# Leveraging Non-lattice Subgraphs to Audit Hierarchical Relations in NCI Thesaurus

**Rashmie Abeysinghe[1,2], Michael A. Brooks, MD[3], Licong Cui, PhD[1,⋆]**
**[1]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX**
**[2]Department of Computer Science, University of Kentucky, Lexington, KY**
**[3]Departments of Radiology and Medicine, University of Kentucky, Lexington, KY**

**Abstract**

*Auditing National Cancer Institute (NCI) thesaurus is essential to ensure that it provides accurate terminology for cancer-related clinical care as well as translational and basic research. We leverage a structural-lexical approach to identify missing hierarchical IS-A relations in NCI thesaurus based on non-lattice subgraphs and derived lexical attributes of concepts. For each concept in a non-lattice subgraph, we use two ways to derive the concept's lexical attributes: (1) inheriting lexical attributes from its ancestors within the subgraph; and (2) inheriting lexical attributes from all its ancestors. For a pair of concepts not having a hierarchical relation, if the lexical attributes of one concept is a subset of that of the other, we suggest there is a potential missing IS-A relation between the two concepts. Our approach identified 547 non-lattice subgraphs in the 19.01d release of NCI thesaurus which revealed a total of 1,022 unique potential missing IS-A relations. A random sample of 100 relations was evaluated by a domain expert. Among these relations, 90 can be obtained by the way of inheriting lexical attributes from ancestors within non-lattice subgraph, among which 76 were confirmed as valid (a precision of 84.44%); and 82 can be obtained by the way of inheriting all ancestors, among which 73 were confirmed as valid (a precision of 89.02%). The results show that our structural-lexical approach based on non-lattice subgraphs is effective for auditing NCI thesaurus.*

## 1 Introduction

The National Cancer Institute thesaurus (NCIt) is a biomedical terminology produced by NCI Enterprise Vocabulary Services containing more than 140,000 concepts[1]. Auditing biomedical terminologies such as NCIt is essential to ensure that it produces an accurate representation of the knowledge of the domain it models. This is especially important because the quality issues in terminologies would cause the applications that use these terminologies to be erroneous as well[2]. As the terminologies are continuously being expanded, their complexity also increases, making the introduction of errors almost unavoidable. Therefore, Terminology Quality Assurance (TQA) has become an important part of the terminological management lifecycle of all modern biomedical terminologies. However, manually reviewing a terminology to perform TQA is impractical due to the increasing size and complexity of modern terminologies. Therefore, automated or semiautomated approaches are needed to perform TQA efficiently and effectively.

In this paper, we introduce a structural-lexical approach based on non-lattice subgraph (NLS) to identify missing hierarchical relations in NCIt. Recently a number of studies have shown that analyzing lexical features in concept labels in NLSs is a promising way to identify different kinds of defects in biomedical terminologies[3–5]. In this work, we further work on this idea to come up with a set of lexical attributes for each concept in an NLS. We leverage these lexical attributes to suggest potential missing hierarchical relations in the NLS. A domain expert reviewed a randomly selected sample from the potential missing relations derived to examine the effectiveness of our approach.

## 2 Background

### 2.1 NCI Thesaurus (NCIt)

NCI Thesaurus is a biomedical terminology which covers vocabulary for cancer-related clinical care, translational and basic research, and public information and administrative activities[1]. It was originally created to facilitate interoperability and data sharing by various components of NCI by incorporating terms used by different components and mapping them to unique concepts[6]. Each concept includes a unique code, a preferred term, abbreviations, synonyms, and definitions[7]. The content of NCIt is organized in a description logic environment with more than 400,000 relations

⋆Corresponding author. Email: licong.cui@uth.tmc.edu

between concepts. NCIt is updated monthly with around 700 new concepts and many additional changes in each new release. It is released in many formats including Ontylog XML, OWL, and flat files. It is also available in both defined and inferred versions. We used the inferred version of the 19.01d release of NCIt in OWL format in this work.

### 2.2 Terminology Quality Assurance

Various methods have been investigated to facilitate quality improvement of biomedical terminologies like NCIt[8]. A type of high-level summary graphs, called abstraction networks, have been widely used to find inconsistencies in many biomedical terminologies[9–12]. Min et al.[13] have applied such an approach to the Biological Process hierarchy of NCIt which has led to the identification of different types of errors such as missing roles, missing concepts, incorrect hierarchical relations etc. Mougin et al.[14] have utilized the relations in UMLS semantic network to audit hierarchical and associative relations in NCIt. He et al.[15,16] have employed topological patterns that exists between NCIt and a reference terminology to import new concepts to NCIt. In previous work[17,18], we introduced a lexical-based inference approach to detect missing and incorrect relations in the Gene Ontology. Zheng et al.[19] have proposed a deep learning-based approach to predict concept names for new concepts that are added to SNOMED CT. Cui et al. investigated different ways of analyzing lexical features of concepts in NLSs to uncover missing hierarchical relations in SNOMED CT[3,5] and in NCIt[4].

### 2.3 Non-Lattice Subgraphs (NLSs)

Being a lattice is considered a desirable property for a well-formed terminology[20]. A terminology is a lattice if any pair of concepts has a unique maximal common descendant and a unique minimal common ancestor. Here, a common descendant $C$ is known as a maximal common descendant of a concept-pair ($A$, $B$), if $A$ and $B$ have no other common descendant $D$ such that $C$ is a descendant of $D$; and similarly, a common ancestor $R$ is known as a minimal common ancestor of a concept-pair ($P$, $Q$), if $P$ and $Q$ have no other common ancestor $S$ such that $R$ is an ancestor of $S$. If a concept pair has more than a single maximal common descendant (or a single minimal common ancestor), it is known as a non-lattice pair[20,21], which may reveal quality issues in terminologies.

It is not economical to separately examine multiple non-lattice pairs which share the same maximal common descendants. To address this, non-lattice subgraphs (NLSs) have been introduced[3]. An NLS can be acquired by a non-lattice pair $(c_1, c_2)$ as follows. Firstly, maximal common descendants of the non-lattice pair, $mcd(c_1, c_2)$, named as the lower bounds, is computed. Then, the minimal common ancestors of the lower bounds, $mca(mcd(c_1, c_2))$, named as the upper bounds, is computed. Finally, all the concepts as well as relations between (and including) lower and upper bounds is aggregated to generate the NLS. The size of an NLS is the number of concepts it contains. For example, in Figure 1, the non-lattice pair $\{1, 2\}$ (alternatively $\{1, 3\}$ or $\{2, 3\}$) yields $\{6, 7\}$ as its maximal common descendants. Reversely computing minimal common ancestors of $\{6, 7\}$ yields $\{1, 2, 3\}$. Then, the concepts $\{4, 5\}$ as well as relations between $\{1, 2, 3\}$ and $\{6, 7\}$ are aggregated to form the given NLS.

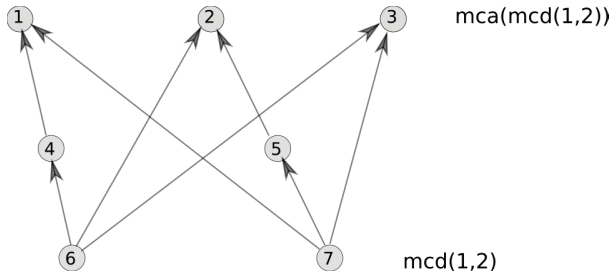



**Figure 1:** An example of an NLS. Nodes of the graph are concepts. The edges indicate hierarchical IS-A relations where the arrowheads point to the parent concept.

NLSs have been utilized to effectively identify defects in biomedical terminologies. Cui et al.[3] have proposed four lexical patterns found in NLSs which suggest missing hierarchical relations and missing concepts in SNOMED CT. In a previous work[4], we introduced two new lexical patterns applying that approach to NCIt. Cui et al.[5] originally introduced an approach combining NLSs and enriched lexical attributes of concepts to identify missing and incorrect

hierarchical relations in SNOMED CT. We also introduced a method to identify similar NLSs in the Gene Ontology to reduce the effort needed by domain experts in reviewing them[22].

## 3 Methods

We first extract all the NLSs in the 19.01d release of NCIt[23]. Then we construct the lexical attributes of concepts in NLSs by two ways: (1) inheriting lexical attributes from ancestors within NLSs; and (2) inheriting lexical attributes from all the ancestors. Based on the lexical attributes, we identify potential missing hierarchical relations between concepts. A random sample of missing relations is evaluated by a domain expert to verify their correctness.

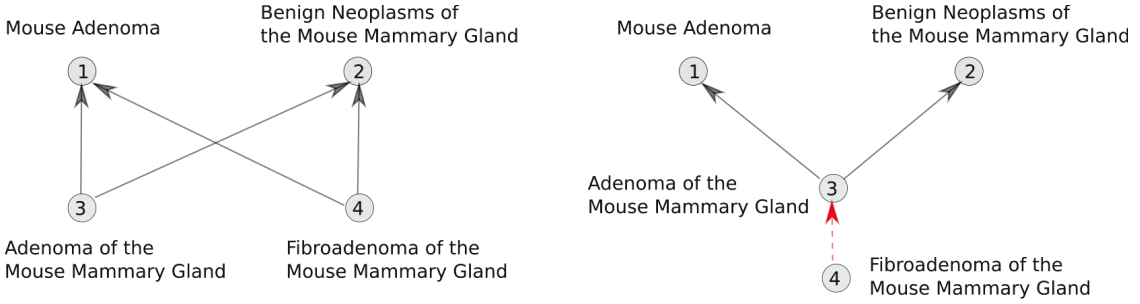
### 3.1 Constructing Lexical attributes of Concepts

Two lexical sources are leveraged to construct the set of lexical attributes for each concept in an NLS. Firstly we consider lexical attributes of the concept itself. The second source relies on the lexical attributes of the concept's ancestors. The second source is obtained in two ways.

1. *Inheriting lexical attributes from ancestors within the NLS*: In this way, we consider concept's ancestors that reside within the NLS to enrich the lexical attributes of a particular concept in the NLS. Note that we consider all the direct and indirect ancestors of a concept. Therefore, we compute the transitive closure of the hierarchical relation within the NLS to obtain indirect (transitive) ancestors.
2. *Inheriting lexical attributes from all the ancestors*: In this way, we consider all the concept's ancestors in the terminology without limiting to the NLS. To obtain indirect (transitive) ancestors, we compute the transitive closure of the hierarchical relation in the entire terminology.

We compare these two ways later in the paper in Section 4. Using these two sources we construct a set of lexical attributes $L_c$ for each concept $c$ in an NLS as follows.

- Load $L_c$ with the set of words contained in the preferred name of $c$.
- For each ancestor $a$ of $c$, add the set of words contained in the preferred name of $a$ to $L_c$. Note that $a$ could be an ancestor within the NLS or an ancestor external to the NLS depending on which way is used as discussed above.




**(A) Original NLS** **(B) Remedied NLS**

**Figure 2:** An NLS of size 4 and its remediation. The suggested remediation here is a missing hierarchical relation: "*C21663: Fibroadenoma of the Mouse Mammary Gland*" IS-A "*C21665: Adenoma of the Mouse Mammary Gland*". This can be obtained by both ways: inheriting lexical attributes from ancestors within the NLS and from all the ancestors.

We demonstrate the construction process using the NLS shown in Figure 2, considering ancestors within the NLS. For each concept $c$ in the NLS, we construct a set of attributes $L_{w_c}$ as follows. We initialize $L_{w_c}$ with the lexical attributes obtained from $c$'s preferred name:

$L_{w_1}$ = {mouse, adenoma}
$L_{w_2}$ = {benign, neoplasms, of, the, mouse, mammary, gland}
$L_{w_3}$ = {adenoma, of, the, mouse, mammary, gland}
$L_{w_4}$ = {fibroadenoma, of, the, mouse, mammary, gland}

If we enrich the above sets with the lexical attributes of the ancestors within the NLS, then the resulting attribute sets for each concept $c$ ($L_{w_c}$) are as follows (newly added attributes are underlined):

$L_{w_1}$ = {mouse, adenoma}
$L_{w_2}$ = {benign, neoplasms, of, the, mouse, mammary, gland}
$L_{w_3}$ = {adenoma, of, the, mouse, mammary, gland, benign, neoplasms}
$L_{w_4}$ = {fibroadenoma, of, the, mouse, mammary, gland, adenoma, benign, neoplasms}

If we use the lexical attributes of all the ancestors in the terminology, then the resulting attribute sets for each concept $c$ ($L_{a_c}$) are as follows:

$L_{a_1}$ = {mouse, adenoma, murine, organism, benign, epithelial, diagnosis, neoplasm, experimental, neoplasms, cell}
$L_{a_2}$ = {benign, neoplasms, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, experimental, murine, disorder, system, neoplasm}
$L_{a_3}$ = {adenoma, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, epithelial, experimental, cell, murine, disorder, system, benign, neoplasm, neoplasms}
$L_{a_4}$ = {fibroadenoma, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, epithelial, experimental, cell, murine, disorder, system, adenoma, benign, neoplasm, neoplasms}

### 3.2 Detecting Missing Relations

For a pair of concepts $c_1$ and $c_2$ in an NLS that are not connected by a hierarchical relation, if $c_2$'s lexical attributes $L_{c_2}$ is a proper subset of the $c_1$'s lexical attributes $L_{c_1}$, then we suggest $c_1$ IS-A $c_2$ (i.e. $c_1$ is the more specific concept). After obtaining all such potential missing relations in an NLS, we remove redundant relations that can be inferred by others. For example, if we suggest $a$ IS-A $b$ and $a$ IS-A $c$ for a particular NLS where $b$ IS-A $c$ already exists in the NLS, then we consider $a$ IS-A $c$ as redundant, since it can be inferred transitively through $a$ IS-A $b$ and $b$ IS-A $c$. Therefore, we remove $a$ IS-A $c$ from the list of suggestions.

For instance, considering ancestors within NLS, for concepts 3 and 4 in Figure 2, $L_{w_3}$ = {adenoma, of, the, mouse, mammary, gland, benign, neoplasms} is a proper subset of $L_{w_4}$ = {fibroadenoma, of, the, mouse, mammary, gland, adenoma, benign, neoplasms}. Also, considering all the ancestors, $L_{a_3}$ = {adenoma, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, epithelial, experimental, cell, murine, disorder, system, benign, neoplasm, neoplasms} is a proper subset of $L_{a_4}$ = {fibroadenoma, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, epithelial, experimental, cell, murine, disorder, system, adenoma, benign, neoplasm, neoplasms}.

Hence, we suggest concept 4 should be more specific than 3, i.e. *Fibroadenoma of the Mouse Mammary Gland* IS-A *Adenoma of the Mouse Mammary Gland*. As discussed above, this can be obtained by both considering ancestors within the NLS and all the ancestors.

Figure 3 contains a size-7 NLS with a potential missing hierarchical relation: "*C4887: Metastatic Malignant Neoplasm in the Trachea*" IS-A "*C4571: Malignant Respiratory Tract Neoplasm*" which can be obtained only by considering ancestors within the NLS for constructing lexical attributes.

Figure 4 contains a size-11 NLS with a potential missing hierarchical relation: "*C5270: Cerebellar Papillary Meningioma*" IS-A "*C3569: Malignant Cerebellar Neoplasm*" which can be only obtained by considering all the ancestors for constructing lexical attributes. This is because inheriting lexical attributes from ancestors within the NLS yields $L_{w_{11}}$ = {malignant, cerebellar, neoplasm, infratentorial, brain, intracranial, central, nervous, system} which is not a proper subset of $L_{w_{10}}$= {cerebellar, papillary, meningioma, grade, iii, malignant, neoplasm}. However, when all the ancestors are considered, $L_{a_{11}}$ = {malignant, cerebellar, neoplasm, disorder, central, system, nervous, infratentorial, intracranial, brain} is a subset of $L_{a_{10}}$ = {cerebellar, papillary, meningioma, infratentorial, intracranial, brain, cell, malignant, disorder, system, central, meningeal, nervous, grade, iii, neoplasm, meningothelial}.

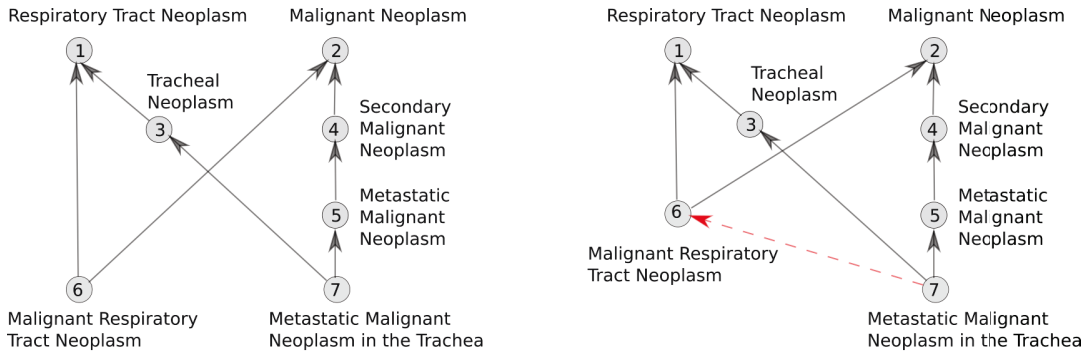



**Figure 3:** An NLS of size 6 and its remediation. The suggested remediation here is a missing hierarchical relation: "*C4887: Metastatic Malignant Neoplasm in the Trachea*" IS-A "*C4571: Malignant Respiratory Tract Neoplasm*". This can only be obtained by considering the ancestors within the NLS for enriching lexical attributes.

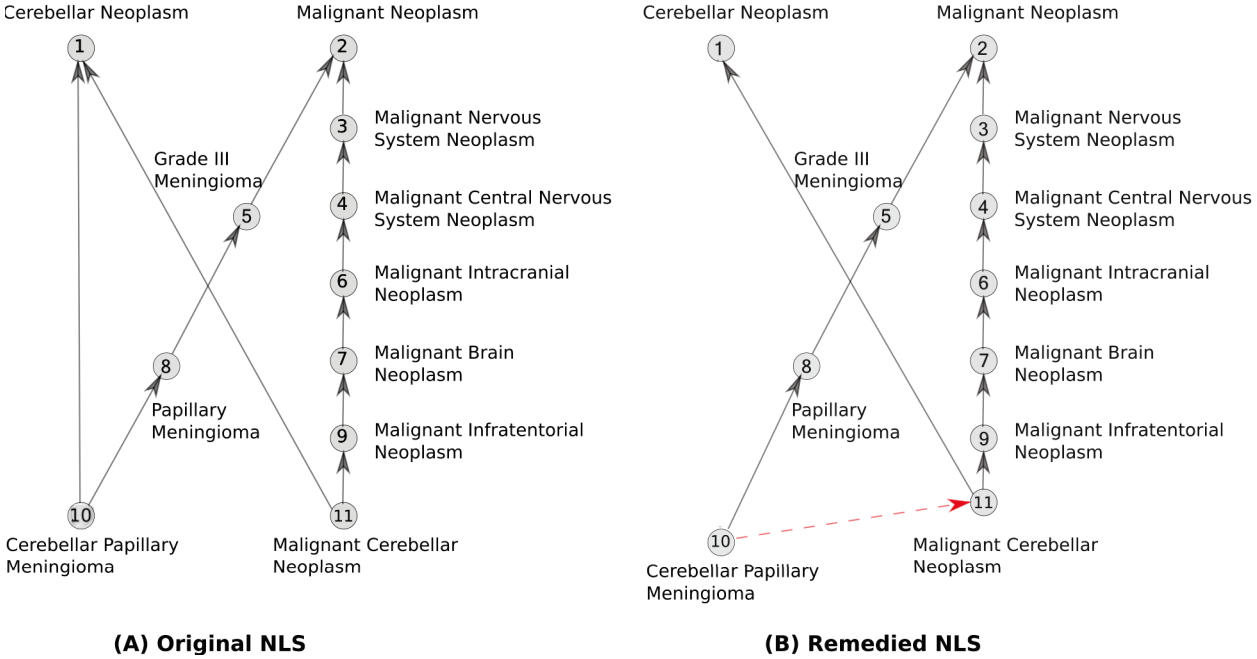



**Figure 4:** An NLS of size 11 and its remediation. The suggested remediation here is a missing hierarchical relation: "*C5270: Cerebellar Papillary Meningioma*" IS-A "*C3569: Malignant Cerebellar Neoplasm*". This can only be obtained by considering all ancestors for enriching lexical attributes.

### 3.3 Filtering

We perform three kinds of filtering to avoid generate erroneous suggestions of potential missing IS-A relations: stop word filtering, antonym filtering, and position filtering.

**Stop word filtering**. Consider the concepts "*C4013: Malignant Head and Neck Neoplasm*" and "*C3260: Neck Neoplasm*". These two satisfy all the requirements to be candidates for a suggestion of a missing hierarchical relations in the form of "*C4013: Malignant Head and Neck Neoplasm*" IS-A "*C3260: Neck Neoplasm*". However, upon close

observation, it can be seen that this suggestion is wrong since it gives the idea of *Head Neoplasm* is a subtype of *Neck Neoplasm*. Existence of such stop words in concepts make them more prone to generate erroneous missing hierarchical relation suggestions. Therefore, If a concept contains such stop words, we do not make any suggestions. Moreover, we also do not consider such concepts to enrich lexical attributes of other concepts. That is, if a concept with stop words exists as an ancestor of another concept, we do not enrich the lexical attributes of the latter with the former. The stop words used to perform this filtering are: "and", "and/or", "or", "no", "not", "without", "due to", "secondary to", "except", "by", "after", "able", "removal", "replacement", "NOS", where "NOS" represents "Not Otherwise Specified".

**Antonym filtering.** If the constructed enriched lexical attributes of a particular concept contains an antonym pair, such concepts are more prone to erroneous suggestions as well. For example, consider the concepts "*C60996: Malignant Epithelial Small Polygonal Cell*" with attributes {*small, cytoplasm, with, large, abundant, polygonal, epithelial, neoplastic, cell, malignant*} and "*C36822: Malignant Epithelial Large Cell*" with attributes {*large, epithelial, neoplastic, cell, malignant*}. Even though attributes of *C60996* is a proper subset of *C36822*, suggesting a hierarchical relation between these two is obviously not accurate since *C60996* is discussing small cells and *C36822* is discussing large cells (note that "small" and "large" is an antonym pair). Therefore, after obtaining the set of attributes, we check the set to ensure that it does not contain an antonym pair. The antonym pairs are obtained from WordNet[24].

**Position filtering**. For concepts with short names, they may appear as a part of other concepts' names in various positions (e.g., beginning, middle, or end). For concepts whose names are not appearing at the end of other concepts' names, it is likely to suggest incorrect missing IS-A relations. For instance, concept "*Fentanyl*" appears at the beginning of concept "*Fentanyl Citrate Pectin-Based Nasal Spray*", and the subset inclusion may wrongly suggest "*Fentanyl Citrate Pectin-Based Nasal Spray*" IS-A "*C494:Fentanyl*". Therefore, we filter out such cases by assigning a constraint such that the shorter concept should always appear at the end of the the longer concept.

### 3.4 Evaluation

To evaluate the performance of our approach in accurately identifying missing hierarchical relations, we randomly selected a sample of missing hierarchical relations from the overall results for evaluation. These samples were provided to a domain expert (author MAB). Existing erroneous hierarchical relations in NCIt may help derive incorrect suggestions for missing hierarchical relations. Therefore, for the potential missing relations identified as incorrect by the domain expert, in a second round of evaluation, we provided the domain expert with existing hierarchical relations that were used to derive the incorrect ones. If the domain expert disagrees with the existing relation as well, then we marked it as an incorrect existing hierarchical relation. For instance, the NLS in Figure 5 denotes such a scenario. "*C3779: Giant Cell Carcinoma*" should not be a subtype of "*C3780: Large Cell Carcinoma*". The existence of this relation derives the incorrect suggestion of "*C4452: Lung Giant Cell Carcinoma*" IS-A "*C4450: Lung Large Cell Carcinoma*".

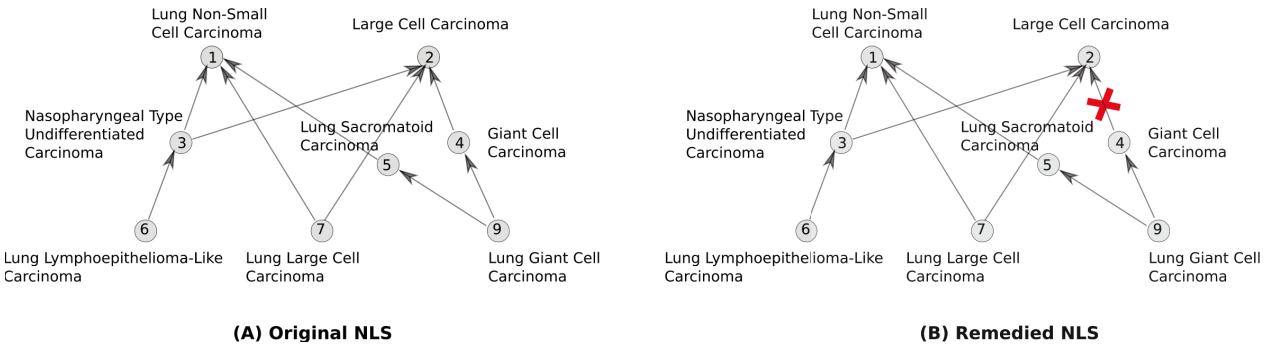



**Figure 5:** An NLS of size 8 and its remediation. The suggested remediation here is an incorrect hierarchical relation: *C3779: Giant Cell Carcinoma* should not be a subtype under *C3780: Large Cell Carcinoma*. This can be obtained by both considering ancestors within the NLS and all ancestors for enriching lexical attributes.

## 4 Results

### 4.1 Summary Results

A total of 9,512 NLSs were extracted from the 19.01d version of NCIt with sizes ranging from 4 to 644. Out of these, our approach identified 547 NLSs with potential missing hierarchical relations. These NLSs contained a total of 1,022 potential missing hierarchical relations (note that an NLS may contain more than one missing hierarchical relation). It can be seen from Table 1 that 441 out of 547 NLSs can be identified by the way of inheriting lexical attributes from ancestors within NLS and suggests 925 potential missing IS-A relations; and 422 out of 547 can be identified by the way of inheriting lexical attributes from all the ancestors and suggests 847 potential missing IS-A relations. The two ways identified 750 potential missing IS-A relations in common.

**Table 1:** The number of NLSs and the number of potential missing hierarchical relations suggested in those NLSs.

| **Type** | **# of NLSs** | **# of potential missing IS-A** |
|---|---|---|
| Inheriting lexical attributes from ancestors within NLS | 441 | 925 |
| Inheriting lexical attributes from all the ancestors | 422 | 847 |

### 4.2 Evaluation

The evaluation sample contained 100 potential missing relations observed in 83 NLSs identified by our approach. The domain expert concluded 85 (85%) of missing hierarchical relations are valid. Table 2 shows 15 examples of valid missing hierarchical relations in the form of subconcept and superconcept, for instance, "*C7155: Primary Central Chondrosarcoma*" IS-A "*C3737: Mesenchymal Chondrosarcoma*". For the 15 invalid ones, the domain expert further inspected the existing hierarchical relations that were used to derive the invalid ones and verified that 8 of them were actually incorrect. Table 3 lists five examples of incorrect existing relations. For example, "*C66775: Borderline Ovarian Mucinous Adenofibroma*" should not be a subtype of "*C4934: Benign Female Reproductive System Neoplasm*", since the word "borderline" indicates that it is on the borderline between benign and malignant, and may exhibit malignant behavior.

We summarize the evaluation result in Table 4 according to the two ways of inheriting lexical attributes. Among 100 NLSs, 90 were identified by the way of inheriting lexical attributes from ancestors within NLS suggesting 76 correct missing IS-A relations (a precision of 84.44%); and 82 were identified by the way of inheriting lexical attributes from all the ancestors suggesting 73 correct missing IS-A relations (a precision of 89.02%).

## 5 Discussion

This paper presents a structural-lexical approach to audit NCIt based on enriched lexical attributes of concepts in NLSs. The results indicate that most missing IS-A relations can be commonly obtained by considering ancestors within the NLSs and all the ancestors to enrich the lexical attributes. The former way identified more potential missing IS-A relations than the the latter did, while the latter achieved a better precision than the former did.

### 5.1 Analysis of Failure Cases

The primary focus of this work was to identify missing hierarchical relations in NCIt. Upon observation of the false positives, it could be noted that a majority of them (53%) occur due to the existing erroneous hierarchical relations in NCIt. For example, in Figure 5, *Giant Cell Carcinoma* is categorized as a subtype of *Large Cell Carcinoma*. However, under the current (2015) WHO classification, Giant Cell Carcinomas are classified as a separate category of tumor. Therefore, *Giant Cell Carcinoma* should not be a subtype of *Large Cell Carcinoma*. Likewise in a separate case, our approach inaccurately identified "*C39951: Testicular Fibroma*" IS-A "*C4092: Benign Epithelial Neoplasm*" as a missing relation. However, it could be seen that this was obtained due to the erroneous existing relation "*C39951: Testicular Fibroma*" IS-A "*C3709: Epithelial Neoplasm*", since a *Testicular Fibroma* does not arise from *Testicular*

**Table 2:** Fifteen examples of valid missing hierarchical relations obtained by our approach.

| **Subconcept** | **Superconcept** |
|---|---|
| C7155: Primary Central Chondrosarcoma | C3737: Mesenchymal Chondrosarcoma |
| C5270: Cerebellar Papillary Meningioma | C3569: Malignant Cerebellar Neoplasm |
| C6430: Thymic Carcinoid Tumor | C3773: Neuroendocrine Carcinoma |
| C133894: Stage 0 Small Intestinal Adenocarcinoma AJCC v8 | C7657: Intestinal Precancerous Condition |
| C39863: Adenocarcinoma of Skene Gland Origin | C6167: Urethral Adenocarcinoma |
| C15385: Excisional Biopsy | C64979: Diagnostic Surgical Procedure |
| C61145: Adenocarcinoma Cell with Eosinophilic Cytoplasm | C53644: Malignant Cell with Eosinophilic Cytoplasm |
| C121571: Leiomyosarcoma of Deep Soft Tissue | C9306: Soft Tissue Sarcoma |
| C6591: Peripheral Neuroblastoma | C4961: Malignant Peripheral Nervous System Neoplasm |
| C64000: Tubulostromal Adenoma of the Rat Ovary | C134942: Rat Neoplasms |
| C3758: Hepatocellular Adenoma | C36207: Digestive System Adenoma |
| C40116: Fallopian Tube Metaplastic Papillary Tumor | C8429: Papillary Epithelial Neoplasm |
| C8961: Fundic Gland Polyp | C4092: Benign Epithelial Neoplasm |
| C9374: Adult Brain Meningioma | C7710: Adult Brain Neoplasm |
| C27404: Childhood Central Nervous System Mature Teratoma | C5591: Benign Childhood Central Nervous System Neoplasm |

**Table 3:** Five examples of incorrect existing hierarchical relations obtained by our approach.

| **Subconcept** | **Superconcept** |
|---|---|
| C66775: Borderline Ovarian Mucinous Adenofibroma | C4934: Benign Female Reproductive System Neoplasm |
| C33149: Muscularis Mucosa | C32209: Bladder Tissue |
| C4826: Central Nervous System Neuroblastoma | C3568: Malignant Brain Neoplasm |
| C38157: Metachronous Osteosarcoma | C4968: Secondary Malignant Neoplasm |
| C39951: Testicular Fibroma | C3709: Epithelial Neoplasm |

**Table 4:** The precision of our approach in two ways to identify missing hierarchical relations based on the evaluation performed by the domain expert.

| **Type** | **# of suggested missing IS-As** | **# of correct suggestions** | **Precision** |
|---|---|---|---|
| Inheriting lexical attributes from ancestors within NLS | 90 | 76 | 84.44% |
| Inheriting lexical attributes from all the ancestors | 82 | 73 | 89.02% |

*Epithelium*, but from the *Stroma*.

Next we give an example of the false positive cases which are not due to the existing erroneous hierarchical relations in NCIt. Our method suggests "*C115093: Recurrent Oropharyngeal Undifferentiated Carcinoma*" as a subtype of "*C9268: Recurrent Malignant Nasopharyngeal Neoplasm*" since it inherits lexical attribute "*malignant*"

from an ancestor "*C150531: Recurrent Malignant Pharyngeal Neoplasm*" and inherits lexical attribute "*nasopharyngeal*" from another ancestor "*C4107: Nasopharyngeal Type Undifferentiated Carcinoma*". However, "*C4107: Nasopharyngeal Type Undifferentiated Carcinoma*" indicates that it looks like nasopharyngeal carcinoma under the microscope, but is not a nasopharyngeal carcinoma. Oropharyngeal carcinoma and nasopharyngeal carcinoma behave differently biologically, with nasopharyngeal carcinoma having a worse prognosis, and they are caused by different types of virus (HPV in oropharyngeal carcinoma, and EBV in nasopharyngeal carcinoma). Therefore, our suggestion is incorrect since our approach is incapable of capturing the subtle difference between "*nasopharyngeal*" and "*nasopharyngeal type*".

### 5.2 Comparison with Previous Work

In our previous work[4], we used six lexical patterns in NLSs to identify missing hierarchical relations in NCIt. One of the patterns was "Containment", where we suggested hierarchical relations if the set of words of a concept is a subset of another. The "Containment" pattern was restricted to lower and upper bounds of the NLS while this work we have no such restriction. Also, we only considered the lexical attributes of the preferred term, while in this work we also enrich it with the lexical attributes of the ancestor terms. Furthermore, we perform three filtering steps to avoid obtaining incorrect suggestions.

The structural-lexical approach based on enriched lexical attributes was first introduced by Cui et al.[5] to audit SNOMED CT. While our approach is similar to theirs, we perform a number of additional steps to improve performance and coverage. We do not skip considering an entire NLS if it contains stop words or antonym pairs as was done previously[5]. Rather, we perform a much fine-grained filtering by considering stop words and antonym pairs at the concept level, not the NLS level. Additionally, we also address an issue mentioned in Cui et al.'s work regarding incorrect suggestions when the set of words of a concept is a subset of another concept's set of words. More importantly, in this work we introduce another way to enrich the lexical attributes of a concept: by considering all its ancestors (not only the ancestors within the NLS). This way was actually found to have a higher precision. Moreover, we do not put any restriction on the sizes of NLSs for evaluation in this work, while the evaluation was limited to small (size 4,5, and 6) NLSs in the previous work.

### 5.3 Limitations and Future Work

While achieving a higher precision in suggesting missing hierarchical relations, our approach only covers a small portion of NLSs in NCIt (547 out of 9,512). New lexical patterns need to be identified to suggest remediations for the remaining unsolved NLSs. In this work, we enriched the lexical attributes of a concept by its ancestor lexical attributes. We expect to investigate into other methods that can be used for enriching, such as synonyms, definitions and other attribute relations. Another limitation of this work is that only one domain expert was involved in the evaluation. We plan to perform future evaluations by multiple domain experts to increase the robustness of the evaluation. In addition, although some of the failure suggestions of missing IS-A relations further revealed incorrect existing relations, it depended on the domain expert's manual review. It would be desirable to develop automated methods to detect incorrect existing relations.

## 6 Conclusion

In this paper, we applied a structural-lexical auditing approach based on enriched lexical attributes of concepts in non-lattice subgraphs to suggest potential missing hierarchical relations in the National Cancer Institute thesaurus. This approach achieved a precision of 84.44% by inheriting lexical attributes from ancestors with NLSs, and a precision of 89.02% by inheriting lexical attributes from all the ancestors in the entire terminology, indicating the effectiveness of our approach. This approach could be generally applied to any biomedical terminology for quality assurance purposes.

### Acknowledgment

This work was supported by the National Science Foundation (NSF) through grants IIS-1816805 and IIS-1931134, and the National Institutes of Health (NIH) National Cancer Institute through grant R21CA231904. The content is

solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH.

## References


1. NCI (NCI Thesaurus) - Synopsis [Internet]. 2019 [cited 4 March 2019]. Available from: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NCI/
2. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. Journal of the American Medical Informatics Association. 2013;21(e1):e11-9.
3. Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. Journal of the American Medical Informatics Association. 2017;24(4):788-98.
4. Abeysinghe R, Brooks MA, Talbert J, Cui L. Quality assurance of NCI thesaurus by mining structural-lexical patterns. In AMIA Annual Symposium Proceedings. 2017:364-73.
5. Cui L, Bodenreider O, Shi J, Zhang GQ. Auditing SNOMED CT hierarchical relations based on lexical features of concepts in non-lattice subgraphs. Journal of biomedical informatics. 2018;78:177-84.
6. Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization of the NCI thesaurus. International Journal of Genomics. 2004;5(8):648-54.
7. Overview of NCI Thesaurus (NCIt) [Internet]. 2019 [cited 5 March 2019]. Available from: https://wiki.nci.nih.gov/pages/viewpage.action?pageId=7472532
8. Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. Journal of Biomedical Informatics. 2009;42(3):413-25.
9. Ochs C, Geller J, Perl Y, Chen Y, Agrawal A, Case JT, Hripcsak G. A tribal abstraction network for SNOMED CT target hierarchies without attribute relationships. Journal of the American Medical Informatics Association. 2014;22(3):628-39.
10. Ochs C, Geller J, Perl Y, Chen Y, Xu J, Min H, Case JT, Wei Z. Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies. Journal of the American Medical Informatics Association. 2015;22(3):507-18.
11. Ochs C, Perl Y, Halper M, Geller J, Lomax J. Quality assurance of the gene ontology using abstraction networks. Journal of Bioinformatics and Computational Biology. 2016;14(03):1642001.
12. Wei D, Bodenreider O. Using the abstraction network in complement to description logics for quality assurance in biomedical terminologies-a case study in SNOMED CT. Studies in Health Technology and Informatics. 2010;160(P2):1070-4.
13. Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. Journal of the American Medical Informatics Association. 2006;13(6):676-90.
14. Mougin F, Bodenreider O. Auditing the NCI thesaurus with semantic web technologies. In AMIA Annual Symposium Proceedings. 2008:500-504.
15. He Z, Geller J. Preliminary analysis of difficulty of importing pattern-based concepts into the National Cancer Institute thesaurus. Studies in health technology and informatics. 2016;228:389.
16. He Z, Chen Y, de Coronado S, Piskorski K, Geller J. Topological-pattern-based recommendation of UMLS concepts for National Cancer Institute thesaurus. In AMIA Annual Symposium Proceedings. 2016:618-627.
17. Abeysinghe R, Hinderer EW, Moseley HN, Cui L. Auditing subtype inconsistencies among gene ontology concepts. In 2017 IEEE International Conference on Bioinformatics and Biomedicine. 2017:1242-1245.
18. Abeysinghe R, Zheng F, Hinderer EW, Moseley HN, Cui L. A lexical approach to identifying subtype inconsistencies in biomedical terminologies. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2018:1982-1989.
19. Zheng F, Cui L. Exploring deep learning-based approaches for predicting concept names in SNOMED CT. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2018:808-813.
20. Zhang GQ, Bodenreider O . Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. In AMIA Annual Symposium Proceedings. 2010:922-6.
21. 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.
22. Abeysinghe R, Qu X, Cui L. Identifying similar non-lattice subgraphs in Gene Ontology based on structural isomorphism and semantic similarity of concept labels. In AMIA Annual Symposium Proceedings. 2018:1186-1195.
23. Zhang GQ, Xing G, Cui L. An efficient, large-scale, non-lattice-detection algorithm for exhaustive structural auditing of biomedical ontologies. Journal of biomedical informatics. 2018;80:106-19.
24. Miller GA. WordNet: a lexical database for English. Communications of the ACM. 1995;38(11):39-41.