An Integrated Database and Smart Search Tool for Medical Knowledge Extraction from Radiology Teaching Files

Priya Deshpande, Alexander Rasin, Eli Brown, Jacob Furst, Daniela S. Raicu
DePaul University
Chicago, IL 60604
pdeshpa1,arasin,ebrown80,jfurst,dstan@depaul.edu

Steven M. Montner, Samuel G. Armato III
University of Chicago, Department of Radiology
Chicago, IL 60637
smontner@radiology.bsd.uchicago.edu,s-armato@uchicago.edu

ABSTRACT

Accurate and timely diagnosis is crucial for an effective medical treatment. Teaching files are widely used by radiologists as a resource in the diagnostic process and to teach students of radiology. Teaching files contain images, recorded discussion and notes, external references, augmenting annotations, and patient history. Most hospitals maintain an active collection of teaching files for their internal purposes, but many publically available teaching files are available through online sources that typically provide a basic keyword search interface but little else that can help physicians find the most relevant examples. Other secondary sources (e.g., journals or radiology textbooks) might also be referenced from a teaching file or provide an independent source of information; however, journal and textbook search capabilities, if available, can be very ad hoc and even more limited than for public teaching file repositories. Therefore, in order to access multiple resources, radiologists need to manually navigate each particular source and aggregate the search results into a full answer.

In this paper, we describe our integration of multiple public data sources into a unified medical resource repository and the design of advanced search features that make it easier to find relevant teaching files as well as journals or textbooks. Our approach supports incorporating diverse public data that can be further combined with a hospital's in-house teaching files to provide an integrated radiological knowledge repository. We tested our Integrated Radiological Image Search (IRIS) engine using a set of representative queries. Our search engine finds more accurate and relevant results compared to search engines available for public data sources. The IRIS engine is tailored to facilitate understanding of natural language queries, including negation statements, synonym terms, adjectives, and different sources of text. In addition, the search engine is designed to allow further integration of a module for image-based search to allow finding of visually similar cases.

KEYWORDS

Radiological Teaching Files, Data Integration, Integrated Radiological Image Search, NLP, Image-based Search

MIH'97, Halifax, Nova Scotia, Canada

DOI: 10.1145/nnnnnnn.nnnnnnn

ACM Reference format:
Priya Deshpande, Alexander Rasin, Eli Brown, Jacob Furst, Daniela S. Raicu and Steven M. Montner, Samuel G. Armato III. 2017. An Integrated Database and Smart Search Tool for Medical Knowledge Extraction from Radiology Teaching Files. In *Proceedings of Workshop on Medical Informatics and Healthcare, Halifax, Nova Scotia, Canada, August 2017 (MIH'97),* 8 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A radiology teaching files system is a collection of important cases for teaching and clinical follow-up, and references to understand the spectrum of a disease [3]. All teaching files share a similar overall structure but significant variations exist even within the same data sources. Teaching files can include information such as patient history, findings, diagnosis, differential diagnosis and images related to clinical reports. Teaching files can be categorized into three types: (1) personal teaching files that are meant for the general use of the teaching file owner, (2) shared in-house teaching files where the owner makes the teaching file content available for viewing within their institution, and (3) public teaching files built on a shared model but with more comprehensive content that may undergo a formal review before publication [2].

A recent national survey assessing the role and desired features of radiology teaching files found that, among the 396 respondents from 115 institutions, 89% use some form of teaching file from which 76% keep a personal teaching file containing a variety of media and 67% use a shared in-house teaching file, while 83 institutions had paid subscriptions to a public teaching file repository [3]. Public teaching file solutions have become increasingly popular, providing users with instant access to thousands of cases (although of inconsistent quality) [23], sometimes for a fee. These solutions include StatDx, RadPrimer, and ACR Learning Files (the most popular among survey respondents) as well as Radiolopolis, BrighamRad, EuroRad, MedPix, AuntMinnie, ACR Learning files, MyPacs.net [7], and RSNA Medical Imaging Resource Community (MIRC) [22]. While all of these public and commercial solutions are available, most do not permit a user to (1) easily submit personal cases to their libraries, (2) perform efficient querying, categorization and search for particular cases, (3) simulate basic PACS (picture archiving and communication system) functionality, or (4) enable self-directed and assessed learning – all very important teaching file features as identified by at least 50% of the survey respondents [3].

Therefore, as the first step to organize and extract medical knowledge from large teaching file repositories, we have 1) developed a database schema for teaching file integration and a framework for smart search, and 2) evaluated the framework on the RSNA

MIRC and MyPacs repositories. We normalize all data sources and augment the integration process with data cleaning and validation, since data comes in different format representation. We also ensure that Health Insurance Portability and Accountability Act (HIPAA) constraints are satisfied by not displaying patients' protected health information. Some teaching file sources (e.g., MIRC) already use Radiology Lexicon (RadLex) based annotations; we automatically annotate all of the imported data in our database.

In Section 2 we present our database schema, the teaching file repositories we integrated, the RadLex and SNOMED ontologies, and the Natural Language Processing (NLP) techniques we applied to perform the smart search. In Section 3 we present some related relevant data sources and other research that motivated our work; in Section 4 we present a comparative analysis of our Integrated Radiological Image Search (IRIS) system to the MIRC search engines as well as customized Google searches. In Section 5 we summarize the steps for data integration and outline both our short- and long-term plans to extend this work.

2 MATERIALS AND METHODS

2.1 Creation of logical schema and data integration

We designed a generalized database logical schema (Figure 1) and populated the database with data extracted from publically available teaching file repositories. The base entry in the center of the schema is a combination of teaching file record and an image entry since teaching files are naturally built around visual examples (e.g., MRI, X-ray images). Each such image is then annotated with a variety of related information, both from the data source (e.g., information from the differential diagnosis, patient history and discussion fields of the teaching file) and derived data (e.g., image properties, indexes or image feature extracts). Each teaching file entry is further linked with references and, diagnosis information as well as patient data, physician data and a pathology report (when available).

To implement the database logical schema, we first compared many relational databases including MySQL, SQL Server, Oracle and PostgreSQL. Based on the radiology teaching files' types of data, we determined that the PostgreSQL database is better suited for heterogeneous database connectivity and data integrity preservation. We then populated the database with publically available data sources from two major repositories.

The Radiology Society of North America (RSNA) Medical Imaging Resource Community (MIRC) [22] is a large repository (over 2,600 public teaching files with more than 12,000 images) with teaching files including patient history, diagnosis, differential diagnosis, findings, and discussion as well as external references (journal articles). Radiological terms are highlighted and linked to the RadLex terms as described in the next section. MIRC links to the RadioGraphics and Radiology journals, which can provide additional data sources for medical knowledge extraction. Although a rich source of medical knowledge, the MIRC repository's search engine does not handle specialized search fields (such as anatomy, age, and imaging modality), does not recognize negation, and does not have the ability to perform query expansion through synonyms to improve the search.

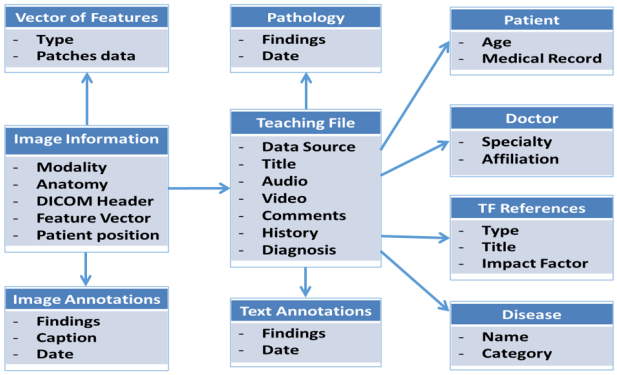


Figure 1: Logical Schema used in IRIS

The MyPacs.net data repository has over 32,396 cases with 202,986 images. Users can search records based on anatomy, pathology, image modality, age, gender, etc. The data repository also links to MedScape for additional supplemental information; however, the built-in search engine cannot handle synonyms and negations, and, similar to the MIRC search engine, it does not have the ability to perform image-based search.

2.2 Text-based processing and indexing

When querying large bodies of unstructured text (e.g., teaching file patient history and discussion categories), indexing is necessary to optimize query response time and to support custom analysis. Text indexing will rely on frequent word use in the data repositories and medical ontologies. Our current indexing implementation integrates two popular ontologies, RadLex and SNOMED.

RadLex [21] is an ontological system that provides a comprehensive lexicon vocabulary for radiologists. The RadLex browser was developed by the RSNA and includes 75,505 defined terms. The RadLex browser also links to articles from journals including the British Institute of Radiology (BIR) and American Journal of Neuroradiology (AJNR).

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [19] ontology provides a standardized, multilingual vocabulary of clinical terminology that is used by physicians and other health care providers for the electronic exchange of clinical health information. The SNOMED ontology follows the National Library of Medicine (NLM) Unified Medical Language System (UMLS) format[24]; it has a hierarchical structure and includes clinical findings, anatomy, test findings, and morphological connections. This ontology covers 106,624 terms with preferred name, synonyms, definition, and semantic meaning.

In terms of text processing, our IRIS engine applies stemming to the keywords in order to effectively use medical ontologies. For example, without changing the meaning of the term, "enlarged heart" and "enlargement of heart" retrieve the same results because they are both searching for "large" in association with "heart".

To keep the IRIS content up-to-date, we propose to update the IRIS database bi-monthly based on our observation that publically available data sources change slowly or might not change at all

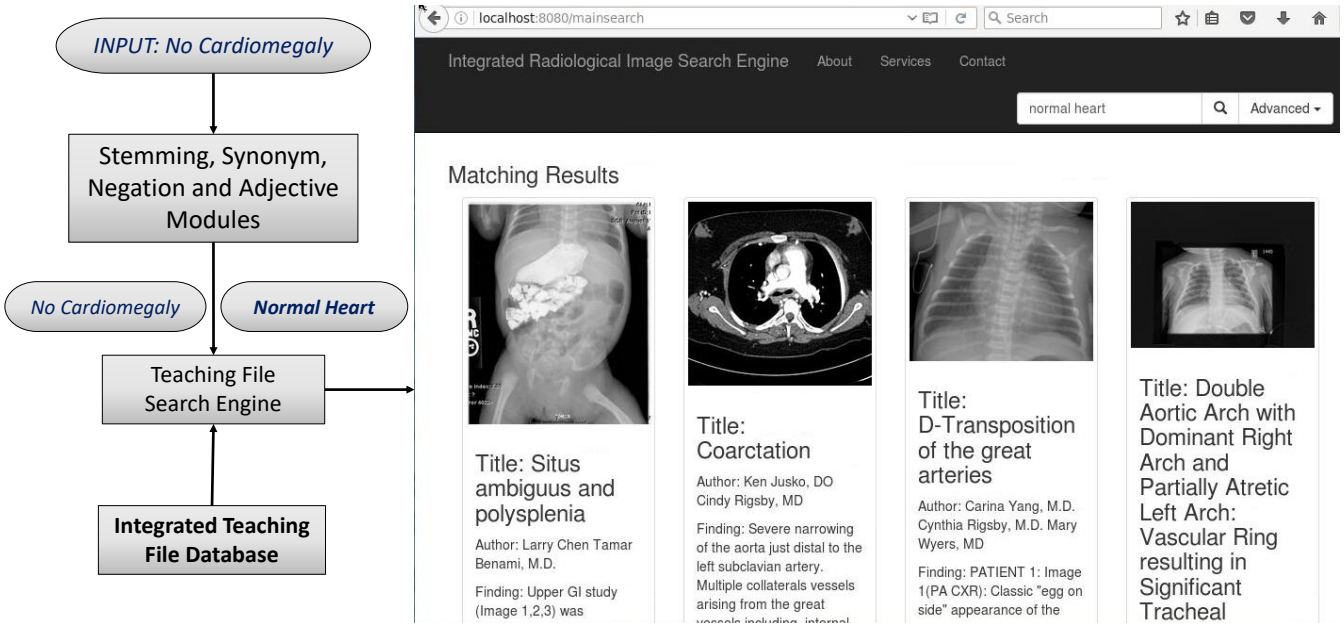


Figure 2: (Left) – Workflow of the IRIS Engine; (Right) – Teaching files and associated images displayed in response to the "no cardiomegaly" and "normal heart" query.

depending on the support and scope of their creation. Given that the database records include the "Date of Modification" attribute, we will compare the two values of the same attribute, one stored in the database and one found online; if the data repository is more recent, IRIS will update the corresponding stored teaching file and all its associated indexes. Similarly, the IRIS content will be updated based on any changes in the SNOMED and RadLex ontologies.

2.3 Smart search through synonym, negation, and adjective interpretation

Rather than limiting searches to an exact match between the query and the data in the integrated database, our search engine performs an automatic query expansion augmenting the search with synonym terms found using RadLex, SNOMED, and the Oxford medical dictionary. For example, if a user searches for "kidney", our IRIS engine will expand the query augmenting the "kidney" search with additional terms such as "kidney structure", "renal structure", and "nephros".

We handle negation by expanding the search from exact matches to queries that contain synonyms for that negation; for example, a typical search for "no X" is expanded with "missing X", "lacking X", "absent X", and "without X". We also substitute negation expressions to better handle these types of queries. For example, query such as "no abnormal renin secretion" will be substituted by the IRIS engine to "normal renin secretion".

To improve quality of the search, we also propose to handle qualitative adjectives through a quantitative equivalent. More specifically, we propose to quantify size adjective terms with values; for example, "smallest" receives a corresponding rating of 1, "smaller" will be 2, "medium" is 3, "large" is 4, and "larger" is 5. If a user searches for a larger lung nodule, then our search engine will search for both "large lung nodule" and "larger lung nodule". In our preliminary adjective quantification evaluation, we will use the semantic characteristics and their associated meanings/ratings for lung nodules based on the NIH/NCI Lung Image Database Consortium (LIDC) [1]; the LIDC semantic characteristics quantified through numerical ratings are: calcification (1-6), internal structure (1-4), and lobulation, malignancy, margin, sphericity, spiculation, subtlety, and texture (each on a scale from 1 to 5) [30].

Figure 2(left) shows the flow of our IRIS engine with the results displayed through a web-based user interface. The flow is illustrated using an example query "no cardiomegaly"; the query was expanded with a search for "normal heart" which is automatically added through the substitution of synonyms and applying negation (described in detail in the Results Section). Based on both of these queries and the expanded query, the IRIS search engine retrieves and displays teaching files images along with the associated text.

3 RELATED WORK

In this section we present a literature review of papers that discuss the need for data integration of radiological sources. We summarize existing radiological sources based on their search engine capabilities, advantages, and limitations. While our current implementation integrates MIRC and MyPacs data sources, as part of our future work we plan to integrate the other publically available radiology teaching file repositories that include categories such as history, findings, diagnosis and differential diagnosis.

Several studies have discussed the need to integrate clinical reports and images into integrated databases with smart search capabilities. Gutmark et al. [8] argued for building a system that reduces errors in radiological images' interpretation using teaching file databases. Easy-to-use computer-based teaching files are useful

for training physicians and serving as a reference tool for experienced physicians with the long-term goal of improving diagnostic accuracy. Talanow et al. [27] discussed how critical radiologic images are for diagnosis, teaching needs, and research. Dos-Santos et al. [6] discussed how the availability of a large and diverse set of clinical cases drives the need for the integration of profiles published by Integrating the Healthcare Enterprise (IHE).

Margolies et al. [16] found that a repository of pathology-proven cases in a dashboard has the potential to enhance and encourage the formation of accurate teaching files, as well as educational publications in the form of case series or "case of the day" submissions. Hwang et al. [10] discussed how the use of positron emission tomography computed tomography (PET-CT) increased the need to retrieve relevant medical images that can assist in image interpretation. Furthermore, Kansagra et al. [12] presented the idea of having a global database that integrates multiple data sources (such as clinical data, patient history, physical exam findings, laboratory data, or imaging data) for more precise and accurate diagnosis.

In the rest of the section, we present a list of currently or previously available data repositories and medical search engines along with their advantages and limitations (summarized in Table 1).

CTisus [9] is a large repository of radiological images, quizzes, and CT protocols. Although there are 237,814 images available along with video files. The repository does not contain case diagnosis, history of the patient, or differential diagnosis, and there is no support for image-based searches. Medscape [15] the latest medical news and information source about drugs and diseases available for radiology students and physicians. Medscape articles are focused on different anatomical structures, including experts' viewpoints and guidance. Even though Medscape is rich in medical data, there is no search engine available nor teaching files containing images, patient history, differential diagnosis or other valuable case information. Radiopaedia [11] is an open-edit radiology resource with 25,640 cases and 10,409 articles. As with all of the other data sources, no image-based search is provided. Gamuts [20] contains a comprehensive list of image differential diagnoses that are linked to symptoms, disease names, and causes. Although images are linked to the GoldMiner Radiology search engine, Gamuts does not offer a text-based search as it has no search engine. The Casimage database within the IRMA framework [28] integrates multimedia teaching and reference data into the PACS environment. The database includes only 8,723 images and it does not feature concept- or image-based retrieval. ImageCLEFmed Teaching Files [17] is a collection of domain– specific photographs for the medical field, which was used in the medical ad hoc retrieval tasks from 2004 to 2007. This medical archive comprises a total of 66,662 images and several composite medical sub-collections provided by independent medical institutions and hospitals. Cases are represented as group of related images and annotations and are meant to provide an evaluation forum for the cross–language annotation and retrieval of images.

Radiology Teacher [27] is a web-based teaching file development and distribution program, which allows authors to create, edit, and delete cases and images with descriptions and annotations. A quiz mechanism and an image annotation feature integrates an interface to the Medical Illustrator software are also provided. On the web portal any user can change stored teaching files. Presently, the Radiology Teacher system contains only 321 cases.

RADTF [5]] is a teaching file solution, which is compatible with RadLex. Differential diagnosis and quiz modes are available. RADTF uses RadLex anatomy concept terms and provides NLP features to process radiologic reports [5], including stemming, ranking of results based on detected negation, hedge, and uncertainty expressions. Although RADTF has been described as open-source [5], others have concluded that RADTF is not publically available [4] and therefore, we cannot evaluate the features it supports.

EURORAD (European Society of Radiology) [18] is a peer-reviewed educational tool based on teaching cases. It contains teaching files with clinical history, image findings, discussion, final diagnosis and differential diagnosis. There are a total of 6,691 teaching cases; users can search based on anatomical structure. EURORAD currently supports three different languages (English, Spanish, and French). Similar to other teaching file sources there is no support for negations,synonyms, or image-based search. Another radiological teaching file system that can be integrated into a PACS environment is RadPix [29]. Complete radiological teaching files can be created by adding text, annotations and images. Annotations are added using built-in drawing software, designed for radiologists. The current public user interface only has 11 teaching cases.

In addition to the work described so far, there were several efforts made to allow image search use text associated with the images through captions or text embedded in the images. Some of these systems proposed for the medical domain are summarized below.

The Biomedical Image Metadata Manager (BIMM) [14] system provides retrieval of similar images using semantic features of image metadata. Based on the imaging observation, 2D regions of interest are stored as metadata, and the system offers content-based image retrieval capabilities although this could not be tested as it is no longer available in the public domain [14].

Khresmoi for Everyone [13] is a medical informatics and retrieval system that provides an access system for online biomedical information and documents. The result of a search is a web link to a discussion forum about diseases and quizzes. However, there is no unified solution with clinical reports with categories, such as history of patient, findings, diagnosis, or differential diagnosis. Furthermore, the search capabilities are limited; for example, they do not support synonym- and negation-based search.

GoldMiner [25] helps users search images and articles from peer-reviewed biomedical journals. It uses the National Library of Medicine to discover medical concepts in figure captions with the final goal of retrieving relevant images. GoldMiner recognizes abbreviations, synonyms, and types of diseases but search results only depend on the explicit presence of specific words in the figure captions.

Yottalook [26] is a radiologist-targeted search engine powered by Google Custom Search that searches a variety of sources such as radiopaedia.org, American Journal of Radiology, University of Michigan Medical School, and MyPacs. It provides users with the ability to choose the category of search (e.g., CT, ultrasound). Although Yottalook claims that aggregates multiple sources, the data sources seem to not be truly integrated – users who choose a search category (e.g., X-ray) are then redirected to the original

source in its specific format (e.g., an external webpage or PowerPoint file).

4 RESULTS

Our preliminary results are based on the MIRC and MyPacs data integration as well as RadLex, SNOMED, and Oxford medical dictionary query expansion. We compared the built-in search engines that are associated with MIRC and MyPacs, the Google search engine, and our IRIS search engine based on several types of queries formulated by expert radiologists and sample queries from related work by De-Arteaga et al. [4] in which the in-house RadTF search engine was compared with Goldminer.

Furthermore, as Google does not offer integrated search for teaching file repositories, we performed custom searches (e.g., Google search for "renal site:www.mypacs.net" to find content related to "renal" in MyPacs.net teaching file repository). This search was chosen over a general Google search because the term "renal" can return a wide variety of results such as Power Point presentations, videos, PDF documents mixed in with the teaching files.

4.1 Smart search through synonym and negation

The evaluation of our IRIS engine consists of comparing query results, including "cardiomegaly" and "no cardiomegaly" queries and compared our results with the ones produced by the MIRC and MyPacs search engines as well as those produced by the Google site search engine and the RSNA journal search. Our search results were significantly improved through applying synonyms and negation interpretation of the search when compared with the other searches. For example, while the built-in search engines for MIRC, MyPacs, and RSNA journals (RadioGraphics and Radiology) retrieved the *same results* for "cardiomegaly" and "no cardiomegaly" (effectively ignoring negation) showing 56 teaching files, 108 teaching files and 519 articles respectively, our search engine differentiated between these two searches by recognizing negation and returning different answers in the negation-based query.

The IRIS engine retrieved 59 cases (3 more) from MIRC, as it also searched for "enlargement of heart" and "enlarged heart" which are synonymous with "cardiomegaly". Using the SNOMED ontology, the IRIS engine also searched for "cardiac dilation" and "congenital cardiomegaly". For MyPacs data repository, our IRIS engine retrieved 99 cases from which 91 were a subset of 108 retrieved by MyPacs search engine; one thing to note that the MyPacs search has access to all 33,000 teaching files versus about 17,000 teaching files that are freely accessible and were integrated by our IRIS search engine (i.e., we integrated roughly half of the MyPacs teaching files).

RSNA has a built-in search engine to retrieve journal articles from its database. It can retrieve results based on entered words or phrases. The result can be filtered to target research articles or case reports with the retrieval results being sorted by relevance or date. Articles with no fees required are highlighted as "free". For the "cardiomegaly" query, the RSNA journals search returned 519 articles. Our IRIS search engine found 130 articles from the subset of RSNA freely available articles. We also used Google site search as one of the alternatives to compare the search results. Google search found 24 teaching files, 72 teaching files, and 201 articles for MIRC, MyPacs, and RSNA Journal site search, respectively.

Figure 3a summarizes the comparative performance of different searches with "cardiomegaly". Our IRIS engine has 50 results that overlap with MIRC results, 91 results that overlap with MyPacs results, 130 overlapping results with RSNA journal search, and 72 overlapping results with the Google search for the same query. There are 21, 71, and 201 result case overlap, respectively, between these search engines. In general, for a simple single-term search our results are very similar to other tested search engines, with the notable difference of being returned within a single integrated search.

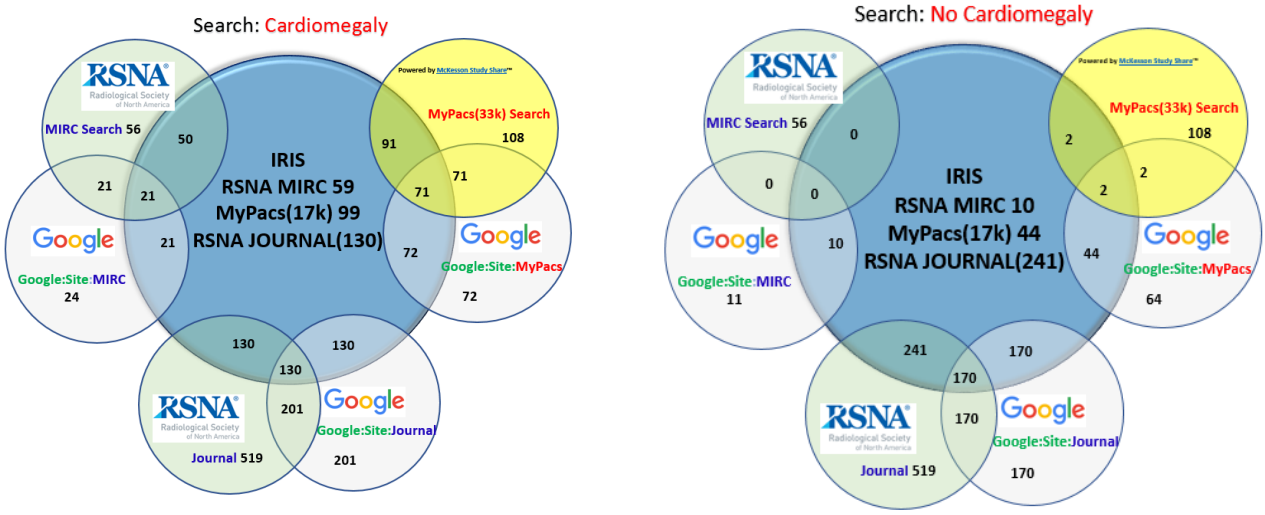
Our next search was based on negation, searching for "no cardiomegaly". Oddly, no search engine of those compared here applies the concept of negation to the search (Figure 3b). MIRC and MyPacs return the exact same teaching files as with "cardiomegaly", meaning that the negation term was not considered in the query. Using the SNOMED ontology, the IRIS search engine replaced "no cardiomegaly" with "normal heart", as "cardiomegaly" was defined as "morphologically abnormal structure of the heart" according to the ontology. IRIS also extended the search with "normal heart structure" and retrieved 10 results from MIRC and 44 from MyPacs. These results are clearly more relevant – negation should produce a different set of cases and must specifically *not* return "cardiomegaly" cases (which we manually verified). Only two cases from our negation-based search matched with the original "cardiomegaly" search; the overlap occurred because discussion referred to "usually normal heart" (i.e., accidental overlap that can be eliminated as a false-positive with additional analysis). Google search showed 11, 64, and 170 results for MIRC, MyPacs and RSNA Journal site search, respectively, but many results were not teaching files at all.

Figure 3b summarizes the comparative performance of different searches for "no cardiomegaly". There is no overlap between the IRIS and MIRC searches and there are two teaching files retrieved both by IRIS and MyPacs with the two teaching files mentioning "normal heart" in a different context (in addition to "cardiomegaly"). For the RSNA journal search, there was an overlap of 241 articles with IRIS and 170 articles with Google. We also looked at the results showing overlaps between MIRC, MyPacs, and RSNA journal with Google search and found 0, 2, and 170 results overlap, respectively, between these search engines (Figure 3b). We then manually inspected the results, and found that the teaching files retrieved by IRIS were more relevant to the "no cardiomegaly" query than for the other search engines.

We discuss another example query, "irregularly shaped", for which MIRC retrieved 7 teaching files and MyPacs returned 47 teaching files with quotations around the search terms. When no quotations were used, MIRC retrieved the same 7 teaching files and MyPACS returned 115 teaching files. This example shows that the search engine structures differ from one repository to another, and further motivates the idea of data integration and smart search capabilities for medical knowledge extraction. IRIS search for "irregularly shaped" used synonym-based query expansion and looked for teaching files also using the expression "abnormally shaped"; the query results produced 11 teaching files from MIRC, 29 teaching file from the MyPacs, and 197 from RSNA. Google only

Table 1: A comparative study of available data sources and search engines

		NLP capabilities					
Search Engine	**Keyword search**	**Synonyms**	**Morphological forms**	**Relationships between terms**	**Spelling error correction**	**Relevance feedback**	**publically available**
RadTF	YES	YES	YES	YES	NO	NO	NO
GoldMiner	YES	NO	NO	NO	YES	NO	YES
Yottalook	YES	YES	NO	NO	YES	NO	YES
Google	YES	NO	NO	NO	YES	YES	YES
MIRC	YES	NO	NO	NO	NO	NO	YES
MyPacs	YES	NO	NO	NO	NO	NO	YES
Gamuts	NO	NO	NO	NO	NO	NO	YES
CTisus	YES	NO	NO	NO	YES	NO	YES
Casimage	NO	NO	NO	NO	NO	NO	NO
RadICS	NO	NO	NO	NO	NO	NO	NO
BIMM	YES	NO	NO	NO	NO	NO	NO
Radiology Teacher	YES	NO	NO	NO	NO	NO	YES
Medscape	YES	NO	NO	NO	YES	NO	YES
ImageCLEFmed	NO	NO	NO	NO	NO	NO	NO
Khresmoi	YES	YES	NO	NO	NO	NO	YES



(a) Search results for "cardiomegaly"

(b) Search results for "no cardiomegaly"

Figure 3: Comparison of negation for "cardiomegaly" (the circles represent the sets of retrieved results and their overlapping sections show the counts of teaching files in the intersection)

showed 4 results in response to this query. Ultimately, the result relevance is more important than the number of results (and the 4 results did not appear as relevant to the search as other cases).

We performed a number of different queries, including queries based on the experimental results by De-Arteaga in [4] where such queries were used to evaluate the proposed RadTF in-house search engine with Goldminer. From the results presented in Table 2, we can conclude that the MIRC, and MyPacs search engines do not recognize the difference between search terms with and without negation. For example, "hepatic adenoma" yields 7, 20, and 15(3+12) results for MIRC, MyPacs, and IRIS, respectively. Alternatively, "no hepatic adenoma" yields 7, 20, and 10 results for MIRC, MyPacs, and IRIS, respectively. Our search engine recognizes "No" in search terms and retrieves teaching files that are different from the ones matching "hepatic adenoma". IRIS search for synonyms such as "ACL tear" is augmented with "tear of ACL" (synonym from RadLex dictionary) and retrieves the same teaching cases. In this particular case, the term is sufficiently unique that even after applying stemming and adding synonyms no new cases matched the search. For "annular pancreas", IRIS returned 7 more results than a MyPacs

Table 2: Examples of queries based on De-Arteaga [4]

Query	IRIS (2k-MIRC data set)	RSNA MIRC (2k)	IRIS (33k-MyPacs data set)	MyPacs.net (33k)	Findings
ACL Tear	3	3	66	96	MIRC, MyPacs same teaching cases with and without negation
No ACL Tear	0	3	14	96	MyPacs.net search by individual term
Appendicitis	40	42	179	160	MIRC, MyPacs same teaching cases with and without negation
No Appendicitis	3	42	1	160	
Hepatic adenoma	3	7	12	20	MIRC, MyPacs same teaching cases with and without negation
No Hepatic adenoma	2	7	8	20	
Annular pancreas	14	16	35	28	2 cases, not publically available with MIRC
No Annular Pancreas	5	16	11	28	MIRC, MyPacs same teaching cases with and without negation
Toxic	48	48	166	98	MIRC, MyPacs same teaching cases with and without negation
No Toxic	12	48	2	98	

search, since IRIS expanded the query with "congenital malformation of pancreas", which is the parent class of "annular pancreas" (a relation from SNOMED) and searches for these two synonymous terms. Note that MyPacs returns more results than any other engine because it will search for individual terms even if the search phrase is in quotation marks. However, this does not warranty that the retrieved teaching files are more relevant to the query.

4.2 Adjective-based search

We also evaluated IRIS performance for adjective-based searches. We have searched for different disease categories: "severe pulmonary edema", "moderate pulmonary edema" and "mild pulmonary edema" – a MIRC search produced 19, 12, and 17 teaching file results, respectively. However, these searches included significant overlap; most of the teaching files appeared three times in all searches. Thus, most of the teaching files were returned for "severe", "moderate", and "mild" searches. On the other hand, an IRIS search produced 23 different teaching files that referenced the phrase "pulmonary edema", but it returned 3 different non-overlapping teaching file sets for "severe", "moderate" and "mild" searches, which is what the user would typically expect. MIRC engine returned 66 teaching files with "pulmonary edema". When using Google search on the teaching files repositories, the search produced 9 results for MIRC (much fewer than the native search) and 210 results for MyPacs (too many, obviously including things not directly related to pulmonary edema).

We assume that all relevant answers should involve the term "pulmonary edema" and that matching "severe" or "pulmonary" terms is not relevant to the search. Through correctly applying qualifier-based partitioning into (mild / moderate / severe) cases, our search can produce far better results than other engines. Out of 23 MIRC cases, 2 teaching files also matched the word "mild" and 6 teaching files used the word "severe", leaving 15 cases that we match to the "moderate" search by default. One of the "mild" cases referenced "mild cough" making it a false-positive match; similarly, one of the "severe" cases referenced "severe hypoxia" as a complication, matching the search by accident. The MyPacs case set produced a total of three matches, so there is not enough data to apply partitioning – all three cases defaulted to the "moderate" category. However, we automatically integrate these results into the total of 26 "pulmonary edema" cases in our search.

Ultimately we plan to be more precise by building a classifier that can partition the "pulmonary edema" cases into severity categories, distinguishing "mild" or "severe" cases. Lacking indication of either, cases can be placed in the "moderate" category. Most importantly, the search cannot, allow an overlap in results of "severe pulmonary edema" and "mild pulmonary edema" as all existing search engines currently do. We further note that this sort of partitioning makes it easy to apply negation. For example, "non-severe pulmonary edema" is to be matched by our engine to "moderate" and "mild" cases.

4.3 User study for IRIS evaluation

In order to further validate the IRIS search engine, we designed a user study to create reference truth for quantifying the search result accuracy. Our survey asked the users (non-medical researchers) to provide feedback on search results by annotating each one of the retrieved teaching file *Relevant*, *Not relevant*, or *Not sure*. Our preliminary evaluation created reference truth for the "cardiomegaly" search term from four human observers.

The four users were asked to provide their feedback on 10 randomly chosen teaching files from the set of retrieved teaching files in response to the "cardiomegaly" query. From the analysis of feedback received, we found that 80% of IRIS results were *Relevant*, 10% of the results were labeled *"not sure"* and 10% of the results labeled were *"not relevant"*. We looked into the *not relevant* results and found that those teaching files include "cardiomegaly" in the discussion category, which was not a direct reference to the case in the corresponding teaching file (e.g., "aside from the findings demonstrated in this patient, patients with progeria can also have "cardiomegaly", kyphoscoliosis, acroosteolysis, ...").

5 CONCLUSION AND FUTURE WORK

Currently, radiologists have access to an internal search engine that helps them access the internal teaching files available at their hospital. The functionality of internal hospital engines is limited

to the in-house teaching files and lacks advanced and analytical search capabilities. To our knowledge, none of the available engines provide radiologists with the ability to integrate in-house data with other data sources, even if these sources are publicly available and are rich om medical content.

In this paper, we described a database and search framework for heterogeneous data integration to facilitate medical knowledge extraction from publically available teaching file repositories. The IRIS engine supports negation, query expansion through synonyms, and recognition of medically relevant adjectives related to the degree of disease severity. Based on our preliminary evaluation, we found that the IRIS engine improves results in two ways: 1) search queries produce more relevant results compared with the existing search tools, and 2) results from multiple data sources can be merged into a single easy-to-query and interpret data source. Although our current implementation uses text-based search for images and teaching files, the existing integrated database schema makes it possible to incorporate image-based search using image features derived from pixel content as part of our future work. Furthermore, we also plan to add another functionality to the IRIS search through which radiologists can mark regions of interest and annotations on images from relevant teaching files. In the long run, the IRIS search engine will provide capabilities for both text and image search and will be tailor for a variety of users, including radiologists, radiology residents, clinicians, and patients.

REFERENCES

[1] Armato, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., et al. (2011). The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931.

[2] Bhargava, P., Dhand, S., Lackey, A. E., Pandey, T., Moshiri, M., and Jambhekar, K. (2013). Radiology education 2.0—on the cusp of change: part 2. ebooks; file sharing and synchronization tools; websites/teaching files; reference management tools and note taking applications. *Academic radiology*, 20(3):373–381.

[3] Dashevsky, B., Gorovoy, M., Weadock, W. J., and Juluru, K. (2015). Radiology teaching files: an assessment of their role and desired features based on a national survey. *Journal of digital imaging*, 28(4):389–398.

[4] De-Arteaga, M., Eggel, I., Do, B., Rubin, D., Kahn, C. E., and Müller, H. (2015). Comparing image search behaviour in the arrs goldminer search engine and a clinical pacs/ris. *Journal of biomedical informatics*, 56:57–64.

[5] Do, B. H., Wu, A., Biswal, S., Kamaya, A., and Rubin, D. L. (2010). Informatics in radiology: Radtf: A semantic search–enabled, natural language processor–generated radiology teaching file 1. *Radiographics*, 30(7):2039–2048.

[6] Dos-Santos, M. and Fujino, A. (2012). Interactive radiology teaching file system: the development of a mirc-compliant and user-centered e-learning resource. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 5871–5874. IEEE.

[7] Group, M. M. I. (May 31, 2017). Mypacs tfs. https://www.mypacs.net/.

[8] Gutmark, R., Halsted, M. J., Perry, L., and Gold, G. (2007). Use of computer databases to reduce radiograph reading errors. *Journal of the American College of Radiology*, 4(1):65–68.

[9] Hospital, J. H. (May 31, 2017). Ctisus. http://www.ctisus.com/.

[10] Hwang, K. H., Lee, H., Koh, G., Willrett, D., and Rubin, D. L. (2016). Building and querying rdf/owl database of semantically annotated nuclear medicine images. *Journal of Digital Imaging*, pages 1–7.

[11] Jones, D. J. (May 31, 2017). Radiopaedia. https://radiopaedia.org/.

[12] Kansagra, A. P., John-Paul, J. Y., Chatterjee, A. R., Lenchik, L., Chow, D. S., Prater, A. B., Yeh, J., Doshi, A. M., Hawkins, C. M., Heilbrun, M. E., et al. (2016). Big data and the future of radiology informatics. *Academic radiology*, 23(1):30–42.

[13] khresmoi (June 28, 2017). khresmoi. http://everyone.khresmoi.eu/hon-search/.

[14] Korenblum, D., Rubin, D., Napel, S., Rodriguez, C., and Beaulieu, C. (2011). Managing biomedical image metadata for search and retrieval of similar images. *Journal of digital imaging*, 24(4):739–748.

[15] LLC, W. (May 31, 2017). medscape. http://www.medscape.com.

[16] Margolies, L. R., Pandey, G., Horowitz, E. R., and Mendelson, D. S. (2016). Breast imaging in the era of big data: structured reporting and data mining. *American Journal of Roentgenology*, 206(2):259–264.

[17] Müller, H., Clough, P., Deselaers, T., Caputo, B., and CLEF, I. (2010). Experimental evaluation in visual information retrieval. *The Information Retrieval Series*, 32.

[18] Neutorgasse, E. (May 31, 2017). Eurorad. http://www.eurorad.org/.

[19] Organization, S. I. I. H. T. S. D. (May 31, 2017). Snomedct ontology. http://www.snomed.org/.

[20] Reeder, M. M. (May 31, 2017). Gamuts. http://gamuts.isradiology.org/Gamuts.htm.

[21] RSNA (May 31, 2017a). Radlex ontology. http://www.radlex.org/.

[22] RSNA (May 31, 2017b). Rsna tfs. http://mirc.rsna.org/query.

[23] Seitz, J., Schubert, S., Völk, M., Scheibl, K., Paetzel, C., Schreyer, A., Djavidani, B., Feuerbach, S., and Strotzer, M. (2003). Evaluation radiologischer lernprogramme im internet. *Der Radiologe*, 43(1):66–76.

[24] SNOMED (July 24, 2017). Snomednlm. https://www.nlm.nih.gov/healthit/snomedct/index.html.

[25] Society, T. A. R. R. (May 31, 2017). Goldminer. http://goldminer.arrs.org/.

[26] Solutions, M. H. (May 31, 2017). Yottalook. http://www.yottalook.com/.

[27] Talanow, R. (2009). Radiology teacher: a free, internet-based radiology teaching file server. *Journal of the American College of Radiology*, 6(12):871–875.

[28] Thies, C., Güld, M. O., Fischer, B., and Lehmann, T. M. (2004). Content-based queries on the casimage database within the irma framework. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 781–792. Springer.

[29] Weadock Software, L. (May 31, 2017). Radpix. http://radpix.com/.

[30] Zinovev, D., Raicu, D., Furst, J., and Armato III, S. G. (2009). Predicting radiological panel opinions using a panel of machine learning classifiers. *Algorithms*, 2(4):1473–1502.