Journal of Digital Imaging (2020) 33:797–813
https://doi.org/10.1007/s10278-020-00331-3

Ontology-Based Radiology Teaching File Summarization, Coverage, and Integration

Priya Deshpande[1] · Alexander Rasin[1] · Jun Son[1] · Sungmin Kim[1] · Eli Brown[1] · Jacob Furst[1] · Daniela S. Raicu[1] · Steven M. Montner[2] · Samuel G. Armato III[2]

Published online: 6 April 2020

Abstract
Radiology teaching file repositories contain a large amount of information about patient health and radiologist interpretation of medical findings. Although valuable for radiology education, the use of teaching file repositories has been hindered by the ability to perform advanced searches on these repositories given the unstructured format of the data and the sparseness of the different repositories. Our term coverage analysis of two major medical ontologies, Radiology Lexicon (RadLex) and Unified Medical Language System (UMLS) Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and two teaching file repositories, Medical Imaging Resource Community (MIRC) and MyPacs, showed that both ontologies combined cover 56.3% of terms in the MIRC and only 17.9% of terms in MyPacs. Furthermore, the overlap between the two ontologies (i.e., terms included by both the RadLex and UMLS SNOMED CT) was a mere 5.6% for the MIRC and 2% for the RadLex. Clustering the content of the teaching file repositories showed that they focus on different diagnostic areas within radiology. The MIRC teaching file covers mostly pediatric cases; a few cases are female patients with heart-, chest-, and bone-related diseases. The MyPacs contains a range of different diseases with no focus on a particular disease category, gender, or age group. MyPacs also provides a wide variety of cases related to the neck, face, heart, chest, and breast. These findings provide valuable insights on what new cases should be added or how existent cases may be integrated to provide more comprehensive data repositories. Similarly, the low-term coverage by the ontologies shows the need to expand ontologies with new terminology such as new terms learned from these teaching file repositories and validated by experts. While our methodology to organize and index data using clustering approaches and medical ontologies is applied to teaching file repositories, it can be applied to any other medical clinical data.

Keywords Unsupervised machine learning · Cluster analysis · Radiology teaching files · Medical ontologies · Coverage analysis · Data integration

Introduction

Teaching files can play an important role in radiology education by serving as reference resources and teaching materials for resident and medical student education. Although each hospital maintains an internal collection of teaching files, public teaching file collections are also available through curated online sources (e.g., Radiology Society

✉ Priya Deshpande
pdeshpa1@depaul.edu

Alexander Rasin
arasin@depaul.edu

Jun Son
json1@depaul.edu

Sungmin Kim
skim1@depaul.edu

Eli Brown
ebrown80@depaul.edu

Jacob Furst
jfurst@depaul.edu

Daniela S. Raicu
dstan@depaul.edu

Steven M. Montner
smontner@radiology.bsd.uchicago.edu

Samuel G. Armato, III
s-armato@uchicago.edu

[1] DePaul University, Chicago, IL, USA

[2] Department of Radiology, University of Chicago, Chicago, IL, USA

of North America Medical Imaging Resource Community (RSNA MIRC) [1], MyPacs [2], and EURORAD [3]). These data sources can be combined and indexed with the help of medical ontologies such as Radiology Lexicon (RadLex) [4] and Unified Medical Language System (UMLS) Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [5] (the terms "UMLS SNOMED CT" and "SNOMED CT" are used interchangeably in this paper). To understand the quality and breadth of content available in different teaching file repositories and ontologies and proliferate their usage as a reference source, the content of these repositories must be systematically quantified. For example, it would be valuable to know what diseases are covered (or not covered) thoroughly in a particular teaching file repository or how comprehensive the terminology in an ontology is for indexing the content of that particular repository.

A radiology teaching file is a collection of radiology images and case-related data from clinical reports corresponding to the images. Teaching file text includes different categories such as patient history, findings, diagnosis, physician discussion, references, and differential diagnosis. These teaching file repositories are, in some cases, further augmented by medical ontologies to help interpret and normalize the data content. Radiology report templates are used by radiologists to create reports in a defined structured format that provides consistency and clarity about the patient diagnosis information. Radiology report provides information about patient demographics, the imaging procedure, imaging observations, and a summary [6].

The RSNA has developed a library of more than 210 reporting templates contributed by radiology societies, institutions, and individuals [6]; radiologists can use these templates to create teaching files using details such as patient history, clinical findings, diagnosis, or treatment suggestions.

The RSNA has also developed the RadLex ontology to meet the terminology challenges of radiology by providing a uniform source of terms and concepts for radiologists [4]. In addition to the RadLex, a more general medical ontology, SNOMED CT [5], has been designated as the recommended clinical terminology reference for clinical information systems around the world [7]. In our previous work, we designed and developed an integrated radiology teaching file search engine, Integrated Radiology Image Search (IRIS) engine [8, 9], which currently encompasses two publicly available heterogeneous data sources, RSNA MIRC [1] and MyPacs [2], and two medical ontologies, RadLex [4] and SNOMED CT [5]. The clustering approach integrated with ontology-based indexing presented here enables categorization of the teaching files in these repositories and their content interpretation using terms from the two medical ontologies. While our analysis of the results will focus on the content richness of the two repositories and the coverage by the ontologies, these results can also be used to advance the IRIS engine functionality from matching terms and their synonyms to topic similarity learned by grouping data into clusters. Although more advanced machine learning approaches such as deep learning [10] can be applied to improve the process of extracting meaningful information from medical reports when a large amount of annotated data is available [7], we show that a simple hierarchical clustering approach can learn topics that can be used to understand the completeness and coverage of the teaching file repositories.

In the long run, our techniques and findings can be used to guide the development of reporting templates by identifying important concepts that occur frequently in radiology reports. These same techniques can further inform the development of future teaching file repositories given the current coverage of available teaching files in terms of diseases, drugs used, and referenced imaging modalities. Our preliminary results on four combinations of data sources and medical ontologies—MIRC with RadLex, MIRC with SNOMED CT, MyPacs with RadLex, and MyPacs with SNOMED CT—provide answers to the following questions that can be posed by the medical imaging community and radiology domain experts when navigating publicly available data repositories: (1) How well do medical ontologies cover popular teaching file data sources? (2) Which terms appear frequently in these data repositories? (3) What types of diseases are covered (and to what degree) in the teaching file repositories? (4) What is the amount of overlap between different medical ontologies and teaching file repositories? Posing and answering these questions not only will inform the need for other teaching file repositories but will also update the need for more effective data source integration. The summarization of the teaching file repositories will provide a glimpse into what information these teaching file repositories contain such as what imaging modalities, anatomical structures, diseases, and differential diagnosis; allowing quick exploration of these repositories at coarse or low level of granularity will improve their use for both educational and research uses. The ontology-based coverage analysis also provides insights into the gap between the terminology used in clinical reports and the various ontologies for the medical domain which could be further exploited to expand these ontologies with new medical terms and relations between terms. Our proposed approach can be used to analyze medical data source properties before they are used for an integration. Deshpande et al. [11] proposed biomedical data integration framework, where our proposed approach to learn dataset properties can be use. Therefore, an integration of data repositories as well as their ontology-based summarization and indexing can benefit the medical community by providing easy access to collective medical knowledge while providing a platform for ontology expansion.

Materials and Methods

In this section, we describe the radiology teaching file repositories, the methodology used for cluster analysis with teaching file repositories, and the coverage analysis of medical ontologies. Figure 1 summarizes the steps in our analysis process.

Radiology Teaching File Repositories and Medical Ontologies

Radiology teaching file repositories (both public and in-house) are heterogeneous data sources that can be used by different users such as radiologists, radiology students, and technicians. Due to data heterogeneity, the need exists to integrate different repositories into a unified resource for easier access (e.g., for IRIS or Open-i [12]). We describe two major, publicly available data repositories and two medical ontologies that we use for content summarization and coverage analysis of radiology data sources.

RSNA MIRC [1] The RSNA is an international society of radiologists, medical physicists, and other medical professionals. The RSNA supported the development of a suite of free software tools for education and research in radiology. Those tools are now available through an open-source development project, with the MIRC as one of the radiology teaching file sources containing reports of imaging studies. The MIRC is a large repository with more than 2500 teaching files and more than 12,000 images including patient history, diagnosis, findings, discussion, and external references (journal articles). Radiological terms are highlighted, linked to the RadLex browser (see discussion about the RadLex below), and used by the built-in text search. An example of teaching file case from the MIRC is shown in Fig. 2.

MyPacs.net [2] The MyPacs is a well-known, publicly available teaching file repository in which radiologists can create, modify, and upload teaching files. More than 37,000 multilingual cases (17,000 of the cases are public) are available with over 200,000 associated images. Users can search records based on anatomy, pathology, modality, age, and other attributes. Teaching file terms are not linked to a medical ontology; however, users can search the repository based on pathology terms.

RadLex [4] The RadLex is an ontological system that provides a comprehensive lexicon vocabulary for radiologists. It was created to make more efficient use of the growing amount of electronic information in the radiology environment and to more accurately search reports and perform data-mining. The RadLex has more than 45,000 terms, including disease, anatomy, and imaging observations. There are a total of 14 term categories: RadLex descriptor, imaging observation, procedure step, process, RadLex non-anatomical set, procedure, object, anatomical entity, report component, clinical finding, temporal entity, imaging modality, property, and non-anatomical substance. Every RadLex term belongs to one of 14 categories [13].The RadLex browser, developed by the RSNA, enables users to view RadLex's structure and content on a web platform; it links to articles from journals including the British Institute of Radiology (BIR) [14] and the American Journal of Neuroradiology (AJNR) [15].

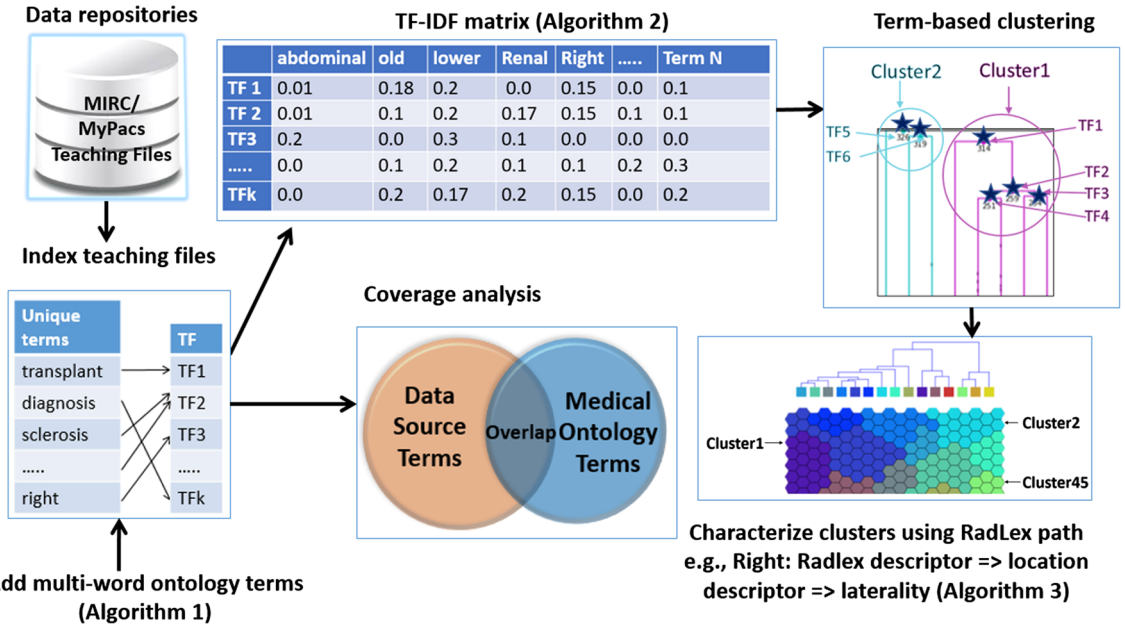

Fig. 1 Overview of proposed methodology

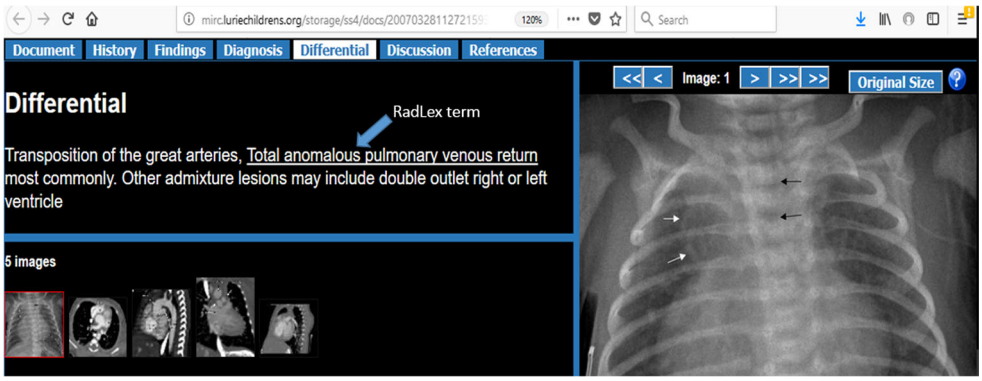

Fig. 2 A sample radiology teaching file case from MIRC with different categories and highlighted RadLex ontology terms

UMLS [16] The Unified Medical Language System (UMLS) Metathesaurus integrates many different vocabularies, including LOINC, MeSH, RxNorm, ICD-10, and SNOMED CT. In this work, we have integrated ontology sources that were most relevant to our data source type—focusing on SNOMED CT and ICD-10. Our IRIS is focused on clinical reports and radiology data sources. The SNOMED CT [17] is used in a variety of healthcare applications and also facilitates the use of coding terminology for clinical information in electronic health records, research, and clinical trials. International Classification of Diseases (ICD) [16] is a widely recognized international system for recording diagnoses using standardized codes. We did not integrate other ontologies from the UMLS such as Logical Observation Identifiers Names and Codes (LOINC) terminology [18] and the Medical Subject Headings (MeSH) [19] because our current IRIS implementation is not focused on patients' billing information or medical journal articles. The LOINC can be used in clinical care for laboratory results, claim management, and managing clinical information using universal codes. The MeSH vocabulary is used for indexing NLM journal collection.

SNOMED CT [5] The SNOMED CT ontology provides a standardized, multilingual vocabulary of clinical terminology that is used by physicians and other healthcare providers for the electronic exchange of clinical health information. The SNOMED CT ontology follows the National Library of Medicine (NLM) UMLS format [17]; it has a hierarchical structure and includes clinical findings, anatomy, test findings, and morphological connections. This ontology covers more than 300,000 terms with specified preferred name, synonyms, definition, and semantic meaning.

RxNorm [20] The RxNorm provides normalized names for clinical drugs and links these names to many of the drug vocabularies such as Gold standard drug Database, Multum [21], and Micromedex [22]. It offers a normalized naming system for generic and branded drugs—a tool built to support semantic interoperation between drug terminologies and pharmacy knowledge base systems. The RxNorm provides normalized names and unique identifiers for medicines and drugs. An electronic health record (EHR) is a digital version of a patient's record. The RxNorm can be used to support practical computing applications (e.g., e-prescriptions and patient medication history) in live EHR systems, which will be helpful to verify the current patient medications.

Data Pre-processing

Understanding the overall structure and the information stored in MIRC and MyPacs repositories is a crucial preliminary step to understanding their coverage quality. Our analysis primarily focused on five major text categories in teaching files: history, findings, diagnosis, differential diagnosis, and discussion. Through a survey of related literature and tools (MIRC, MyPacs, and EURORAD all use these categories) and by consulting expert radiologists, we determined that these five categories are the most important from a clinical perspective. Figure 2 shows an example of teaching file case from the MIRC repository with different categories: document (information about the author), history, findings, diagnosis, differential, discussion, and references, as well as a collection of images; the highlighted and underlined term in the differential category (see Fig. 2) shows an example of a RadLex ontology term linked to the RadLex browser. Among the many repositories reviewed, only the MIRC automatically indexes RadLex terms. All teaching files in the MIRC include text from the five categories; for example, in the MIRC dataset, the terms in these five categories constitute over 73% of all terms. The MyPacs repository contains over 17,000 publicly available teaching files; however, almost 1000 of those teaching files do not have any text from any of the five categories (history, findings, diagnosis, differential diagnosis, and discussion). We only considered teaching files with at least some content in these five categories, which yielded a total of 15,904 teaching files for MyPacs.

As part of the pre-processing of teaching file sources, we performed data cleaning, removed stop words, and applied stemming to terms in the text. However, we excluded some stop words from the standard removal list because they were

medically (i.e., diagnostically) relevant. For example, "no," "with," and "without" may be medically relevant and could appear within the ontology terms (e.g., "no sedation" and "dementia with Lewy bodies" are both RadLex terms). We identified all unique terms in the MIRC and MyPacs repositories to calculated term frequency and inverse document frequency (TF-IDF) matrix. We also augmented the TF-IDF matrix with multi-word medical terms from the RadLex and SNOMED CT ontologies using multi-word ontology term algorithm. In this algorithm, we augmented the term collection by appending multi-word terms from ontologies that appear in teaching file datasets (in addition to the individual terms). For example, "renal artery," "renal," and "artery" are three distinct terms and are thus treated separately when calculating unique term count and ontology overlap. We also used both multi-word and single terms for coverage analysis of ontology terms occurring in teaching file repositories. Algorithm 1 outlines the creation of the TF-IDF matrix with all unique terms (single- and multi-word); a term frequency weight represents the importance of the term for the given corpus. Table 1 summarizes the number of teaching files and the term counts for these teaching files.

Clustering of MIRC and MyPacs Teaching Files

Using the TF-IDF [23] matrix representation of the teaching files, we grouped each data source into a set of k clusters using the $k$means algorithm [24] to produce an initial grouping of the teaching files in each dataset. We considered values of k between 2 and 150 and used classification and regression tree (CART) [25] classifier to determine the best initial number of clusters where the classes were the k clusters, the splitting criterion was entropy, and minimum number of samples per node was varied to avoid overfitting. We selected the minimum value of k for which there was a significant decrease in the performance of the classifier when the number of clusters was increased to $k+1$. This process resulted in grouping each repository into 45 clusters—we then further clustered these 45 clusters into fewer clusters (making it easier to interpret) using a hierarchical clustering with Ward's linkage distance [26], which minimizes the total within-cluster variance. As a result, the MIRC dataset was finally clustered into six large clusters. By applying the same hierarchical clustering process, we generated 12 large clusters for the MyPacs dataset. The steps of this analysis are described in "MIRC and MyPacs Teaching File Clustering."

Table 1 MIRC and MyPacs term content summary

Dataset	MIRC	MyPacs
No. of teaching cases	2319	15,904
Single-word terms	15,944	87,272
Multi-word terms	1796	3141
Multi-word terms with RadLex	1780	3142
Multi-word terms with SNOMED CT	1830	3338
Total unique terms	17,740	90,413

Interpreting MIRC and MyPacs Clusters Using RadLex and SNOMED CT Ontologies

We interpreted the resulting clusters using the top 10 most frequent terms in each cluster for both RadLex and SNOMED CT ontologies. A RadLex ontology path traces the levels of the term hierarchy from leaf via its parent terms (e.g., "right ⟹laterality ⟹location descriptor ⟹ RadLex descriptor ⟹ RadLex entity"). We used the RadLex hierarchy to compute the term path for each RadLex term. RadLex path analysis was then applied to each cluster (i.e., each cluster generated 10 RadLex ontology paths, one for each of the top 10 RadLex terms). We also performed similar term analysis using the SNOMED CT; however, the SNOMED CT structure differs substantially from RadLex and no exact equivalent of path generation could be applied (i.e., there is no unique leaf-to-parent path present in the SNOMED CT ontology). For example, the term "Heart disease" has three parents in SNOMED CT ("Cardiac finding," "Disorder of cardiovascular system," and "Disorder of mediastinum") instead of offering a single path through parent terms as in the RadLex; therefore, for the SNOMED CT ontology, we used the ontology terms as is.

The steps used to generate a RadLex term path are summarized in Fig. 3.Our algorithm iterates through all RadLex terms: each RadLex entity has an associated unique identifier (RadLex ID) and a corresponding parent ID. For each RadLex term, we iterate through parent terms until we reach the root, capturing the path taken from the leaf node to the root term. For example, as shown in Fig. 3"Tegmental nuclei" is the (initial) leaf term. Our algorithm identifies the parent term from the leaf until it finds the root term ("Anatomical entity" in this example). The maximum term path length is 20 (i.e., RadLex terms may be up to 20 levels away from the root). For each of the clusters in the MIRC and MyPacs datasets, we determined the paths for the ten most frequent RadLex terms in every cluster. For example, in one of the MIRC clusters (containing 115 teaching files), "right" is the most frequent term, appearing 72 times. This term "right" belongs to the RadLex "location description" of patient anatomical structure; similarly, each of the other nine most frequent RadLex terms in a given cluster was represented by its RadLex path for the clustering analysis and interpretation.

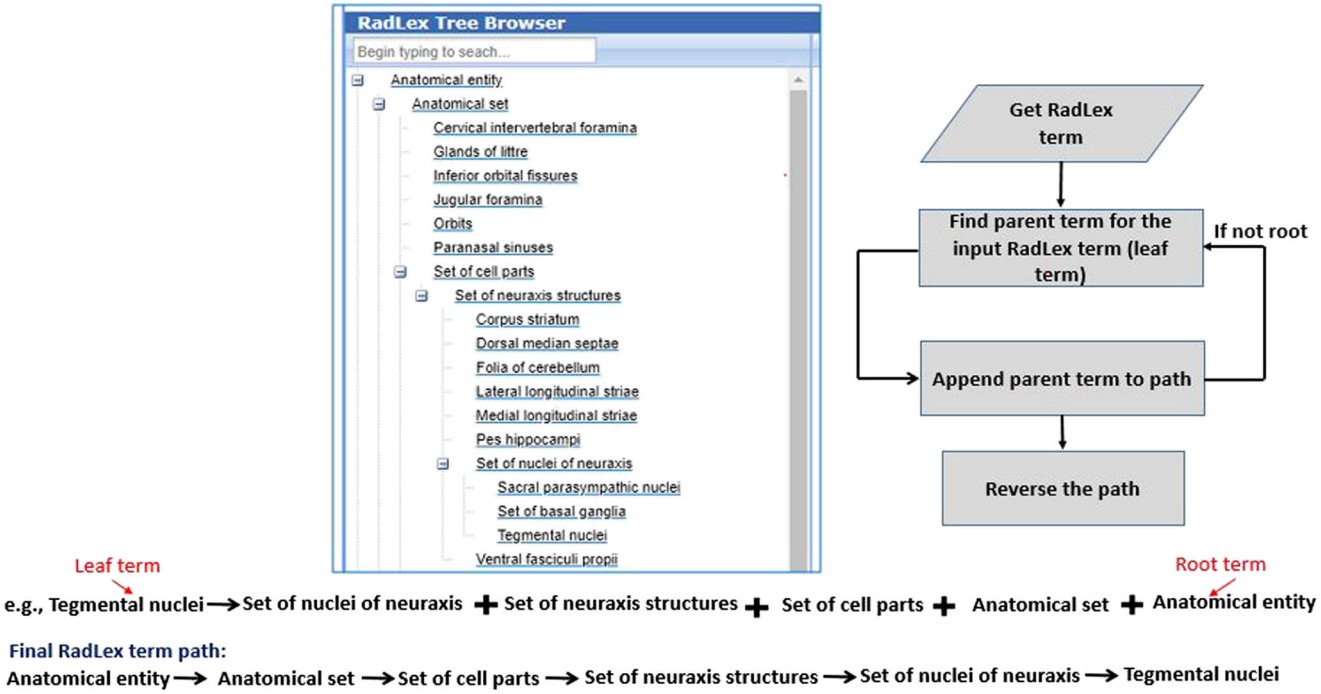

Fig. 3 RadLex path generation flow

Coverage and Overlap Analysis

Our analysis measured and compared medical ontology coverage of data sources. To compute term coverage, we evaluated how many of the defined ontology terms belong to our radiology datasets and further computed the term overlap between the medical ontologies themselves (i.e., overlap between the RadLex and SNOMED CT). If we denote an ontology with "O" and radiology dataset with "R," the coverage of O over R is defined as the percentage of terms from R that also appears in O. First, we calculate the unique terms in both sets—$UniqueTerms_O$ and $UniqueTerms_R$. We also performed coverage analysis for image modality (e.g., CT, MRI) distribution on both the MIRC and MyPacs to better understand the types of diseases and medical tests they cover. "Modality Distribution Analysis" describes our modality analysis. The coverage of O over R can then be described using Eq. 1. Our coverage results are presented in "Coverage Analysis."

$$TermCoverage_{o\to R} = \frac{UniqueTerms_O \cap UniqueTerms_R}{UniqueTerms_R}. \quad (1)$$

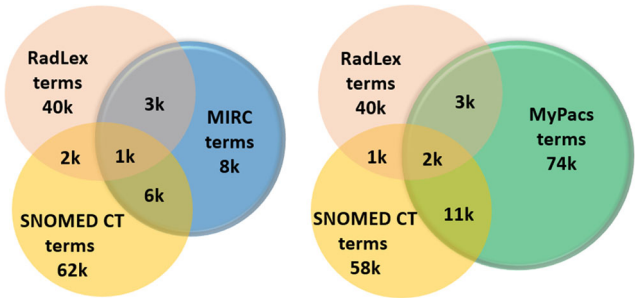

Fig. 4 Coverage analysis of RadLex and SNOMED CT ontologies with MIRC and MyPacs datasets (terms rounded to thousands—actual numbers are in "Discussion")

Results

Coverage Analysis

There are 45,000 unique terms in the RadLex and 71,000 unique terms in the SNOMED CT. Interestingly, only about 2000 of these terms are common, resulting in an overlap of less than 5% between the RadLex and SNOMED CT (details are shown in Fig. 4). The small overlap indicates that each ontology focuses on different types of terms, making the case for ontology integration when indexing content of medical

Table 2 RadLex entity terms distribution

Percentile of RadLex term	RadLex entity name
83.5	Anatomical entity
2–3	RadLex descriptor, property, imaging observation
1–2	Procedure
< 1	Objectives, clinical findings, procedure step, imaging modality, process, RadLex non-anatomical set, report component, temporal entity non-anatomical substance

Algorithm 1 TF-IDF matrix generation algorithm

```
TFiles ← Teaching files
MultiWordTerms ← ontology terms #from multi-word ontology term algorithm
Stopwords ← Stopword list
function TermFreq(termdic,docList)
    TFDict ={}
    N ← Length of the document list
    for word, count in termdic do
        TFDict[word] = count/N
    end for
    return TFDict
end function
function InvDocFreq(docList)
    IDFVal ={}
    N ← Length of the document list
    for doc in docList do
        for word, val in doc do
            if val > 0 then
                IDFVal[word]=IDFVal[word]+1
            end if
        end for
    end for
    for word, val in IDFVal do
        IDFVal[word]=log10 N/val
    end for
    return IDFVal
end function
function TFIDF_Generator(termdic,docList)
    return TermFreq(termdic,docList) × InvDocFreq(docList)
end function
AllTerms =[]
for TF in TFiles do
    tfTerms = []
    for term in TF do
        if term not in Stopwords then
            tfTerms.append(term)
        end if
    end for
    AllTerms.append(tfTerms)
end for
for term in MultiWordTerms do
    AllTerms.append(term)
end for
AllTermsIndex =[]
for term,termindex in AllTerms do
    AllTermsIndex.append([term, termindex])
end for
return TFIDF_Generator(AllTerms, AllTermsIndex)
```

Table 3 Expanding search query terms using RadLex and UMLS SNOMED CT ontologies with MIRC and MyPacs datasets

Query	mirc(2 k)				MyPacs(17 k)			
	No ontology	RadLex	SNOMED CT	RadLex + SNOMED CT	No ontology	RadLex	SNOMED CT	RadLex + SNOMED CT
Cardiomegaly	59	59	63	63	99	99	106	106
Bronchus intermedius	1	1	3	3	2	2	2	2
Chiari	19	19	38	38	133	134	153	154
Angiosarcoma	1	1	30	30	26	26	96	96
Cystitis cystica	0	0	3	3	0	0	2	2
Cystitis	7	7	10	10	95	95	96	96
Cystitis glandularis	2	2	5	5	0	0	2	2
Innominate vein	0	39	39	39	10	68	95	95
Innominate artery	0	238	238	238	0	855	866	866
Varicocele	2	2	4	4	24	24	28	28
Irregularly shaped	11	11	11	11	20	20	20	20
Acl tear	9	9	9	9	145	145	145	145
Study	117	117	117	117	776	776	776	776
Appendicitis	40	40	40	40	176	176	176	176
ACL graft tear	7	7	7	7	85	85	85	85
Hepatic adenoma	74	74	74	74	360	360	360	360
Annular pancreas	14	14	14	14	36	36	36	36
Perthe	20	20	20	20	63	63	63	63
Mega cisterna magna	3	3	3	3	9	9	9	9
Vertebra	243	243	243	243	753	753	753	753
Tracheal dilation	131	131	200	200	627	627	786	786
Toxic	48	48	48	48	165	165	165	165
Buford complex	43	43	43	43	178	178	178	178
Baastrup disease	0	0	0	0	1	1	1	1
Limbus vertebra	0	0	0	0	5	5	5	5
Splenic hemangioma	0	0	0	0	2	2	2	2
Double duct sign	0	0	0	0	0	0	0	0
Thornwaldt cyst	0	0	0	0	6	60	6	6

data repositories. Leveraging the hierarchical structure of the RadLex to determine the types of terms it covers, we found that the RadLex ontology primarily contains anatomical category terms (83.5%). Table 2 summarizes the distribution of all 14 RadLex entities. We also performed coverage and overlap analysis of the most frequent entities in the RadLex with respect to SNOMED CT ontology terms and found an overlap of only 13.5% with terms from the SNOMED CT.

We computed the coverage of the two data repositories based on the terms from each ontology. The MIRC repository has a coverage of 20% using RadLex terms and 42% coverage using SNOMED CT terms; when combined, both ontologies can offer an MIRC coverage of 56%. For the MyPacs repository, the RadLex and SNOMED CT combined coverage is 18% with almost no overlap between the two ontologies (2%).

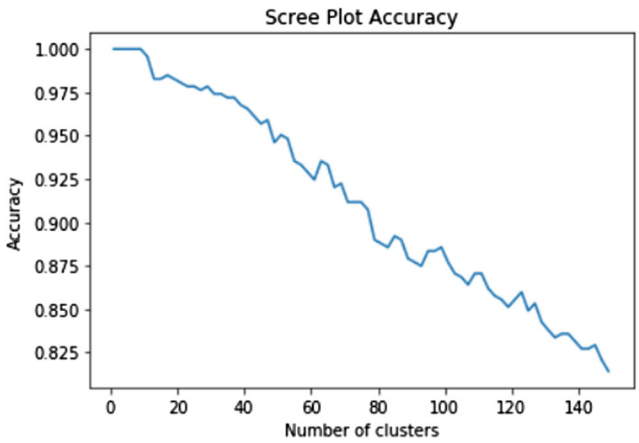

Fig. 5 MIRC cluster membership accuracy scree plot

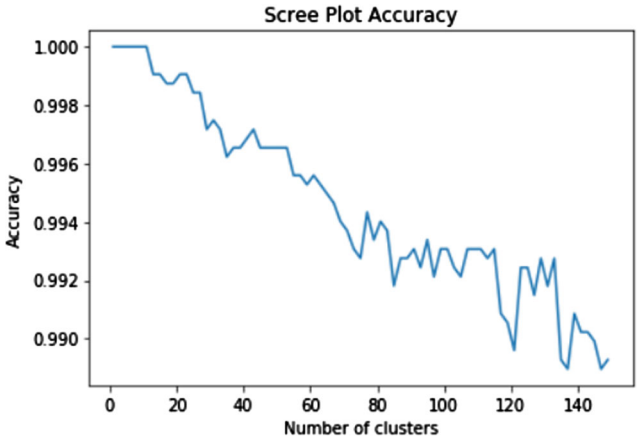

Fig. 6 MyPacs cluster membership accuracy scree plot

The breakdown of coverage and overlap numbers is shown in Fig. 4.

Note that a coverage of 100% is hard to achieve because the teaching files also include non-medical terms. Thus, in order to better understand ontology coverage, we repeated the same analysis with only the text from the "Diagnosis" category as this category is the least likely to contain non-medical discussion terms. The combined ontology coverage for the MIRC "Diagnosis" category increased to 88% (with 22% overlap between the RadLex and SNOMED CT), and the coverage for MyPacs content increased to 45% (with still only an 8.3% overlap between the two ontologies). These results confirmed that combining ontologies greatly increased the coverage, even after excluding most of the non-medical terms. MyPacs coverage is relatively low compared with MIRC because it includes foreign language terms (there are about 7000 non-English cases in the MyPacs repository) and because some of the teaching files in the MyPacs do not include diagnosis data.

We also performed coverage analysis using RxNorm ontology with MIRC and MyPacs data repository. RxNorm contains more than 160,000 unique terms related to drug information. Our coverage analysis shows that only 370 RxNorm terms appear in the MIRC repository and 1060 RxNorm terms occur in the MyPacs repository. Thus, only 2.3% terms are covered in the MIRC and 1.2% terms are covered in the MyPacs repository—which is not surprising. The MIRC and MyPacs are radiology teaching datasets and are not meant to contain medication information for the patient. Based on this minimal coverage from RxNorm, we did not perform an in depth analysis with RxNorm ontology and teaching file repositories. As RxNorm coverage is negligible (1–2%) in MIRC and MyPacs repositories, we did not perform further analysis on this ontology.

Ontology Coverage with Query Indexing on MIRC and MyPacs Datasets

We performed repository indexing analysis using the UMLS SNOMED CT and RadLex ontology on the MIRC and MyPacs datasets. We used 28 sample queries received from radiologists at a well-known medical hospital and from an extensive literature survey [27]. We executed these 28 queries against the MIRC and MyPacs database repository. Table 3 shows the documents retrieved using four types of search support ("no ontology," "with RadLex," "with SNOMED CT," and "with RadLex + SNOMED CT") from the MIRC and MyPacs datasets. Our results show how much integrating ontologies improved radiology teaching case retrieval. Initially, we observed that without integrated ontology support, many documents were missing from the search. After adding support for the RadLex ontology, search returns

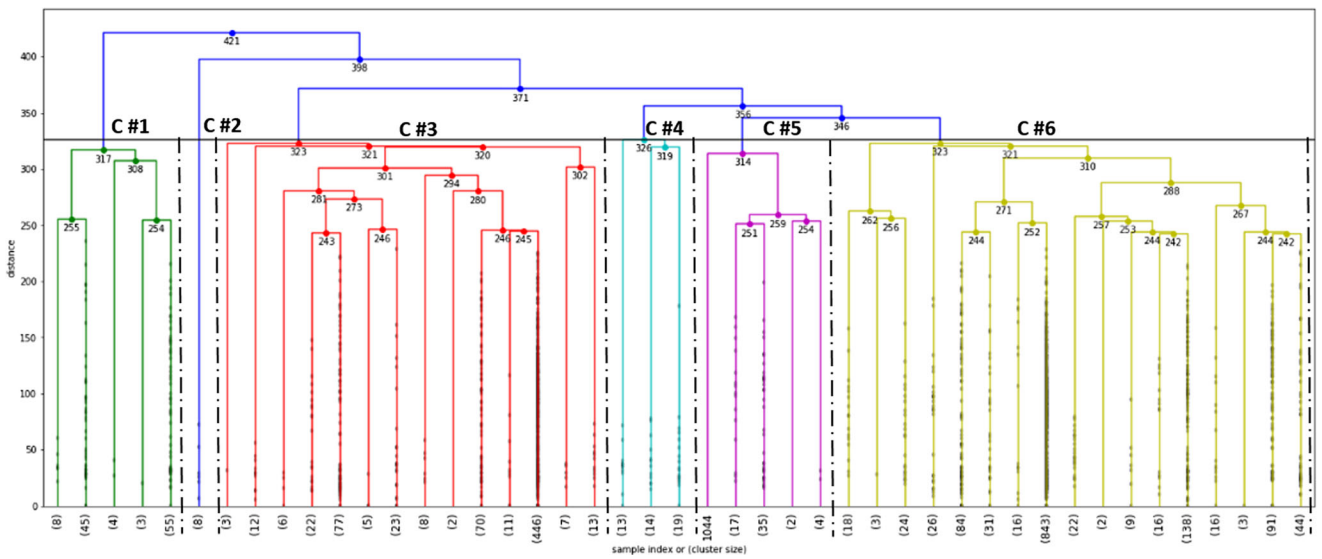

Fig. 7 Hierarchical clustering dendrogram for MIRC 45 clusters. *X*-axis is cluster size, *Y*-axis is the Ward distance. The vertical dash-dot-dash lines mark the boundaries between the final six clusters.

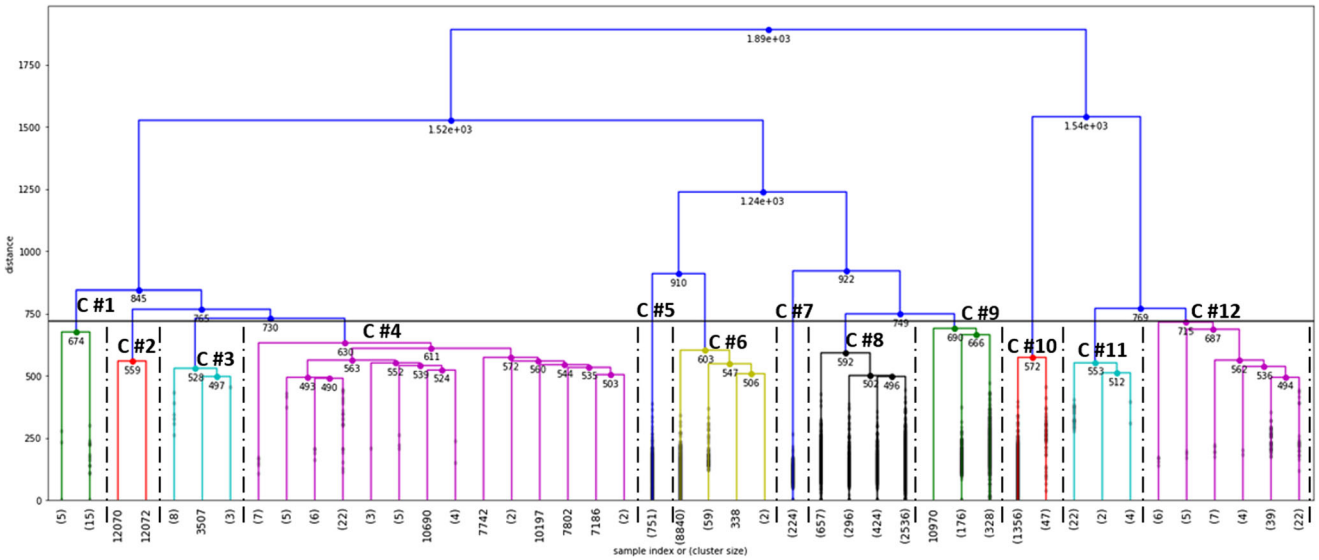

Fig. 8 Hierarchical clustering dendrogram for MyPacs 45 clusters. *X*-axis is cluster size, *Y*-axis is the Ward distance. The vertical dash-dot-dash lines mark the boundaries between the final six clusters.

several documents that were previously missing with no ontology support (e.g., query results for "Chiari," "Innominate vein").

We also performed document search operation with SNOMED CT ontology that returned new teaching cases for which there were no synonyms from the RadLex ontology (e.g., "cardiomegaly"). However, adding one ontology is not sufficient to fetch all relevant documents to query term; we expanded our ontology integration by including both the RadLex and UMLS SNOMED CT ontology terms. Results show an improvement in number of hits between these 4 types of search. There are cases where synonyms for a query term are present in both ontologies, but synonym term is not present in our the MIRC and MyPacs corpus. For example, for "Angiosarcoma," RadLex synonym terms are "Malignant hemangioendothelioma" and "angiosarkom." These synonym terms are not present in our corpus and do not contribute towards number of results. This analysis also shows that radiology teaching cases have less coverage of ontology terms and that can be improved by usage of medical ontologies in report or cases writing.

MIRC and MyPacs Teaching File Clustering

We used the CART classification (where the clusters represented the classification classes) to learn the best starting point for our clustering analysis. The scree plots in Figs. 5 and 6 show the classification accuracy for different numbers of clusters for MIRC and MyPacs teaching files, respectively. We chose 45 clusters as it results in a good accuracy with respect to the number of clusters (approximately 95% of teaching files are correctly classified into their chosen clusters at this point) and because it is the minimum value of k for which there is a significant decrease in the performance of the CART classifiers for both teaching file repositories.

The interpretation of the clusters can be done at the finest level (45 clusters) or at a coarse level of granularity where aggregated clusters provide a more abstract view of the data. We aggregated the 45 clusters by applying hierarchical clustering, and, using Ward's linkage distance to minimize the total within-cluster variance, we show different levels of aggregation through the dendrograms in Figs. 7 and 8 for the MIRC and MyPacs, respectively. We noticed that with six

Table 4 MIRC dataset cluster analysis using RadLex and SNOMED CT terms

Cluster	No. of subclusters	No. of teaching files	Total of words	No. of unique words	Average no. of unique words	No. of unique RadLex terms	No. of unique SNOMED CT terms
1	5	115	19,833	3133	27.24	1131	1812
2	1	8	2453	547	68.37	261	484
3	14	705	103,600	10,013	14.20	2455	4547
4	3	46	9914	1464	31.82	716	1096
5	5	59	11,255	2694	45.66	810	1692
6	17	1386	183,136	14,922	10.76	3162	6515

Table 5 MyPacs dataset cluster analysis using RadLex and SNOMED CT terms

Cluster no.	No. of subclusters	No. of teaching files	Total no. of words	No. of unique words	Average no. of unique words	No. of unique RadLex terms	No. of unique SNOMED CT terms
1	2	20	12,233	2814	140.7	702	1898
2	2	2	2624	2127	1063.5	231	1182
3	3	12	11,425	4812	401.0	1029	2947
4	14	61	44,945	10,617	174.05	1727	5149
5	1	751	76,239	4349	5.79	685	2145
6	4	8902	268,570	32,759	3.68	3295	8001
7	1	224	28,885	4297	19.18	700	1543
8	4	3913	565,149	51,746	13.22	4161	10,580
9	3	505	142,564	22,374	44.3	2754	7169
10	2	1403	147,894	14,510	10.34	451	2020
11	3	28	16,451	5848	208.86	260	1189
12	6	83	40,628	6006	72.36	234	1114

aggregated clusters (shown using different colors and the dotted vertical lines), we can still preserve the same top 10 most frequent ontology terms as were observed in the initial 45 clusters. Table 4 presents a summary of the final six MIRC clusters used in our interpretation analysis. We followed the same clustering process for both the MIRC and MyPacs. However, the MyPacs dataset is significantly larger than MIRC and covers a wider number of different diseases, anatomical structures, and cases written in different languages. We therefore chose to partition the MyPacs into 12 final clusters. We manually inspected the resulting clusters and concluded that, with fewer than 12 clusters, the clusters tend to contain an overly diverse variety of teaching files. Figure 8 shows the dendrogram for the MyPacs dataset clustering, and Table 5 presents the summary of the 12 final MyPacs clusters. Some interesting facts to note are that certain clusters have very little ontology coverage. For example, cluster no. 12 has only 3% of the RadLex terms; this can be explained by the fact that this cluster has 56 cases out of 83 cases being in French. Clusters no. 10, no. 11, and no. 12 also have a significant presence of French and Spanish teaching files and thus low ontology coverage.

MIRC Summarization Using RadLex and SNOMED CT

We used the RadLex path of the 10 most frequent terms and the 10 most frequent SNOMED CT terms to interpret the types of teaching files that each cluster contains. We provide a summary of the six clusters for MIRC teaching file repository in Table 6. Note that due to space considerations, Table 6 only includes the five most frequent ontology terms for both the RadLex and SNOMED CT, even though our summary analysis is based on a total of 20 terms (with 10 most frequent terms taken from the RadLex and 10 most frequent terms taken from the SNOMED CT). Notably, the bulk of summary information for each cluster was derived from the RadLex ontology because the RadLex path associated with each term offers a great deal of additional information (placing the term into its context). The contribution of the top 10 SNOMED CT terms (that do not have an associated path) was much less useful in generating an informative cluster summary. Cluster no. 2 has a large average number of unique words compared with other clusters even though this cluster has only eight teaching files that describe cases with "developmental dysplasia of the hip." The fact that the unique term average is high for this cluster may be explained by the fact that all cases are related to the same disease.

Using our approach, we were able to summarize 2319 teaching cases in one table with six clusters. From the overall cluster analysis, we can conclude that the current MIRC data repository has substantial coverage of pediatric and female patient diseases. Anatomical entities covered are heart, bones, and malignant tissues with "right" describing the most typical location, and cases that discuss a history of patient diseases are included. From the SNOMED CT with the MIRC dataset, our terms analysis concluded that these clusters cover cases with abdominal pain, heart diseases, and fracture cases.

MyPacs Summarization Using RadLex and SNOMED CT Ontologies

We analyzed the 12 MyPacs clusters using the 10 most frequent terms from each RadLex and SNOMED CT ontologies. As in Table 6, only the five most frequent terms from each ontology per cluster are shown in Tables 7 and 8. We summarized 15,904 teaching cases using 12 clusters as shown in Tables 7 and 8. RadLex terms are still a much better source of information (compared with SNOMED CT terms) due to the additional context contained within the RadLex path of each term. From the analysis of combined RadLex and

Table 6 MIRC clusters summary

RadLex and SNOMED CT terms	Cluster summary
•Right: RadLex descriptor ⟹ location descriptor ⟹ laterality •Thorax: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ subdivision of cardinal body part ⟹ subdivision of body proper ⟹subdivision of trunk ⟹ subdivision of trunk proper •Congenital: RadLex descriptor ⟹disease origin descriptor •Heart: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ Organ ⟹ cavitated organ ⟹ organ with cavitated organ parts •Present: RadLex descriptor ⟹ quantity descriptor •SNOMED CT: Heart, Right, Pulmonary, Left, Age	Focusing on anatomical structures of a body. Descriptors tell us about location of body parts (right or left), origin behind diseases, possible different diseases, and patient gender. Based on anatomical structure, location of disorder, and gender of patient, radiologists discuss conditions related to congenital heart, thorax, and aorta. Most of the teaching files within cluster no. 1 are about a female child with an abnormal structure of the heart, cases with enlarged heart, or problems related to blood vessels.
•Female: RadLex descriptor ⟹ patient descriptor ⟹ gender •Normal: RadLex descriptor ⟹ normality descriptor •Femur: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ Organ ⟹ cavitated organ ⟹ organ with cavitated organ parts ⟹ bone organ ⟹ long bone •Possibly: RadLex descriptor ⟹ certainty descriptor •Dysplasia: clinical finding ⟹ pathophysiologic finding •SNOMED CT: Abnormal, Adequate, Infant, Congenital, Dysplasia	Female patients with a possibility of having dysplasia. Anatomical entities focus on femur and tissues related to different organs. Based on clinical finding, treatment is provided under procedure category. Most of the teaching cases are related to female patient with dysplasia—an abnormal growth of tissues.
•Possibly: RadLex descriptor ⟹ certainty descriptor •Present: RadLex descriptor ⟹ quantity descriptor •Right: RadLex descriptor ⟹ location descriptor ⟹ laterality •Normal: RadLex descriptor ⟹ normality descriptor •Set of bone organs: Anatomical entity ⟹ anatomical set ⟹ set of organs ⟹ set of bone organs •SNOMED CT: Normal, Fracture, Proximal, Old, Image	Anatomical entities related to bones. Based on RadLex descriptors, the presence of abnormality in bone tissues is discussed. Treatment is typically suggested under the procedure category. Most of the cases are related to complaint about bone diseases of female patients discussed with uncertainty in diagnosis
•Treatment: procedure •Neoplasm: clinical findings ⟹ pathophysiologic findings ⟹ proliferation •Possibly: RadLex descriptor ⟹ quantity descriptor •Present: RadLex descriptor ⟹ quantity descriptor •Right: RadLex descriptor ⟹ location descriptor ⟹ laterality •SNOMED CT: Lesion, Disease, Male, Skin, Old	Growth in lesion under imaging observation category. Based on clinical finding, neoplasm and infection are observed; treatment is discussed. Most of the teaching files with female patients and diseases are related to neoplasm, abnormal growth of tissues, and infection in body parts.
•Intestine: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ subdivision of organ system ⟹ subdivision of alimentary system ⟹ subdivision of gastrointestinal system ⟹ subdivision of gut •Possibly: RadLex descriptor ⟹ quantity descriptor •Obstruction: clinical finding ⟹ pathophysiologic findings ⟹ mechanical disorder ⟹ flow disorder •Diagnosis: property •Proximal: RadLex descriptor ⟹ location descriptor •SNOMED CT: Abdominal, Obstruction, Proximal, Old, Patient	Clinical findings show mechanical disorder in body parts. Anatomical entities are related to abdominal organs such as the intestine, colon, and abdomen. Cases with findings of obstruction are observed in the colon body part.
•Possibly: RadLex descriptor ⟹ certainty descriptor •Right: RadLex descriptor ⟹ location descriptor ⟹ laterality •Present: RadLex descriptor ⟹ quantity descriptor •Treatment: procedure •Female: RadLex descriptor ⟹ patient descriptor ⟹ gender •SNOMED CT: Right, Lesion, Female, Old, Tissue	Both diagnosis and treatment are suggested based on clinical findings which are typically neoplasm. Patient gender is typically female.

SNOMED CT terms, we observed that MyPacs covers infection, neoplasm, vascular, and congenital diseases along with face, eye, neck, heart, and breast cases. Clusters no. 10, no. 11, and no. 12 contain most of the non-English cases in this repository. From this analysis, one can conclude that MyPacs repository content is neither about a particular disease nor focused on any gender, age, or even a particular anatomical structure. These clusters contain cases with diagnosis related to ear, face, head, heart, neck, chest, breasts, skeletal system, etc. There are also teaching cases that are related to pediatric patients. Anatomical entity references covered a variety of entries including heart, bones, and leg injuries.

Modality Distribution Analysis

To further understand the types of cases that the repositories contain, we performed an imaging modality distribution analysis on both repositories. We used 87 known modality terms [28] and looked for the frequency of occurrence for those modality terms in MIRC and MyPacs

Table 7 MyPacs clusters summary no. 1

RadLex and SNOMED CT terms	Cluster summary
•Stat: RadLex descriptor ⟹ imaging procedure descriptor ⟹ orderable priority •Cutaneous: RadLex descriptor ⟹ anatomically related descriptor •Complication: Property ⟹ interventional outcome ⟹ morbidity •Tin: Non-anatomical substance ⟹ chemical element •Gastrointestinal surgery: procedure ⟹ treatment ⟹ surgical procedure •SNOMED CT: Age, Air, Patient, Man, Face	Skin diseases are diagnosed, and surgical procedure is suggested as treatment. Tin mentions are significant because it is a trace element that is required in bone formation. Cases with fracture of the neck are discussed along with bone and joint injury cases.
•Computed tomography: Imaging modality ⟹ tomography •Plication: Procedure ⟹treatment ⟹ surgical procedure ⟹ gastrointestinal surgery ⟹ small bowel surgery •Routine: RadLex descriptor ⟹ imaging procedure descriptor ⟹ orderable priority •Stat: RadLex descriptor ⟹ imaging procedure descriptor ⟹ orderable priority •Abnormal: RadLex descriptor ⟹ normality descriptor •SNOMED CT: Abdominal, Agnostic, Male, Age, Diagnostic	This cluster has only 2 teaching files. A 16-year-old male with the left hemidiaphragm has been plicated, and the movement is limited, appearing to move passively related to movement of the right hemidiaphragm. A 13-year-old female patient with transjugular intrahepatic portosystemic shunt was observed. Gastric decompression treatment was suggested.
•Computed tomography: Imaging modality ⟹ tomography •Wide: RadLex descriptor ⟹ width descriptor •Low: RadLex descriptor ⟹ generalized descriptor •Appearance: property •Rib: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ organ ⟹ cavitated organ ⟹ organ with cavitated organ parts ⟹ bone organ ⟹ long bone •SNOMED CT: Infection, Normal, Drain, Face, Proximal	Computed tomography modality is used to perform the diagnosis as related to bone organs and descriptor that shows the size of the nodule. Cases with infection related to the brain, muscle, and other tissues are discussed.
•Ear: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ subdivision of cardinal body part ⟹ subdivision of head •Male: RadLex descriptor ⟹ patient descriptor ⟹ gender •Female: RadLex descriptor ⟹ patient descriptor ⟹ gender •Left: RadLex descriptor ⟹ location descriptor ⟹ laterality •Thin: RadLex descriptor ⟹ thickness descriptor •SNOMED CT: Diagnostic, Injection, Tumor, Man, Vascular	Male and female patient cases with severe back and neck pain are discussed. Cases related to the ear, aorta, blood vessel, and heart problem also discussed.
•No: RadLex descriptor ⟹ certainty descriptor •Fracture: clinical findings ⟹ pathophysiologic findings ⟹ injury •Ear: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ subdivision of cardinal body part ⟹ subdivision of head •Joint: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ anatomical cluster ⟹ anatomical junction •Injury: Clinical finding ⟹ pathophysiologic finding • SNOMED CT: Normal, Ultrasound, Fracture, Imaging, Chest	Cases with joint fracture and diseases related to the ear are discussed. Clinical findings show injury at joint anatomical entity. No term shows certainty descriptor about findings for those cases.
•No: RadLex descriptor ⟹ certainty descriptor •with: RadLex descriptor ⟹ concomitance descriptor •Pain: Clinical finding ⟹ symptom •Male: Patient descriptor ⟹ gender •Right: RadLex descriptor ⟹ location descriptor ⟹ laterality •SNOMED CT: Sign, Old, Late, Patient, Tissue	This cluster has cases with ear diseases and lesion mass description cases. Male patient of old age and with clinical findings such as pain with left laterality (location)

datasets. The top 5 most frequently referenced modality terms used by MIRC and MyPacs are as follows: MIRC ("CT," "MRI," "MR," "US," "X-ray") and MyPacs ("CT," "MRI," "MR," "US," "PET"). Another finding resulting from our modality analysis is that disease type is often connected to the type of diagnosis in the case file. For example, when MIRC and MyPacs cases are related to the diagnosis of "tumors," "bone injuries," "vascular condition," "spinal injuries," "breast cancer," and "heart abnormal conditions," specific modalities such as CT, MRI, or PET are used to diagnose these diseases; these correlations between diagnosis and imaging modality are also validated by the work in [29, 30]. For example, MRI was used to evaluate "Blood vessels" and "Abnormal tissue," while CT was used to evaluate "Vascular condition/blood flow" and "Pulmonary embolism." We further note that, in practice, radiologists use abbreviations for modality (e.g., "CT" for "computed tomography" and "US" for "ultrasound"); however, while the RadLex defines "Computed Tomography," "CT" is not an ontology term and thus it is not included in the coverage analysis or in the indexing of a teaching file using the RadLex.

Table 8 MyPacs clusters summary no. 2

RadLex and SNOMED CT terms	Cluster summary
•With: RadLex descriptor ⟹ concomitance descriptor •Neck: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ subdivision of cardinal body part ⟹ subdivision of body proper •Ear: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ subdivision of cardinal body part ⟹ subdivision of head •Old: RadLex descriptor ⟹ temporal descriptor •Head: Anatomical entity ⟹ cardinal body part •SNOMED CT: Normal, Ear, Abnormal, Man, Patient	A variety of diseases related to the neck, ear, and head are discussed. Cases related to brain diagnosis involved with observation done for the head and neck to see any abnormal manifestations related to these body parts.
•With: RadLex descriptor ⟹ concomitance descriptor •Male: Patient descriptor ⟹ gender •Cyst: clinical finding ⟹ pathophysiologic finding ⟹ proliferation ⟹ focal proliferation • Mass: Imaging observation ⟹ enhancement ⟹ lesion •Old: RadLex descriptor ⟹ temporal descriptor •SNOMED CT: Lateral, Benign, Vessel, Male, Normal	A variety of diseases related to cystic mass and abdominal mass are discussed. Cases related abdominal distension with displacement and left laterally (location) are discussed.
•No: RadLex descriptor ⟹ certainty descriptor •Ear: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ subdivision of cardinal body part ⟹ subdivision of head •Low: RadLex descriptor ⟹ generalized descriptor •Lateral: RadLex descriptor ⟹ location descriptor •Pain: Clinical finding ⟹ symptom •SNOMED CT: Fracture, Lateral, Pain, Right, Joint	A variety of diseases related to the ear and head are discussed. Location descriptor describes the laterality of diseases.
•No: RadLex descriptor ⟹ certainty descriptor •Face: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹subdivision of cardinal body part ⟹ subdivision of head •Male: RadLex descriptor ⟹ patient descriptor ⟹ gender •Stat: RadLex descriptor ⟹ imaging procedure descriptor ⟹ orderable priority •Normal: RadLex descriptor ⟹ normality descriptor •SNOMED CT: Ear, Pain, Male, Normal, Female	Findings show certainty about the malignancy of tissues or cells. Descriptor shows that findings are normal and several cases with the face and ear as an anatomical entity, related to male patients.
•Diagnostic: RadLex descriptor ⟹ quality descriptor •Stat: RadLex descriptor ⟹ imaging procedure descriptor ⟹ orderable priority •Vein: anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ cardinal organ part ⟹ organ region ⟹ organ segment ⟹ region of vascular tree ⟹ segment of venous tree organ •Abdomen: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ subdivision of cardinal body part ⟹ subdivision of body proper ⟹ subdivision of trunk ⟹ subdivision of trunk proper •Calcification: clinical finding ⟹ pathophysiologic finding ⟹ degenerative disorder ⟹ deposition ⟹ mineral deposition disorder •SNOMED CT: Neck, Abdomen, Chest, Liver, Head	Cases related to vascular malformation and abdomen are discussed. Cases with neck and head anatomical structure are discussed. Clinical findings show calcification in the abdomen body part. Procedure Stat is used in medical emergency and shows that cases discussed in this cluster are more severe.
• Injection: RadLex descriptor ⟹ imaging procedure descriptor ⟹ substance administration attribute ⟹ route of administration •Hyperintense: RadLex descriptor ⟹ imaging observation descriptor ⟹ modality-related characteristic ⟹ signal characteristic •Dens: Anatomical entity ⟹ material anatomical entity ⟹ anatomical structure ⟹ cardinal organ part ⟹ organ region ⟹ organ segment ⟹ segment of vertebra •Hypointense: RadLex descriptor ⟹ imaging observation descriptor ⟹ modality-related characteristic ⟹ signal characteristic •Vascular: RadLex descriptor ⟹ anatomically related descriptor •SNOMED CT: Patient, Male, Lateral, Diagnosis, Abnormal	Diseases related to brain cysts and lesions in the nervous system are discussed. Male patient with bladder incontinence is discussed.

Discussion

In this proposed work, we addressed two important questions: quantifying the coverage of medical ontologies in radiology data sources and interpreting radiology data sources to help researchers understand contents of different radiology data sources. This analysis would be useful for radiology data source integration to determine sources which would best

increase case coverage. Cluster analysis accuracy is verified using the scree plot accuracy. The scree plot shows that our cluster classification accuracy is above 95%. For cluster interpretation, we used the RadLex ontology that determines the meaning behind that term. We interpreted cluster content based on the top frequent terms that appear in these clusters. We further generated RadLex ontology paths that provide us with the meaning and root path for each defined term. We verified our RadLex path generation by manually comparing our RadLex term path with RadLex browser [4].

Our evaluation of term coverage by RadLex and SNOMED CT ontologies showed that both ontologies combined cover 56.3% of terms in the MIRC and only 17.9% of terms in the MyPacs. These findings indicate that there is a need to expand these ontologies to improve coverage and enable better indexing of medical repositories by the relevant medical ontologies. Woods and Eng [31] and Sandhu et al. [32] reported similar findings while studying the completeness of the RadLex ontology. Woods and Eng [31] estimated the completeness of the RadLex in the chest radiography domain and determined that, despite the large number of terms in the RadLex, gaps still exist. Sandhu et al. [32] investigated RadLex ontology uses in body imaging structured reports and concluded that the RadLex encompasses approximately 50% or less of terminology used in body imaging structured report templates, suggesting an opportunity for expanding RadLex content towards better coverage.

Consequently, several recent studies have proposed to learn new RadLex terms such as from mammography reports [33] and contextual patterns in radiology reports [34]. The use of contextual patterns and terms from mammography reports may also help disambiguate competing and imprecise RadLex definitions encountered in some instances [13], such as "mass" defined differently in Breast Imaging Reporting and Data System [35] versus the Fleischner definition [36]. Chan and Kahn have evaluated the completeness of a radiology glossary using iterative refinement [37]. Focusing on the readability of these reports by the layperson, Martin-Carreras and Kahn [38] studied coverage and readability of information resources from MedlinePlus, RadLex, and PORTER to help patients understand radiology reports and found that RadLex and PORTER offered significantly greater coverage than MedlinePlus.

An expansion of any prominent ontology (such as the RadLex) can also lead to improved indexing, structuring, and summarization of medical data repositories. Hong et al. [39] showed that there exists a substantial overlap between the terms used in the structured reporting templates and the RadLex and introduced techniques to establish global benchmarks for reporting templates [40, 41]. As a step towards free-text radiology report summarization, Goff et al. [42] used NLP and machine learning techniques to automatically extract asserted and negated disease entities from free-text radiology reports. Chen et al. [43] also applied machine learning techniques to categorize the oncologic response in radiology reports. Khan [38] used ontology-based knowledge to identify patients with rare diseases and estimated the frequency of those diseases in a large database of radiology reports. Hassanpour and Langlotz [44] used an unsupervised machine learning approach to identify topics in a radiology report repository and categorized the reports into nineteen major radiology categories with the number of clusters determined by an expert.

Similar to Hassanpour and Langlotz, we applied an unsupervised machine learning approach to partition and categorize teaching file repositories, but we algorithmically determined the number of clusters that could help with the automation of the categorization and therefore easier translation into clinical practice. We further automatically augmented the cluster interpretation with the term paths as defined by the RadLex and the definitions and synonyms in the SNOMED CT.

Our approach generated a number of clusters related to certain diseases, imaging modalities, body parts, and patient demographics by using two radiology teaching file repositories (MIRC and MyPacs) and two ontologies (RadLex and SNOMED CT). Ultimately, in order to show that our results are generalizable, we would need a much larger set of teaching file repositories and ontologies. Future work can be done to integrate other publicly available teaching file repositories such as the European Society of Radiology (EURORAD) [3], abstracts and images from the open-source literature, and biomedical image collections as enabled by the Open-i service of the National Library of Medicine [12]. To expand the coverage of medical data repositories with medical terms, other ontologies and vocabularies also can be integrated. For example, the use of RxNorm [20] can provide normalized names for clinical drugs and link these names to many of the drug vocabularies such as Gold standard drug Database [45], Multum [21], and Micromedex [22].

Conclusions

We presented a data-driven unsupervised machine learning approach to organize and summarize radiology teaching file repositories by their content augmented with terms from medical ontologies. Quantifying and automatically analyzing the content of publicly available teaching file repositories can broaden their use as educational resources and references in the diagnostic process because users will be able to access these repositories either at a high level of granularity (cluster topics) or a low level of granularity (teaching files within a certain cluster).

Our results revealed that the MIRC repository focuses on pediatric and female patients, with heart-, chest-, and bone-related diseases, while the MyPacs contains a range of

different diseases with no focus on a particular disease category, gender, or age group. These findings are important because they inform how these teaching file repositories can be further expanded and diversified either through integration or by collecting new cases that correspond to the existing clusters or will form new cluster topics. Furthermore, our results showed that the RadLex and SNOMED CT ontologies have low coverage with respect to each other as well as individually with each of the teaching file repositories. Therefore, these ontologies can be expanded by adding terms from the teaching files such as the top RadLex and SNOMED CT terms within each one of the clusters. Leveraging newer deep learning topic modeling techniques such as word2vec, relationships between terms can be discovered and integrated into the clustering approach. While the proposed approach was demonstrated on two teaching file repositories and two ontologies, it can be applied to any other publicly available or in-house teaching file repositories, clinical reports, or medical ontologies.

References

1. RSNA: Rsna tfs. http://mirc.rsna.org/query, 2018
2. McKesson Medical Imaging Group: Mypacs tfs. https://www.mypacs.net/, 2018
3. European Society of Radiology Neutorgasse: Eurorad. http://www.eurorad.org/, 2018
4. RSNA: RadLex ontology. http://www.radlex.org/, 2018
5. SNOMED International International Health Terminology Standards Development Organization: Snomedct ontology. http://www.snomed.org/, 2018
6. Heilbrun ME, Kahn CE, Applegate KE: From guidelines to practice: How reporting templates promote the use of radiology practice guidelines. J Am Coll Radiol:268–273, 2013. https://doi.org/10.1016/j.jacr.2012.09.025
7. Lee D, Cornet R, Lau F, de Keizer N: A survey of SNOMED CT implementations. J Biomed Inform 46(1):87–96, 2013. https://doi.org/10.1016/j.jbi.2012.09.006 ISSN 1532-0464. http://www.sciencedirect.com/science/article/pii/S1532046412001530.
8. Deshpande P, Rasin A, Brown E, Furst J, Raicu D, Montner S, Armato S III: An integrated database and smart search tool for medical knowledge extraction from radiology teaching files. 69:10–18, 2017. http://proceedings.mlr.press/v69/deshpande17a.html
9. Deshpande P, Rasin A, Sriram Y, Fang C, Brown E, Furst J, Raicu DS: Multimodal ranked search over integrated repository of radiology data sources. KDIR 372–383, 2019
10. Kent J: Machine learning, ehr big data analytics predict sepsis. https://healthitanalytics.com/news/machine-learning-ehr-big-data-analytics-predict-sepsis, 2018
11. Deshpande P, Rasin A, Furst J, Raicu D, Antani S: Diis: A biomedical data access framework for aiding data driven research supporting fair principles. Data 4(2):54, 2019
12. NIH: Openi. https://openi.nlm.nih.gov/, 2018.
13. Heilbrun ME: Evaluating RadLex and real world radiology reporting. Acad Radiol, 2013. https://doi.org/10.1016/j.acra.2013.09.011
14. BIR: Bir. https://www.bir.org.uk/, 2018.
15. AJNR:Ajnr. http://www.ajnr.org/, 2018.
16. UMLS: Umls. https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/, 2018
17. SNOMED: Snomednlm. https://www.nlm.nih.gov/healthit/snomedct/index.html, 2017
18. NLM UMLS: Umls loinc. https://www.nlm.nih.gov/research/umls/loinc_main.html, 2019
19. NLM UMLS: Umls mesh. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSH/, 2019
20. U.S. National Library of Medicine: Rxnorm. https://www.nlm.nih.gov/research/umls/rxnorm/, 2018
21. Cerner: Drug CERNER database. https://www.cerner.com/solutions/drug-database, 2018
22. Micromedex: Drug Micromedex database. https://www.micromedexsolutions.com/home/dispatch, 2018
23. Ramos J, et al.: Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning, volume 242,. Piscataway, NJ, 2003, pp 133–142
24. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 24(7):881–892, 2002. ISSN 0162-8828. https://doi.org/10.1109/TPAMI.2002.1017616
25. Stuart L: Crawford. Extensions to the cart algorithm. Int J Man Mach Stud 31(2):197–217, 1989
26. Murtagh F, Legendre P: Ward's Hierarchical agglomerative clustering method: which algorithms implement Ward's Criterion? J Classif 31(3):274–295, 2014
27. De-Arteaga M, Eggel I, Bao D, Rubin D, Kahn, Jr CE, Muller H: Comparing image search behaviour in the ARRS GoldMiner search engine and a clinical PACS/RIS. J Biomed Inform 56:57–64, 2015
28. DICOM Library: Modality. https://www.dicomlibrary.com/dicom/modality/, 2018
29. National Electrical Manufacturers Association: Mita. http://www.medicalimaging.org/about-mita/medical-imaging-primer/, 2018
30. WHO: Who-imaging modalities. http://www.who.int/diagnostic_imaging/imaging_modalities/, 2018
31. Woods RW, Eng J: Evaluating the completeness of RadLex in the chest radiography domain. 20:1329–1333, 11 2013.
32. Wang KC, Sandhu RS, Shin J, Shih G: RadLex and structured reporting in body imaging. 2017.
33. Bulu H, Sippo DA, Lee JM, Burnside ES, Rubin DL: Proposing new RadLex terms by analyzing free-text mammography reports. J Digit Imaging:1–8, 2018
34. Percha B, Zhang Y, Bozkurt S, Rubin D, Altman RB, Langlotz CP: Expanding a radiology lexicon using contextual patterns in radiology reports. J Am Med Inform Assoc 25(6):679–685, 2018
35. ACR: Acr-birads. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads, 2018
36. MacMahon H, McLoud TC, MAijller NL, Remy J, Hansell DM, Bankier AA: Fleischner society: glossary of terms for thoracic imaging. Radiology, 2008. https://doi.org/10.1148/radiol.2462070712 https://www.ncbi.nlm.nih.gov/pubmed/18195376
37. Chan PYW, Kahn CE: Evaluating completeness of a radiology glossary using iterative refinement. J Digit Imaging:1–3, 2018
38. Martin-Carreras T, Kahn, Jr CE: Coverage and readability of information resources to help patients understand radiology reports. J Am Coll Radiol, 2017
39. Hong Y, Zhang J, Heilbrun ME, Kahn CE: Analysis of RadLex coverage and term co-occurrence in radiology reporting templates. J Digit Imaging 25(1):56–62, 2012. ISSN 1618-727X. https://doi.org/10.1007/s10278-011-9423-9
40. Hong Y, Kahn CE: Content analysis of reporting templates and freetext radiology reports. J Digit Imaging 26(5):843–849, 2013. ISSN 1618-727X. https://doi.org/10.1007/s10278-013-9597-4
41. Hong Y, Zeng ML, Zhang J, Dimitroff A, Kahn, Jr CE: Application of standardized biomedical terminologies in radiology reporting templates. Inf Serv Use 33(3-4):309–323, 2013 ISSN 0167-5265. http://dl.acm.org/citation.cfm?id=2596874.2596884

42. Goff DJ, Loehfelm TW: Automated radiology report summarization using an open-source natural language processing pipeline. J Digit Imaging:1–8, 2017
43. Chen P-H, Zafar H, Galperin-Aizenberg M, Cook T: Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. J Digit Imaging:1–7, 2017
44. Hassanpour S, Langlotz CP: Unsupervised topic modeling in a large free text radiology report repository. J Digit Imaging 29(1):59–62, 2016
45. Elsevier: Drug database. https://www.elsevier.com/solutions/drug-database, 2018.