# The UMLS Coverage of Clinical Radiology

Carol Friedman Ph.D.

Queens College of the City University of New York
Columbia University, New York

## Abstract

*The informational content of clinical radiology reports was examined to determine the coverage of the Unified Medical Language System (UMLS) in relation to the terminology used by physicians in the Radiology Department of Columbia Presbyterian Medical Center (CPMC). The UMLS semantic network contained 17 semantic types which were compatible with the types of clinical information in the reports. The type of semantic categories missing from the UMLS consisted mainly of modifier information relating to certainty, degree, and change type of information. This type of information formed a substantial part of the domain. Although most of the informational categories were found in the UMLS semantic network, most of the domain terms were not. Our results strongly suggest that the UMLS could be a significant tool for developing clinical text processing applications if it were extended to cover clinical domains.*

## 1 Introduction

The Unified Medical Language System (UMLS) was established by the National Library of Medicine [7] to further the development of automated biomedical information systems. One type of medical information system deals with the extraction of relevant clinical information from narrative reports [11, 1, 5]. The salient clinical information is transformed into a structured form containing controlled vocabulary terms, making the clinical data accessible for further computerized applications, such as automated quality assurance, clinical decision support, and biomedical research. In order for the information to be transformed, there must be a controlled vocabulary to represent the clinical concepts, and the words and phrases found in the domain must be semantically categorized and assigned corresponding target terms from the controlled vocabulary.

The lack of an adequate standardized biomedical vocabulary impedes the development of such systems because a controlled vocabulary and a system of semantic classification has to be developed anew for each domain. The UMLS offers the potential of alleviating this bottleneck because it is a knowledge source of biomedical terms which are already semantically categorized. The UMLS could be used to build the vocabulary of the domain, to obtain semantic categories for the terms of the vocabulary, and to associate phrases of the domain with controlled vocabulary terms. The effectiveness then of the UMLS for clinical text processing applications depends on adequate coverage of the clinical domain.

Other studies have reported on the utility of the UMLS in clinical applications [6, 3], and on the utility of the UMLS for text processing systems [8, 2]. In this paper, the coverage of the UMLS in relation to clinical information found in the domain of radiology reports is presented, and an evaluation of the utility of the UMLS in the development of a text processor for clinical radiology is also discussed. The type of reports studied consist of the impression section of chest x-rays because they cover a broad array of clinical information and they are readily available in electronic form as part of the Clinical Information System at Columbia Presbyterian Medical Center (CPMC).

## 2 Background

Building a text processor is an inherently difficult task. In order to understand the information in text, humans utilize a broad array of general knowledge, which includes knowledge concerned with the syntax and semantics of the underlying language, in addition to general knowledge about the world and the specialized domain of the text. Several text processing systems have been developed, encompassing varying degrees of knowledge. The methodologies utilize pattern matching [10, 5], conceptual and/or semantic modelling [1, 9, 12], and comprehensive natural language processing [11, 4].

We are developing a natural language system which incorporates biomedical semantic knowledge along with general syntactic knowledge of English. The semantic knowledge component consists of semantic classifications for clinically relevant words and terms, their corresponding controlled vocabulary terms, rules specifying well defined semantic patterns found in the text, and the corresponding semantic interpretations of the patterns. This paper focuses on a discussion of the semantic aspects of the processor that are relevant to the UMLS knowledge sources. In order to evaluate the clin-

0195-4210/92/$5.00 ©1993 AMIA, Inc.

ical coverage of radiology in the UMLS, a description of the informational content of the domain is presented. This description was obtained by manually analyzing texts of the domain.

The present study consists of the impression section of 600 randomly chosen chest x-ray reports containing a total of 1387 sentences which encompass a vocabulary of 918 distinct terms, 516 of which were deemed relevant to this study because they contain clinically salient information; the terms considered irrelevant represent general English terminology. A distinct term consists of either one word or a phrase of several words where phrase cannot be decomposed into a sequence of individual words without losing the underlying meaning. For example *sickle cell disease* is a unique term, whereas *mild cardiomegaly* consists of two terms *mild*, and *cardiomegaly*.

# 3 Types of Information

A preliminary manual analysis of the informational content of the text was performed by the author and a physician from the Radiology Department of CPMC. The different types of information in the text were initially grouped into broad informational units as follows: 1)descriptions and interpretation of the findings - *multiple opacities, interstitial markings* 2) previous therapeutic procedures or medical devices seen on or inferred from the x-ray - *mastectomy, groshong catheter* 3)comparisons made between the current examination and a previous one - *heart appears larger, unchanged since previous exam* 4)information concerning technique - *poor inspiration, hazy film* 5)patient management issues - *follow up suggested, clinical correlation recommended.*

Only the first three types of information above were included in the present study because these form the bulk of the clinical information. In order to differentiate between our semantic categories and those of the UMLS semantic network, we created our own general semantic category covering the first three types of information called Rad-exam-findings.

On a subsequent, more detailed manual analysis of the informational content of the Rad-exam-findings, the semantic categories associated with them were divided into finer informational units which generally correspond to individual words of the reports. For example, *mild pleural effusion* consists of a degree type word *mild*, a body region type of word *pleura*, and a Rad-finding type of word *effusion*.

Each Rad-exam-finding basically consists of a Rad-finding and modifiers. The Rad-findings are equivalent to the informational types shown above exclusive of the modifiers. The modifiers consist of the following types of information: 1)**certainty**: *probable, possible, no*, 2)**degree**: *mild, extensive, severe*, 3)**change**: *improved, increased*, 4)**status**: *active, acute*, 5)**type**: *focal* as in *focal infiltrate*, and *carcinoid* as in *carcinoid tumor*, 6)**body part**: *lung, aorta*, 7)**body part region**: *left lower lobe, hemidiaphragm.*

# 4 UMLS Semantic Coverage

The semantic types in the UMLS semantic network were manually examined to determine which types were relevant to the clinical radiology domain, and whether there were any informational types in the domain that were not represented in the semantic network. There are 12 semantic types in the UMLS semantic network that are relevant to the clinical findings we call Rad-findings. They are **Finding** *(effusion)*, **Sign** *(adenopathy)*, **Organ and Tissue Function** *(aeration)*, **Pathologic Function** *(cardiomegaly)*, **Disease or Syndrome** *(pneumothorax)*, **Injury or Poisoning** (*fracture)*, **Therapeutic or Preventative Procedure** *(mastectomy)*, **Medical Device** *(metal clips)*, **Qualitative Concept** *(normal sized heart)*, **Quantitative Concept** *(4 cm mass)*, **Virus** *(viral pneumonia)*, **Congenital Abnormality** *(opacity)*, and **Acquired Abnormality** *(scarring)*. In addition, there are several types which are relevant to the modifiers of a Rad-finding; these are related to body parts or body part regions, and to terms corresponding to temporal concepts such as *postoperative examination.* These types are **Fully Formed Anatomical Structure, Body part, Organ, or Organ Component** *(heart)*, **Tissue** *(scar tissue)*, **Body Location or Region** *(chest)*, **Body Space or Junction** *(interstitial)*, and **Temporal Concept** (*postoperative interval*). The radiology examination itself is classifiable as a **Diagnostic Procedure.**

All the Rad-finding terms were covered by the UMLS semantic types, but the modifier information was generally not covered. This is not surprising since the UMLS was primarily designed to consist of complete medical concepts. The modifiers operate on concepts and often substantially change the underlying meanings, but they are basically not complete concepts by themselves.

There are appropriate places in the network where the modifier concepts can be handled. The type **Qualitative Concept** is defined as "A concept which involves primarily a judgment, rather than a direct measurement". This type can be the parent of new types needed to cover modifier information, if an extension is warranted. The new types could be **Degree Concept**, **Certainty Concept**, **Status Concept** and **Change Concept**. Another new type **Bodypart region** is also needed, which could be a child of **Anatomical Structure**. If the UMLS is to be applicable to the clinical domain, the semantic network would have to be extended to represent new types of information, and the appropriate concepts would have to be included in Meta-1, because these types of information form a substantial portion of the clinical domain.

In order to determine the amount and frequency of the different types of information found in clinical ra-

| *Type of Information* | *Number of Terms* | % | *Frequency* | % |
|---|---|---|---|---|
| Rad-findings | 190 | 37 | 1210 | 26 |
| Bodypart | 48 | 9 | 755 | 16 |
| Bodypart region | 115 | 22 | 794 | 17 |
| Status | 21 | 4 | 347 | 7 |
| Degree | 24 | 5 | 396 | 8 |
| Certainty | 42 | 8 | 909 | 19 |
| Change | 29 | 6 | 255 | 5 |

Table 1: **Types of Information in Clinical Radiology**

diology, a statistical analysis of the radiology reports was performed to obtain the number of words or terms which corresponded to each type of informational category, and to determine the frequency of their occurrence. The entire body of text contained 4,666 occurrences of relevant terms. Table 1 shows the results of the study. The first column contains the type of semantic information, the second column contains the number of unique terms for that category, and the third column contains the percentage of the unique terms. The total number of unique clinically relevant terms (516) was used to obtain the percentage, although only the most significant informational categories were studied. The fourth column contains the total number of occurrences of the corresponding informational category, and the fifth column contains its frequency. The number of medically relevant occurrences (4,666) was used to compute the percentage for the frequency. The four modifier categories **Degree, Certainty, Status**, and **Change** together account for 23% of the vocabulary and occur in the reports with a frequency of 39%. The **Bodypart** and **Bodypart region** categories together account for 31% of the vocabulary and occur with a frequency of 33%. The Rad-findings account for 39% of the vocabulary and occur with a frequency of 26%. This study indicates that modifier information forms a substantial portion (54%) of the clinical vocabulary in radiology texts.

Since this study was performed for clinical radiology only, we can only surmise that these types of modifier information are found in other clinical domains as well. These types are equivalent to modifier information regularly found in texts of other clinical domains that were processed by the Linguistic String Project [11].

# 5 The UMLS Semantic Network

When terms of the radiology domain were not found in Meta-1, they were manually assigned semantic classes based on the UMLS definitions of the semantic types. However, the UMLS definitions themselves contain inherent overlaps, which made the classification of new concepts somewhat uncertain. For example, according to the semantic definition of the UMLS types, the term *cardiac enlargement* is classifiable as a **Pathologic Function**, which is defined as *A disordered process, activity, or state of the organism as a whole, of a body system or systems, or of multiple organs or tissues.* However, the semantic type **Acquired Abnormality**, which is defined as *An abnormal structure, or one that is abnormal in size or location, found in or deriving from a previously normal structure*, also applies, as does the type **Congenital Abnormality**, whose definition is similar except that the abnormality *is present at birth.* In addition, the semantic type **Sign**, which is defined as *An observable manifestation of a disease or condition based on clinical judgment* also applies. Looking at the semantic classification of several other randomly chosen concepts in the UMLS did not clarify the issue. For example, *effusion* is classified both as a **Finding**, and as a **Pathologic Function**, but *edema* is classified as a **Sign** and a **Pathologic Function**.

The classification of clinical findings would be facilitated if there were an additional semantic type **Abnormal Finding** to cover abnormal conditions of the whole, part, or function of the organism. Then, the UMLS semantic types **Pathologic Function, Acquired Abnormality, Congenital Abnormality**, and **Finding**, which are typical findings of radiology exams, could be defined as children or descendants of **Abnormal Finding** in the semantic network. This representation would be accurate from the viewpoint of modelling clinical findings, and would simplify the classification of terms not in Meta-1, because a term could be classified as having the general category *Abnormal Finding* when the more specific categories overlap.

Presently, the addition of a node denoting **Abnormal Finding** would be problematic because of the current organization of the network. **Acquired Abnormality** and **Congenital Abnormality** are children of **Anatomical Structure, Finding** is a root node, and **Pathologic Function** is a child of **Biologic Function**. Therefore, there is no place in the network to put **Abnormal Finding** so that these four types are the children. This suggests that it would be appropriate to allow multiple hierarchies in the semantic network. If a type could have more than one parent, it would be possible to have different conceptual groupings for different viewpoints. Although a network with multiple inheritance is more complicated than a simple network, it would be a richer and more robust model in which to represent biomedical concepts.

Since the network could not be changed, we resolved the problem another way. New semantic types were created specifically for the radiology domain when applicable, and certain UMLS types were made to correspond to the new domain types. For example, a new semantic category **Abnormal Finding** was defined. The UMLS semantic categories **Congenital Abnormality**, **Acquired Abnormality**, **Finding**, and **Pathologic Function** were made to correspond to **Abnormal Finding**.

This technique also resolved another problem that occurred because of the way structural abnormalities are covered in the UMLS. Currently, a structural abnormality has to be classified as a **Congenital Abnormality** and/or an **Acquired Abnormality**. However, in radiology, a term such as *opacity* denotes a more general class, which is a **Structural Abnormality**, but that class does not exist in the UMLS network. Instead, *opacity* is classified as corresponding to both semantic types. In our application, both **Congenital Abnormality** and **Acquired Abnormality** are mapped into one common type **Abnormal Finding**. In this case, it would be easy to change the UMLS to add a new type **Structural Abnormality** as a parent of **Congenital Abnormality** and **Acquired Abnormality**.

# 6 The Controlled Vocabulary

The UMLS is an extensive source of a biomedical vocabulary, which also has the potential of alleviating the burden of creating a controlled vocabulary for a domain. Each biomedical concept has a list of synonymous terms, along with the preferred name for the concept. If a term in the radiology domain matches a concept in Meta-1 exactly, the preferred term could then be used as the target form which the original term should be translated into. In addition, synonymous terms could be obtained from Meta-1 and added to the vocabulary of the text processor; additionally, their target forms would also be the same preferred term. Thus, in theory, the UMLS could be used to build up the vocabulary of the domain, and to establish a standardized controlled target vocabulary.

In reality, however, there were very few exact matches between the terms of clinical radiology and Meta-1 concepts. Finding an equivalent concept in Meta-1 for a word or term based on a partial match was a difficult and laborious matching problem which required manual review. One problem occurred because all the words of the sample texts could not be used to automatically search Meta-1. General English words or terms had to be manually identified and excluded from the matching procedure; otherwise, partial matches would be found in Meta-1 for words, such as *the, an*, and *however*. English words constituted about 44% of the text. In addition, a morphological component was needed to stem words, so that a word such as *atherosclerotic* would match the UMLS concept *atherosclerosis*. About 14% of the clinically relevant words in the text had to be stemmed so that matches could be obtained.

Other difficulties occurred because some terms in the domain were more general than the UMLS concepts. This typically occurred when the domain term was found to be a word of a UMLS concept consisting of several words. For example, *tuberculosis* is a domain term, which is contained in the UMLS concepts *bone tuberculosis, joint tuberculosis, silicotuberculosis, and pulmonary tuberculosis* ; however it is more general than the UMLS concepts. In this case, a suitable term *pulmonary tuberculosis* was manually chosen from the alternatives because we knew the domain consisted of chest exams. Another problem occurred because some of the words in the domain term match some (but not all) words in a UMLS concept. This was the worst situation. For example, the term *interstitial markings* partially matched 20 UMLS concepts containing the word *interstitial*, and one UMLS concept containing the word *marking*. Thus, a list of 21 concepts had to be manually reviewed. In this case a few concepts were related in some way to the original term, such as *emphysema interstitial*, *idiopathic interstitial fibrosis of lung syndrome*, and *lung disease interstitial*, but did not have quite the same meaning as *interstitial markings*. The term containing *marking* was *denture identification marking*, which is a completely different concept.

Not surprisingly, terms corresponding to modifier type of information (except for words corresponding to body parts) were not in the UMLS. This constituted 45% of the vocabulary, as shown in Table 1. Matches were made for the remaining clinical terms, which consisted of body parts and Rad-findings. Only 45 terms of the 190 terms in the Rad-finding category matched a concept in the UMLS exactly, and 43 terms partially matched a UMLS concept that was appropriate. Therefore only 46% of the terms corresponding to Rad-findings were in the UMLS. Body part terms were represented somewhat better because 28 body part terms matched a UMLS concept exactly, and 12 terms partially matched an appropriate concept. Therefore 40 out of 48 (83%) of the body part terms were in the UMLS. However, a total of 110 terms corresponding to the Rad-finding and body part categories did not have an equivalent concept in the UMLS. The overall results show that a total of 66% of the clinically relevant terms were not in the UMLS.

# 7 Discussion

It is important to note that the algorithms used for matching the domain terms to UMLS concepts were preliminary and very simplistic. A more sophisticated matching algorithm, especially one with a comprehensive morphological component, would have simplified the task of finding equivalent concepts. Although better matching algorithms would have facilitated the task, it is unlikely that the task could be completely automated or that significantly more matches would be found.

Whenever a term from clinical radiology was found in the UMLS, information concerning its semantic categorization, its preferred form, and its synonymous forms were effectively used to obtain semantic knowledge for the text processor. If the UMLS were extended to include terminology from the domain, then it would definitely be a significant tool that could be used to develop a text processor for that domain. The effort expended for matching domain terms with UMLS terms would be worthwhile because that task would still be less time-consuming and would also require less expertise than developing a controlled vocabulary and semantic classification system from scratch. Considering that clinical radiology constitutes such an important part of clinical information, it would be very beneficial for the NLM to expand its coverage in this direction.

Ancther reason that it would be desirable to have UMLS terminology available for clinical applications is that the UMLS was designed to facilitate access to information sources for literature searches. If the target structures of a text processor consisted of preferred terms from the UMLS, the clinical information extracted from the text could be added to a clinical database in a form that could be used for literature searches to retrieve citations about clinical findings. Similarly, the findings could be used by medical decision support applications or by statistical applications for research or quality assurance. Using the UMLS to obtain a unified controlled vocabulary is an important step towards facilitating the integration of clinical findings with other automated information processes.

## Acknowledgements

This work was supported in part by grant number R29 LM05397 from the National Library of Medicine and by grant number 443206 from the Research Foundation of the City University of New York. The UMLS was provided by the National Library of Medicine under an experimental agreement. The author would like to thank Dr. Philip Alderson of the Radiology Department at CPMC, and her colleagues Dr. James Cimino, Dr. George Hripcsak, and Dr. Steven Johnson at the Center for Medical Informatics at CPMC for their contributions.

## References

[1] K. Canfield, B. Bray, S. Huff, and H. Warner. Database capture of natural language echocardiology reports: A UMLS approach. In *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, pages 350–353, Los Alamitos, CA, 1990. IEEE Computer Society Press.

[2] C.G. Chute, Y. Yang, and D.A. Evans. Latent semantic indexing of medical diagnoses using UMLS structure. In *Proceedings of the 15th Annual Symposium on Computer Applications in Medical Care*, pages 185,189, NYC, New York, 1991. McGraw-Hill, Inc.

[3] J.J Cimino. Representation of clinical laboratory terminology in the unified medical language system. In *Proceedings of the 15th Annual Symposium on Computer Applications in Medical Care*, pages 199–203, NYC, New York, 1991. McGraw-Hill, Inc.

[4] D. Evans. Pragmatically structured, lexical-semantic knowledge bases for unified medical language systems. In *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care*, pages 169–173, Washington, D.C., 1988. IEEE Computer Society Press.

[5] E. Gabrieli and D. Speth. Computer processing of discharge summaries. In *Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care*, pages 137–140, Washington, D.C., 1987. IEEE Computer Society Press.

[6] S.M. Huff and H.R. Warner. A comparison of Meta-1 and HELP terms: implications for clinical data. In R.A. Miller, editor, *Proceedings of the 14th Symposium of Computer Applications in Medical Care*, pages 161–165, 1990.

[7] B.L. Humphreys. *UMLS Knowledge Sources - 2nd Experimental Edition.* National Library of Medicine, Bethesda, Maryland, 1991.

[8] A. McCray. Extending a natural language parser with UMLS knowledge. In P.D. Clayton, editor, *Proceedings of the 15th Symposium of Computer Applications in Medical Care*, pages 194–198, 1991.

[9] D. Ranum. Knowledge based understanding of radiology text. In *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care*, pages 141–145, Washington, D.C., 1988. IEEE Computer Society Press.

[10] D. Rothwell, L. Hause, and C. Frey. Lab management memo. *Pathologist*, May 1982.

[11] N. Sager, C. Friedman, and M.S. Lyman et al. *Medical Language Processing: Computer Management of Narrative Data.* Addison-Wesley, Reading, MA, 1987.

[12] J. Wagner, R. Baud, F. Borst, C. Kohler, and J. Scherrer. A knowledge-based system for interactive medical diagnosis encoding, expert systems and decision support in medicine. In *33rd Annual Meeting of the GMDS EFMI Special Topic Meeting, Peter L. Reichertz Memorial Conference*, Hannover, West Germany, 1988.