.

**Containment:** The set of words for one concept in the upper bounds is contained in the set of words for another concept in the upper bounds; or the set of words for one concept in the lower bounds is contained in the set of words for another concept in the lower bounds. This situation generally suggests a **missing hierarchical relation between concepts in the upper bounds (or in the lower bounds)**. For instance, in the lower bounds of the non-lattice subgraph in Figure 3A, {*duodenal, ulcer, with, perforation, and, obstruction*} is contained in {*chronic, duodenal, ulcer, with, perforation, and, obstruction*}. Here, there is a missing hierarchical relation between concepts in the lower bounds, because *Chronic duodenal ulcer with perforation AND obstruction* is more specific than *Duodenal ulcer with perforation AND obstruction*. Of note, for this pattern, we specifically excluded non-lattice subgraphs with concepts that contain negation words such as *not, no, without, absence, and except*, because a missing hierarchical relation would be wrongly suggested between the concept with the negation and the same concept without negation. For example, the set of words for the concept *Anemia during pregnancy - baby not yet delivered* contains that of the concept *Anemia during pregnancy - baby delivered* as a subset, but the two concepts are obviously not hierarchically related.

**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. This situation generally suggests a **missing hierarchical relation between concepts in the upper bounds**. For example, in Figure 3C, the intersection of {*irritable, bowel, syndrome, variant, of, childhood*} and {*irritable, bowel, syndrome, with, diarrhea*} is {*irritable, bowel, syndrome*}, which is equal to the set of words for

the concept *Irritable bowel syndrome* in the upper bounds. Here, there is a missing hierarchical relation between concepts in the upper bounds, because *Irritable bowel syndrome* is more specific than *Disorder of colon*.

**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. This situation generally suggests a **missing hierarchical relation between concepts in the lower bounds**. For instance, in Figure 3E, the union of {*epithelial, neoplasm, of, skin*} and {*malignant, neoplasm, of, skin*} is {*malignant, epithelial, neoplasm, of, skin*}, which is equal to the set of words for the concept *Malignant epithelial neoplasm of skin* in the lower bounds. Here, there is a missing hierarchical relation between concepts in the lower bounds, because *Squamous cell carcinoma of skin* is more specific than *Malignant epithelial neoplasm of skin*.

**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. This situation generally suggests a **missing intermediary concept between the upper bounds and the lower bounds**. For instance, in Figure 3G, the union of {*neoplasm, right, upper, lobe, of, lung*} and {*malignant, neoplasm, upper, lobe, of, lung*} is {*malignant, neoplasm, right, upper, lobe, of, lung*}, which is equal to the intersection of {*secondary, malignant, neoplasm, right, upper, lobe, of, lung*} and {*primary, malignant, neoplasm, right, upper, lobe, of, lung*}. Here, there is a missing concept, *Malignant neoplasm of right upper lobe of lung*, representing the features common to the two concepts in the lower bounds (*Primary malignant neoplasm of right upper lobe of lung* and *Secondary malignant neoplasm of right upper lobe of lung*), inherited from both concepts in the upper bounds (*Malignant neoplasm of upper lobe of lung* and *Neoplasm of right upper lobe of lung*).

### Analyzing non-lattice subgraphs with lexical patterns

As shown above, these patterns may suggest remediation strategies for transforming a non-lattice subgraph into a lattice subgraph. For example, for the non-lattice subgraph in Figure 3A exhibiting a Containment pattern (indicative of a missing hierarchical relation between concepts in the upper bounds or in the lower bounds), there is indeed a missing hierarchical relation between the two lower bound concepts *Duodenal ulcer with perforation AND obstruction* and *Chronic duodenal ulcer with perforation AND obstruction*, because the added notion of chronicity makes the latter more specific. The suggested correction is to add the relation *Chronic duodenal ulcer with perforation AND obstruction is-a Duodenal ulcer with perforation AND obstruction* (see Figure 3B).

For the non-lattice subgraph in Figure 3C exhibiting an Intersection pattern (indicative of a missing hierarchical relation between concepts in the upper bounds), there is indeed a missing hierarchical relation between the two upper bound concepts *Irritable bowel syndrome* and *Disorder of colon*, because the colon is the anatomical location of this syndrome. The suggested correction is to add the relation *Irritable bowel syndrome is-a Disorder of colon* (see Figure 3D).

For the non-lattice subgraph in Figure 3E exhibiting a Union pattern (indicative of a missing hierarchical relation between concepts in the lower bounds), there is indeed a missing hierarchical relation between the two lower bound concepts *Malignant epithelial neoplasm of skin* and *Squamous cell carcinoma of skin*, because squamous cell carcinoma is a type of malignant epithelial neoplasm. The suggested correction is to add the relation *Squamous cell carcinoma of skin is-a Malignant epithelial neoplasm of skin* (see Figure 3F).

For the non-lattice subgraph in Figure 3G exhibiting a Union-Intersection pattern (indicative of a missing intermediary concept between the upper bounds and the lower bounds), a concept

*Malignant neoplasm of right upper lobe of lung* is indeed missing between the concepts in the lower bounds and the concepts in the upper bounds (see Figure 3H), because the characteristics malignant (neoplasm) and right (upper lobe), each represented by one concept in the upper bounds are both shared by the two concepts in the lower bounds.

It is worth noting that **smaller non-lattice subgraphs** may be contained in larger subgraphs. As a consequence, correcting errors in smaller non-lattice subgraphs will mechanically result in the correction of the same errors in larger subgraphs that contain these smaller subgraphs. For instance, Figure 4A shows an example of a size-9 non-lattice subgraph that contains a size-5 non-lattice subgraph (in dashed red circle). A possible correction for this size-5 non-lattice subgraph (exhibiting a Containment pattern) is to add the relation *Malignant hypertensive end stage renal disease on dialysis is-a Malignant hypertensive end stage renal disease*. Applying this correction in the size-5 non-lattice subgraph will also eliminate the same error (dashed red circle) in the larger non-lattice subgraph in Figure 4A. Moreover, the larger subgraph may become a smaller non-lattice subgraph after correction (which may contain additional errors). For example, Figure 4B shows the resulting size-7 non-lattice subgraph obtained after applying the above-mentioned correction, which exhibits a Union pattern. A possible further correction for this size-7 non-lattice subgraph is to add two relations *Hypertensive renal disease with end stage renal failure is-a Hypertensive end stage renal disease*, and *Malignant hypertensive end stage renal disease is-a Hypertensive end stage renal disease*. Figure 4C presents the resulting graph after fixing all possible errors in the size-9 non-lattice subgraph in Figure 4A. Note that this subgraph, shown in Figure 4C, is no longer a non-lattice subgraph, i.e., it does not violate the lattice property.

In this paper, we focused our investigation on small non-lattice subgraphs of size 4, 5, or 6. These small subgraphs are easier to inspect visually, and they are embedded in nearly 50% of all non-lattice subgraphs (see the Results section for details).

## **Evaluation**

To assess the effectiveness of our method in identifying real errors in SNOMED CT, we focused on small non-lattice subgraphs following any of the four lexical patterns. A random sample of 100 such subgraphs was selected from the two largest subhierarchies: *Clinical finding* and *Procedure*. The sample non-lattice subgraphs were rendered in Scalable Vector Graphics (SVGs) to facilitate visualization and evaluation by experts.

To minimize the time and effort needed by the experts to review the subgraphs, author GQZ first triaged the 100 non-lattice subgraphs, eliminating the most complex cases (e.g., subgraphs with multiple problems), as well as cases for which IHTSDO would be unlikely to integrate the suggested correction. For example, the triaged subgraphs include those with terms containing “AND/OR,” which are progressively being eliminated by IHTSDO. Other examples include cases requiring systematic pre-coordination, which IHTSDO tends to avoid (e.g., “missing” intermediary concept *Tobramycin measurement in blood* between the lower bounds *Serum tobramycin measurement* and *Plasma tobramycin measurement*, and the upper bounds *Measurement of level of drug in blood* and *Tobramycin measurement*).

Authors JTC and OB, clinical experts familiar with SNOMED CT, independently reviewed the erroneous subgraphs selected by GQZ and their suggested remediation. Differences between the two experts were resolved by discussion.

## RESULTS

### Identifying non-lattice pairs and subgraphs

631,006 non-lattice pairs were found in the September 2015 version of the SNOMED CT (U.S. edition). From these pairs, 171,011 non-lattice subgraphs were generated, whose sizes ranged from 4 to 5,137. About 90% of the non-lattice subgraphs had sizes 4 to 100 (see online supplementary appendix I for the distribution of non-lattice subgraphs by size), with size 6 being the most frequent (6,541).

**Small non-lattice subgraphs.** A total of 3,339 non-lattice subgraphs of size 4 were contained in 28,292 larger subgraphs, 3,773 subgraphs of size 5 were contained in 34,808 larger subgraphs, and 5,342 subgraphs of size 6 were contained in 40,404 larger subgraphs. In total, 70,250 distinct larger non-lattice subgraphs contained smaller subgraphs of size 4, 5, or 6. Moreover, none of the size-4 non-lattice subgraphs were contained in any size-5 subgraphs, and none of the size-5 subgraphs were contained size-6 subgraphs. Only 197 size-4 non-lattice subgraphs were contained in size-6 subgraphs. Overall, nearly half of the non-lattice subgraphs are related to subgraphs of size 4, 5, or 6 (i.e., either they are size-4, size-5, or size-6 non-lattice subgraphs themselves, or they are larger non-lattice subgraphs containing these smaller subgraphs).

### Analyzing non-lattice subgraphs with lexical patterns

6,801 non-lattice subgraphs were found exhibiting any of the four lexical patterns, among which 2,046 were small non-lattice subgraphs (of size 4, 5, and 6). These small subgraphs exhibiting any of the four lexical patterns were contained in 15,776 larger non-lattice subgraphs. Table 1 shows the distribution of small non-lattice subgraphs exhibiting each pattern by size. The Intersection pattern accounted for the largest proportion (1,085). Table 2 presents the distribution of small non-lattice subgraphs exhibiting any of the four lexical patterns by SNOMED CT

subhierarchy. *Clinical finding*, the largest subhierarchy in SNOMED CT, accounted for the largest number. Of the 2,046 smaller subgraphs, 1,300 were in in two classes, namely *Clinical Finding* (728) and *Procedure* (572).

**Table 1** Number of small non-lattice subgraphs exhibiting any of the four lexical patterns (Containment, Intersection, Union, and Union-Intersection) according to the size of non-lattice subgraphs.

	Number of non-lattice subgraphs				
	Containment	Intersection	Union	Union-Intersection	Total
<b>Size 4</b>	160	336	31	17	544
<b>Size 5</b>	229	291	75	13	608
<b>Size 6</b>	347	458	58	31	894
<b>Total</b>	736	1,085	164	61	2,046

**Table 2** Number of small non-lattice subgraphs (of size 4, 5, and 6) exhibiting any of the four lexical patterns according to the SNOMED CT subhierarchy.

Subhierarchy	Total
<i>Clinical finding</i>	728
<i>Procedure</i>	572
<i>Body structure</i>	267
<i>Pharmaceutical/biologic product</i>	202
<i>Substance</i>	115
<i>Physical object</i>	71

<i>Qualifier value</i>	20
<i>Specimen</i>	19
<i>Organism</i>	17
<i>Social context</i>	15
<i>Observable entity</i>	9
<i>Situation with explicit context</i>	7
<i>Environment or geographical location</i>	2
<i>Event</i>	1
<i>Physical force</i>	1
<b>Total</b>	2,046

## **Evaluation**

Of the 100 subgraphs randomly selected from the 1,300 small-size subgraphs from the two main hierarchies on SNOMED CT, 65 were in the *Clinical finding* subhierarchy and 35 in the *Procedure* subhierarchy. Of these subgraphs, 37 exhibited the Containment pattern, 46 the Intersection pattern, 13 the Union pattern, and 4 the Union-Intersection pattern.

Of the 100 non-lattice subgraphs, 59 were triaged for review by the medical experts. In each case, the experts confirmed the existence of an error. Therefore, the error rate among the 100 subgraphs is at least 59%, since some erroneous subgraphs may not have been selected for review during the triage process.

Among the 59 erroneous subgraphs examined, 34 exhibited a Containment pattern, 14 an Intersection pattern, 8 a Union pattern, and 3 a Union-Intersection pattern. These 59 erroneous subgraphs were contained in 656 larger non-lattice subgraphs, indicating that fixing errors in

these 59 subgraphs will automatically eliminate similar errors in 656 larger subgraphs (although additional errors may remain in the larger subgraphs).

For 6 of the erroneous non-lattice subgraphs, although the experts acknowledged the existence of an error, they rejected the suggested remediation, because manual examination revealed deeper modeling issues in SNOMED CT that needed further investigation. Analysis of the 53 other erroneous subgraphs resulted in a total of 61 verified errors (see online supplementary appendix II for the visualized non-lattice subgraphs and corrections). Figure 5 shows four examples of the non-lattice subgraphs that were evaluated, as well as their verified corrections. Note that an erroneous non-lattice subgraph may reveal multiple errors and suggested changes. For example, Figure 5E is a non-lattice subgraph of size 5, and its analysis revealed two missing *is-a* relations: *Nevus of choroid of left eye is-a Nevus of choroid*, and *Nevus of choroid of right eye is-a Nevus of choroid* (see Figure 5F).

Among the 61 suggested corrections, 59 were missing *is-a* relations and 2 were missing concepts. Table 3 lists 10 examples of the verified missing *is-a* relations (see online supplementary appendix III for a complete list of corrections). We will submit these suggested corrections to IHTSDO through the regular content request submission process for inclusion in the ongoing internal quality improvement activities being undertaken by IHTSDO.

**Table 3** Ten examples of missing *is-a* relations in the SNOMED CT, along with the lexical patterns of their corresponding non-lattice subgraphs and the location of the missing relation (LB: lower bound; UB: upper bound).

Child	Parent	Pattern	Location of the missing

			<b>relation</b>
<i>Acute exacerbation of chronic obstructive bronchitis</i>	<i>Acute exacerbation of chronic bronchitis</i>	Containment (Figure 5: A, B)	LB→LB
<i>Compartment syndrome of abdomen due to trauma</i>	<i>Abdominal compartment syndrome</i>	Intersection	LB→LB
<i>Recurrent rheumatic heart disease</i>	<i>Chronic rheumatic heart disease</i>	Union (Figure 5: C, D)	LB→LB
<i>Removal of foreign body of cornea by incision</i>	<i>Incision of cornea</i>	Intersection	LB→LB
<i>Acute endometritis</i>	<i>Acute uterine in inflammatory disease</i>	Intersection	UB→UB
<i>Nevus of choroid of left eye</i>	<i>Nevus of choroid</i>	Containment (Figure 5: E, F)	LB→LB
<i>Nevus of choroid of right eye</i>	<i>Nevus of choroid</i>	Containment (Figure 5: E, F)	LB→LB
<i>Acromioclavicular joint pain</i>	<i>Shoulder joint pain</i>	Union	LB→LB
<i>Benign neoplasm of skin of forearm</i>	<i>Benign neoplasm of soft tissue of forearm</i>	Intersection (Figure 5: G, H)	LB→LB
<i>Cervical spondylosis with myelopathy</i>	<i>Cervical spondylosis</i>	Containment	LB→LB

## DISCUSSION

### Significance

In this paper, we mined non-lattice subgraphs exhibiting four lexical patterns to uncover missing hierarchical relations or missing concepts in SNOMED CT. Our approach not only uncovered novel (i.e., unreported) SNOMED CT errors, but also suggested appropriate remediation in many

cases. While most approaches to quality assurance in SNOMED CT merely indicate the presence of a possible error, our hybrid approach overlays lexical information onto structural information to analyze the precise nature of the error and propose a correction. The ability to suggest remediation for the errors we identify, sets us apart from other methods and will likely drive adoption. Focusing on non-lattice subgraphs of smaller size provides an effective way of auditing hierarchical relations in SNOMED CT. Not only is it easier for experts to review and examine these graphs, but also the errors found in small graphs are mechanically propagated to larger graphs. Since virtually all biomedical ontologies are organized into subsumption hierarchies and have concept names, our non-lattice-based approach can be generalized and applied to other biomedical terminologies for quality assurance purposes.

### **Practical quality impact of suggested SNOMED CT remediation**

Addressing quality issues in SNOMED CT can improve the quality in downstream information systems and tools relying on its hierarchies.[3] Practical areas of impact include value set definition for EHR decision support, quality reporting and cohort selection.[4] Value sets are increasingly defined in intension, i.e., as the list of concepts sharing some common feature (e.g., all descendants of *Malignant epithelial neoplasm of skin*). *Squamous cell carcinoma of skin* is currently not listed as one of its descendants, and would thus be missing from the corresponding intensional value set. As a consequence, patients with *Squamous cell carcinoma of skin* would not be selected for a cohort of patients with *Malignant epithelial neoplasm of skin*. Of note, some of the errors we identified involve concepts from the widely used Clinical Observations Recordings and Encoding (CORE) Problem List Subset of SNOMED CT,[27] which contains concepts widely used across many healthcare institutions. For example, *Shoulder joint pain* and *Acromioclavicular joint pain* are two concepts from the CORE subset of SNOMED CT, but a

missing *is-a* relation between the two concepts was identified in this work (see the 43th non-lattice subgraph in the online supplementary appendix II).

### **Generalization**

Most existing approaches on quality assurance of SNOMED CT typically take advantage of specific knowledge in the terminology, such as lexical information [11,12,14] or structural information on specific subhierarchy,[15-19] but have limitations in scalability and applicability.

This work not only leverages both structural and lexical information and is not limited to a specific subhierarchy, but it is also scalable and widely applicable to other terminologies. The scalability of exhaustive computation of non-lattice subgraphs has been demonstrated in our previous work [25,26], where eight versions of SNOMED CT based on all subhierarchies have been used for experiments.

Abstraction networks (AbNs) have been systematically explored for quality assurance of biomedical ontologies from a structural point of view.[15-18] The AbNs approaches utilize “hypergraphs” where each node contains a collection of concepts sharing some common attributes, used for summarizing structural information. Distinct from AbNs, the non-lattice subgraphs in this work are directly based on the same concrete level of the underlying graph structure, rather than on an abstraction thereof. Moreover, our non-lattice subgraphs are generated from the hierarchical *is-a* relationships, while AbNs rely on outgoing attribute relationships for grouping concepts into areas or partial areas. Since concept names and an *is-a* taxonomy, i.e., a hierarchical backbone, are present in virtually all biomedical terminologies, our hybrid approach combining non-lattice subgraph and lexical information is widely applicable for quality assurance purposes.

Another key distinction of this work from other existing terminology quality assurance work, which requires domain experts' manual review to uncover potential errors, is the potential of automatically suggesting remediation for potential errors uncovered, saving domain experts' manual review effort.

### **Failure analysis of complex cases**

It is worth noting that the remediation suggested by the presence of a lexical pattern is not always accurate. For example, for the non-lattice subgraph with an Intersection pattern in Figure 5G, the correction associated with the pattern is a missing hierarchical relation between concepts in the upper bounds. In this case, however, the missing hierarchical relation is between concepts in the lower bounds instead. In this example, a related fact is that *Benign neoplasm of skin of forearm is-a Benign neoplasm of soft tissues of upper limb*, which indicates that *skin* is a kind of *soft tissue*, and therefore, the correction is to add the relation *Benign neoplasm of skin of forearm is-a Benign neoplasm of soft tissue of forearm*.

Also note that even though erroneous non-lattice subgraphs may reveal modeling problems in SNOMED CT, they may not be easily fixed by adding a missing *is-a* relation or a missing concept. For instance, Figure 6A presents an erroneous non-lattice subgraph. Here again, the Intersection pattern suggests a missing hierarchical relation between concepts in the upper bounds, i.e., between *Evoked magnetic fields* and *Procedure on central nervous system*. However, *Evoked magnetic fields* is a primitive concept. While adding a hierarchical relation would make this subgraph a lattice, a more sensible solution is to create a complete logical definition for *Evoked magnetic fields*, from which the description logic classifier would simply infer a hierarchical relation to *Procedure on central nervous system*.

## Limitations and future work

A limitation of this work is that our suggested remediation (e.g., to add missing hierarchical relations) is based on the inferred concept hierarchy of SNOMED CT. Since this hierarchy is produced by the description logic classifier based on the logical definitions for the concepts, a more meaningful remediation would be to modify the logical definitions, so that the appropriate hierarchy can be inferred. When we submit the missing hierarchical relations we identified to the IHTSDO, we expect that the IHTSDO editors will address the root cause (i.e., incomplete logical definitions) rather than simply add the relations.

As mentioned earlier, due to the organization of the evaluation, we can only report the lower bound of the rate of identified errors, because there may be errors in the subgraphs that were not selected for review. While this may seem suboptimal, our choice was justified by (1) the need to minimize the workload of medical experts in this labor-intensive review process; (2) the purpose of the evaluation was to show the promise of combining non-lattice subgraphs and lexical patterns to not only detect potential errors in SNOMED CT, but also facilitate remediation (as a proof of principle). A larger, more thorough evaluation is planned.

Leveraging lexical patterns proved an effective way to identify potential errors in non-lattice subgraphs. However, the four patterns we consider in this investigation only cover some of the subgraphs. It would be interesting to investigate additional patterns or new lexical approaches. For example, the non-lattice subgraph shown in Figure 6B does not follow any of the four patterns. However, if we considered *neoplasm* and *tumor* as synonyms, it would exhibit the Intersection pattern. Figure 2 illustrates another such example. Finally, we also plan to use all the synonyms in SNOMED CT, as well as additional synonyms from the Unified Medical Language

System (UMLS) Metathesaurus, to complement the fully specified terms used in this investigation.

## **CONCLUSIONS**

In this paper, we introduced a novel hybrid approach leveraging non-lattice subgraphs and lexical information in concept names for detecting missing hierarchical relations or missing concepts in the SNOMED CT. Our approach differs from other quality assurance methods in that we also suggest remediation for the errors identified. We showed that identifying and analyzing small non-lattice subgraphs in the SNOMED CT with lexical patterns is a simple and effective quality assurance technique.

## **Acknowledgement**

The work was supported by University of Kentucky Center for Clinical and Translational Science (Clinical and Translational Science Award UL1TR001998). This work was also supported by the Intramural Research Program of the NIH, National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

## **Contributors**

LC, GQZ, and OB conceptualized and designed this study. LC implemented the idea, generated the auditing results, and prepared the sample non-lattice subgraphs for evaluation. WZ and ST assisted with rendering SVG graphs to facilitate the review and evaluation. GQZ reviewed the random collection of non-lattice subgraphs and proposed initial corrections for evaluation. JTC and OB reviewed and verified the proposed corrections, and suggested additional corrections.

## References

1. Geller J, Perl Y, Halper M, Cornet R. Special issue on auditing of terminologies. *J Biomed Inform*. 2009;42(3):407-11.
2. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*;2008:67-79.
3. Lee D, de Keizer N, Lau F, et al. Literature review of SNOMED CT use. *J Am Med Inform Assoc* 2014;21(e1):e11-9.
4. Winnenburt R, Bodenreider O. Metrics for assessing the quality of value sets in clinical quality measures. *AMIA Annu Symp Proc*;2013:1497-1505.
5. Health Information Technology for Economic and Clinical Health (HITECH) Act. 2009. [http://www.healthit.gov/sites/default/files/hitech\\_act\\_excerpt\\_from\\_arra\\_with\\_index.pdf](http://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf) [Accessed 6 April 2015].
6. ONC Stage 2 Meaningful Use Final Rule. 2012 [Accessed 6 April 2015]. <http://www.gpo.gov/fdsys/pkg/FR-2012-09-04/pdf/2012-20982.pdf>
7. SNOMED CT Starter Guide. 2014 [Accessed 6 April 2015]. [http://ihtsdo.org/fileadmin/user\\_upload/doc/download/doc\\_StarterGuide\\_Current-en-US\\_INT\\_20141202.pdf](http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_StarterGuide_Current-en-US_INT_20141202.pdf)
8. Cimino JJ, Hripcsak G, Johnson S, et al. Designing an introspective, controlled medical vocabulary. *In Proceedings of the Thirteenth Annual SCAMC*;1989:513-518.
9. Cimino JJ. Auditing the unified medical language system with semantic methods. *J Am Med Inform Assoc* 1998;5(1):41-51.
10. Zhu X, Fan JW, Baorto DM, et al. A review of auditing methods applied to the content of controlled biomedical terminologies. *J Biomed Inform* 2009;42(3):413-25.

11. Bodenreider O, Burgun A, Rindflesch TC. Assessing the consistency of a biomedical terminology through lexical knowledge. *Int J Med Inform* 2002;67(1-3):85-95.
12. Agrawal A, Elhanan G. Contrasting lexical similarity and formal definitions in SNOMED CT Consistency and implications. *J Biomed Inform* 2014;47:192-8.
13. Jiang G, Chute CG. Auditing the semantic completeness of SNOMED CT using formal concept analysis. *J Am Med Inform Assoc* 2009;16(1):89-102.
14. Rector A, Iannone L. Lexically suggest, logically define: quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. *J Biomed Inform* 2012;45(2):199-209.
15. Wang Y, Halper M, Min H, et al. Structural methodologies for auditing SNOMED. *J Biomed Inform* 2007;40(5):561-81.
16. Wang Y, Halper M, Wei D, et al. Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED. *J Biomed Inform* 2012;45:15-29.
17. Wang Y, Halper M, Wei D, et al. Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. *J Biomed Inform* 2012;45:1-14.
18. Ochs C, Geller J, Perl Y, et al. Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies. *J Am Med Inform Assoc* 2014;amiajnl-2014-003151.
19. Ochs C, Geller J, Perl Y, et al. A tribal abstraction network for SNOMED CT target hierarchies without attribute relationships. *J Am Med Inform Assoc* 2014;amiajnl-2014-003173.
20. Zhang GQ and Bodenreider O. Using SPARQL to Test for Lattices: application to quality assurance in biomedical ontologies. *The Semantic Web-ISWC* 2010;273-288.

21. Zhang GQ, Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. *AMIA Annu Symp Proc*;2010:922-26.
22. Zweigenbaum P, Bachimont B, Bouaud J, et al. Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods of information in medicine* 1995;34:15-24.
23. Ganter B and Wille R. Formal concept analysis. Springer Berlin 1999.
24. Troy AD, Zhang GQ, Tian Y. Faster concept analysis. In *Conceptual Structures: Knowledge Architectures for Smart Applications 2007*;pp. 206-219.
25. Zhang GQ, Zhu W, Sun M, et al. MaPLE: A MapReduce Pipeline for Lattice-based Evaluation and Its Application to SNOMED CT. *IEEE BigData 2014*;754-9.
26. Cui L, Tao S, Zhang GQ. Biomedical Ontology Quality Assurance Using a Big Data Approach. *ACM Transactions on Knowledge Discovery from Data* 2016;10(4):41.
27. The CORE Problem List Subset of SNOMED CT. 2016 [Accessed 3 October 2016].  
[https://www.nlm.nih.gov/research/umls/Snomed/core\\_subset.html](https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html)

## Figure legends

Figure 1: (A) is an example of a non-lattice pair *Irritable bowel syndrome variant of childhood* and *Irritable bowel syndrome with diarrhea* (in purple) sharing two minimal common ancestors, *Irritable bowel syndrome* and *Disorder of colon* (in green). (B) is a suggested correction for (A). By making *Irritable bowel syndrome* a child of *Disorder of colon*, the subgraph is transformed into a lattice.

Figure 2: An example of non-lattice graph. Three pairs of concepts (among the 3 upper-bound concepts in green) share the same maximal common descendants (the two lower-bound concepts in purple).

Figure 3: Examples of non-lattice subgraphs exhibiting the patterns Containment (3A), Intersection (3C), Union (3E), and Union-Intersection (3G), along with suggested remediation (right-hand side). Left-hand side: the upper-bound concepts are in green, and the lower-bound concepts are in purple. Right-hand side: the suggested missing hierarchical relations and concepts are in red.

Figure 4: (A): An example of a size-9 non-lattice subgraph containing a size-5 non-lattice subgraph (in dashed red circle), where the upper-bound concepts are in green, and the lower-bound concepts are in purple. (B): The resulting size-7 non-lattice subgraph obtained by fixing the size-5 non-lattice subgraph contained in (A). (C): The resulting graph after fixing all possible errors in (A).

Figure 5: Examples of evaluated non-lattice subgraphs (left-hand side), as well as their remediation (right-hand side). The verified correction highlighted in red.

Figure 6: (A): An example of problematic non-lattice subgraph revealing modeling problems. (B): Non-lattice subgraph pattern for which new lexical patterns would be required (e.g., leveraging synonymy between *neoplasm* and *tumor*). For each non-lattice subgraph, the upper-bound concepts are in green, and the lower-bound concepts are in purple.

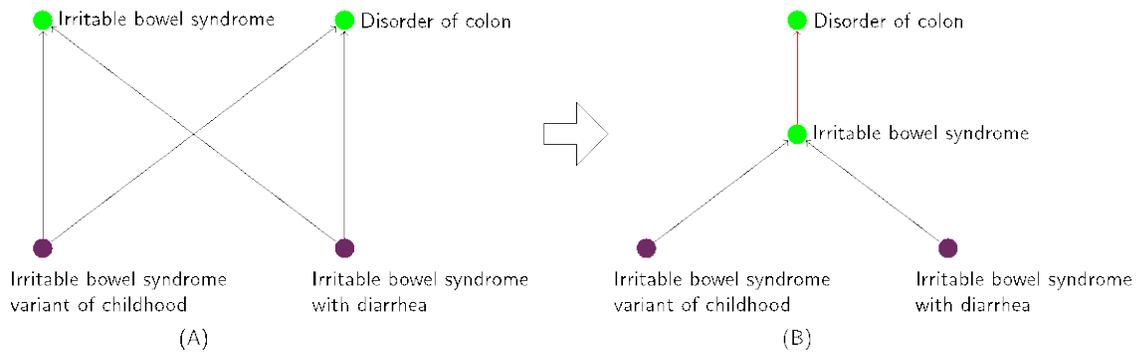


Figure 1: (A) is an example of a non-lattice pair Irritable bowel syndrome variant of childhood and Irritable bowel syndrome with diarrhea (in purple) sharing two minimal common ancestors, Irritable bowel syndrome and Disorder of colon (in green). (B) is a suggested correction for (A). By making Irritable bowel syndrome a child of Disorder of colon, the subgraph is transformed into a lattice.

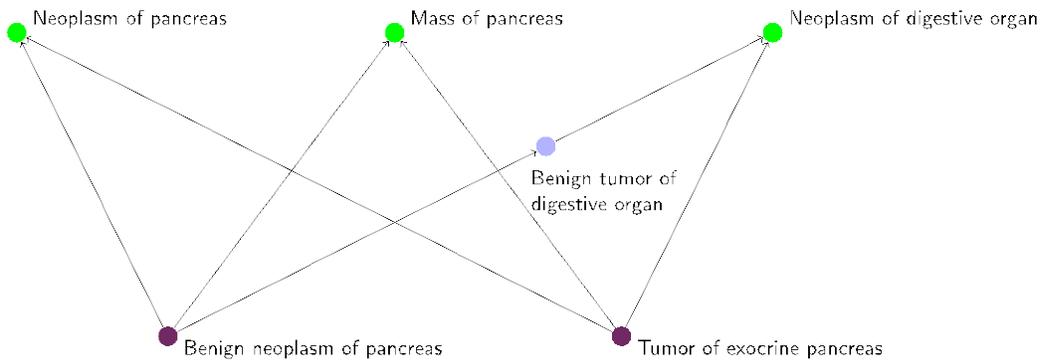
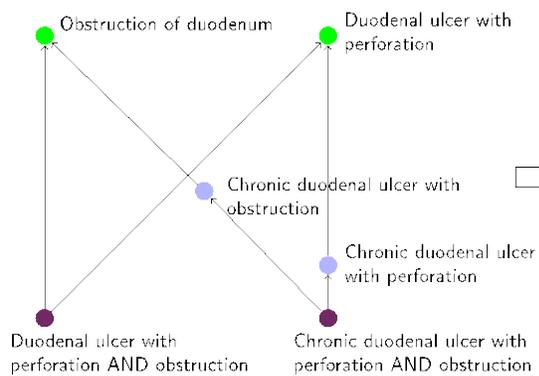
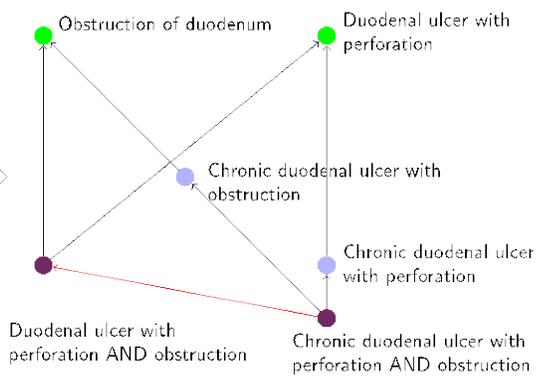


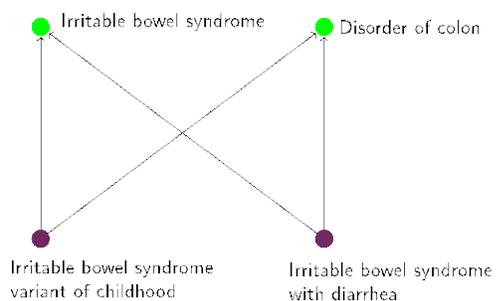
Figure 2: An example of non-lattice graph. Three pairs of concepts (among the 3 upper-bound concepts in green) share the same maximal common descendants (the two lower-bound concepts in purple).



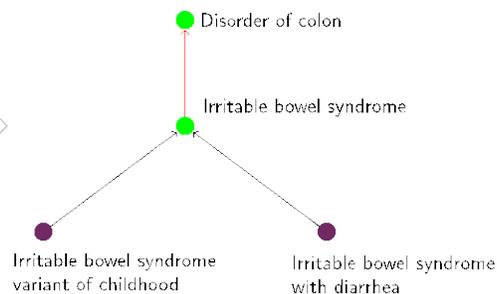
(A) - Pattern Containment



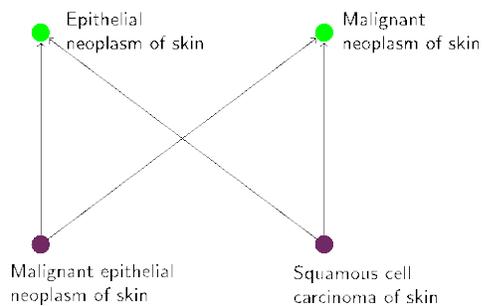
(B)



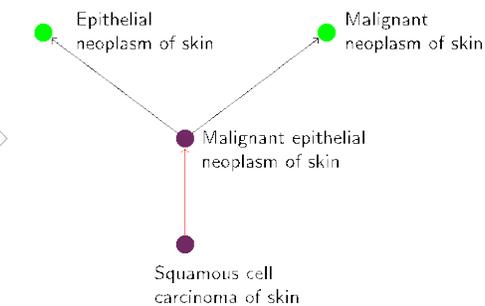
(C) - Pattern Intersection



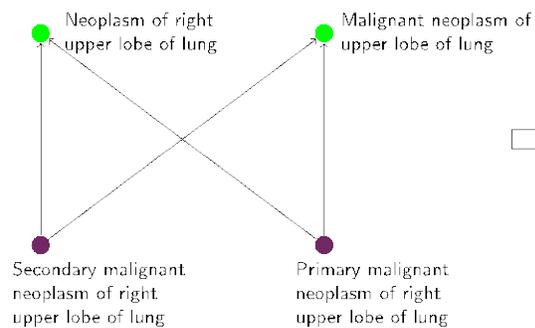
(D)



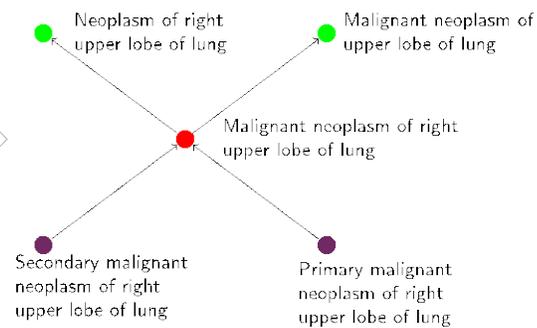
(E) - Pattern Union



(F)

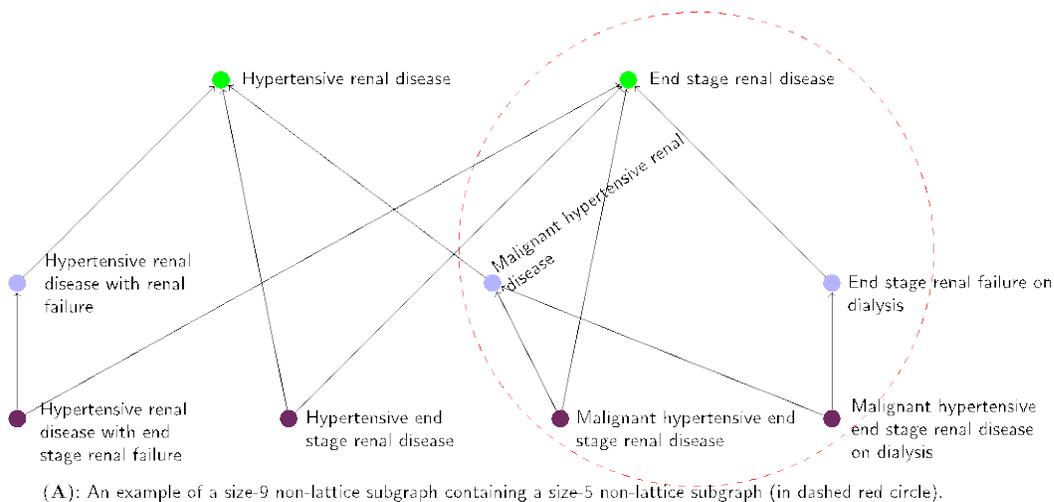


(G) - Pattern Union-Intersection

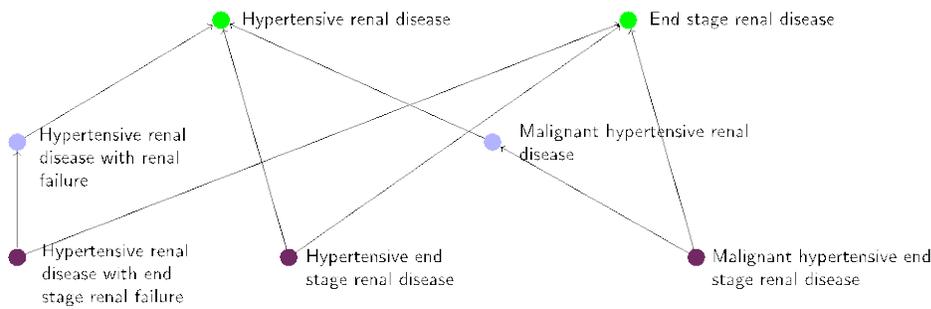


(H)

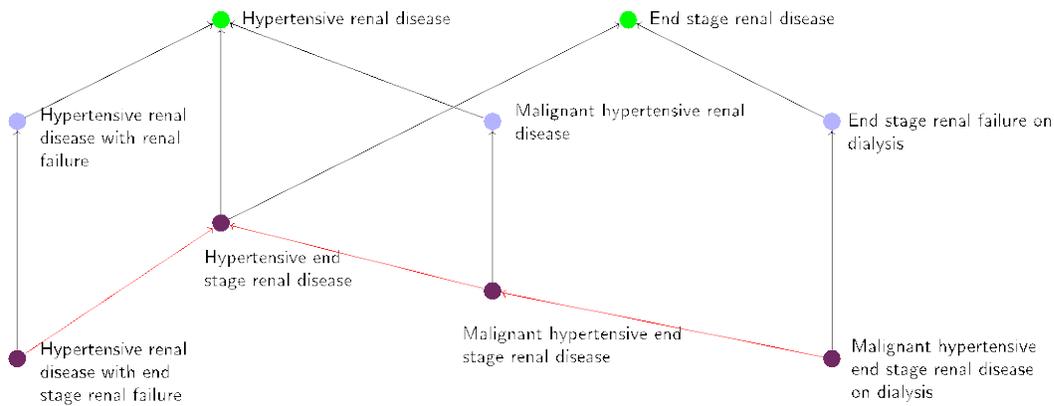
Figure 3: Examples of non-lattice subgraphs exhibiting the patterns Containment (3A), Intersection (3C), Union (3E), and Union-Intersection (3G), along with suggested remediation (right-hand side). Left-hand side: the upper-bound concepts are in green, and the lower-bound concepts are in purple. Right-hand side: the suggested missing hierarchical relations and concepts are in red.



(A): An example of a size-9 non-lattice subgraph containing a size-5 non-lattice subgraph (in dashed red circle).



(B): Resulting size-7 non-lattice subgraph after fixing the size-5 non-lattice subgraph contained in (A). The concepts *End stage renal failure on dialysis* and *Malignant hypertensive end stage renal disease on dialysis* are no longer part of the non-lattice subgraph.



(C): Resulting subgraph after fixing all possible errors in (A). This no longer violates the lattice principle.

Figure 4: (A): An example of a size-9 non-lattice subgraph containing a size-5 non-lattice subgraph (in dashed red circle), where the upper-bound concepts are in green, and the lower-bound concepts are in purple. (B): The resulting size-7 non-lattice subgraph obtained by fixing the size-5 non-lattice subgraph contained in (A). (C): The resulting graph after fixing all possible errors in (A), with corrections highlighted in red.

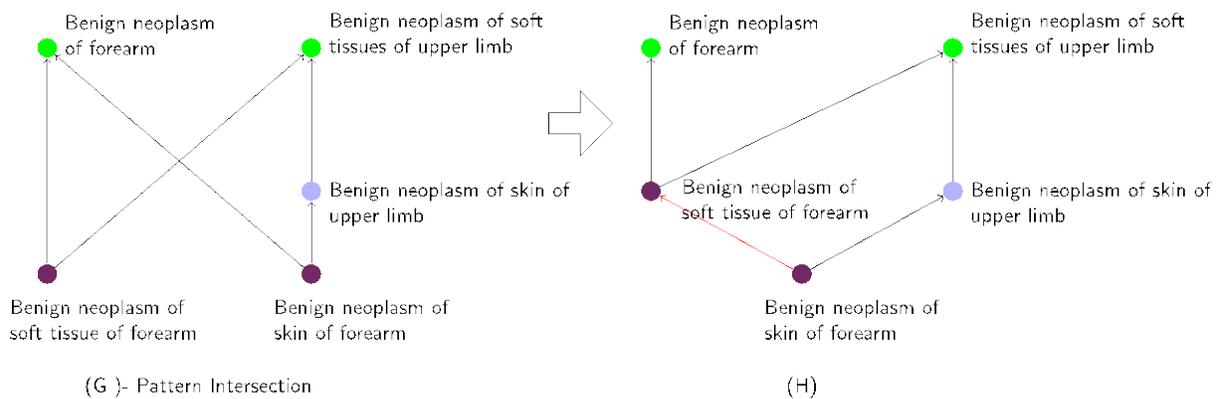
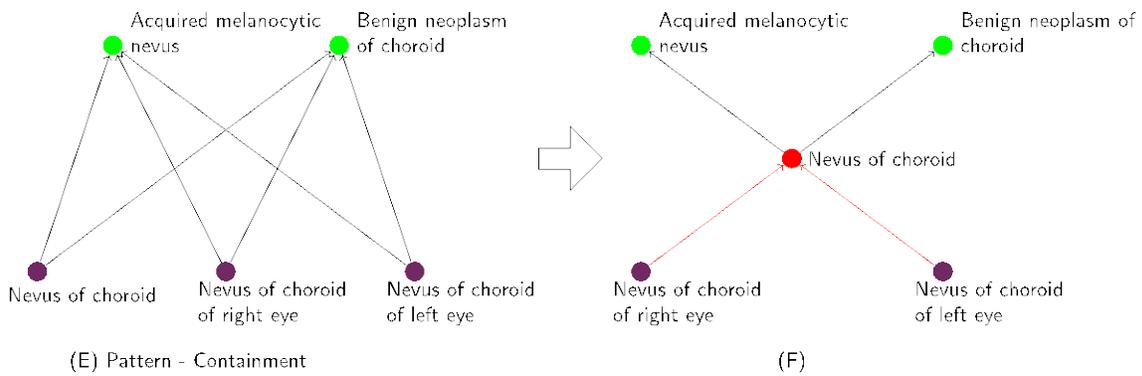
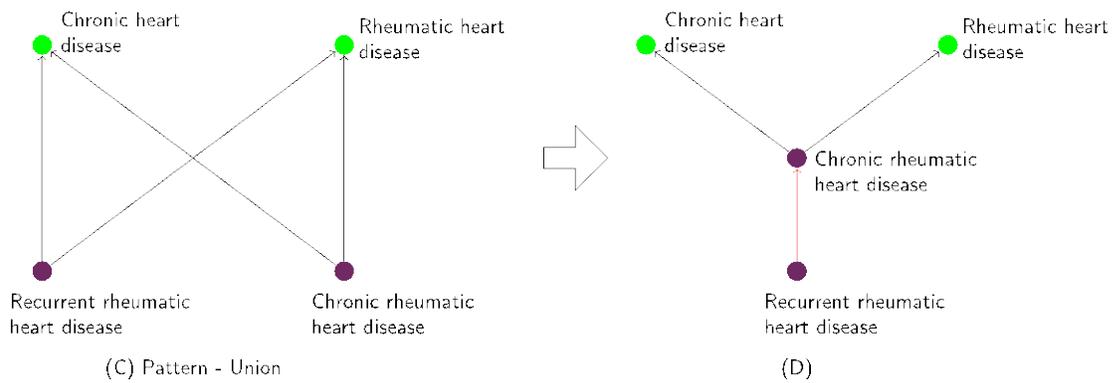
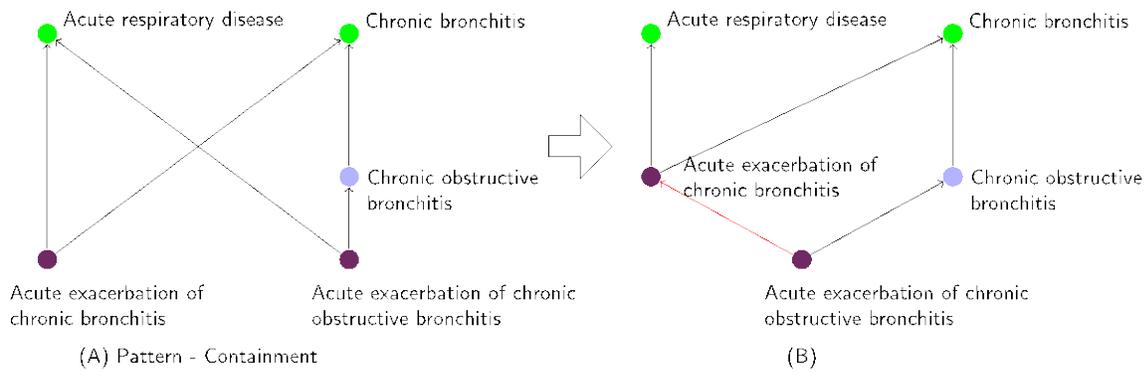


Figure 5: Examples of evaluated non-lattice subgraphs (left-hand side), as well as their remediation (right-hand side). The verified correction highlighted in red.

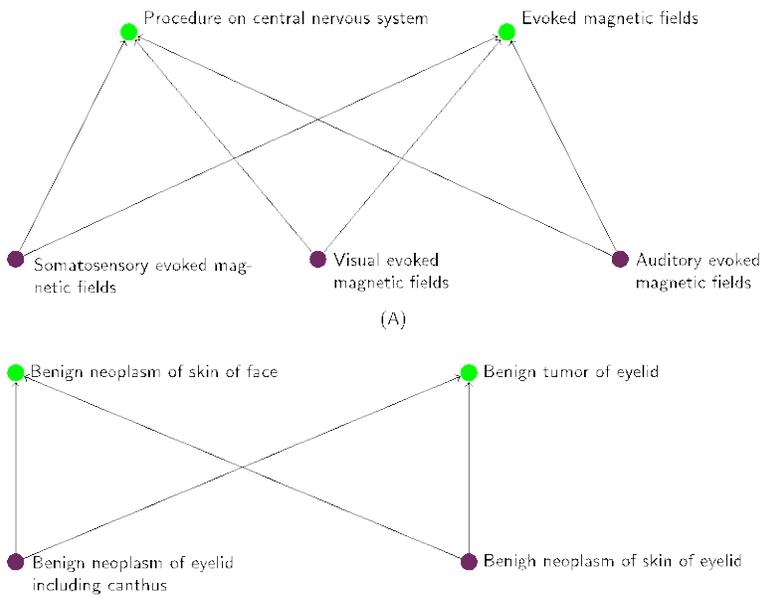


Figure 6: (A): An example of problematic non-lattice subgraph revealing modeling problems.

(B): Non-lattice subgraph pattern for which new lexical patterns would be required (e.g., leveraging synonymy between neoplasm and tumor). For each non-lattice subgraph, the upper-bound concepts are in green, and the lower-bound concepts are in purple.