Journal of Digital Imaging (2018) 31:596–603
https://doi.org/10.1007/s10278-018-0064-0

Proposing New RadLex Terms by Analyzing Free-Text Mammography Reports

Hakan Bulu[1] · Dorothy A. Sippo[2] · Janie M. Lee[3] · Elizabeth S. Burnside[4] · Daniel L. Rubin[1]

Published online: 20 March 2018

Abstract
After years of development, the RadLex terminology contains a large set of controlled terms for the radiology domain, but gaps still exist. We developed a data-driven approach to discover new terms for RadLex by mining a large corpus of radiology reports using natural language processing (NLP) methods. Our system, developed for mammography, discovers new candidate terms by analyzing noun phrases in free-text reports to extend the mammography part of RadLex. Our NLP system extracts noun phrases from free-text mammography reports and classifies these noun phrases as "Has Candidate RadLex Term" or "Does Not Have Candidate RadLex Term." We tested the performance of our algorithm using 100 free-text mammography reports. An expert radiologist determined the true positive and true negative RadLex candidate terms. We calculated precision/positive predictive value and recall/sensitivity metrics to judge the system's performance. Finally, to identify new candidate terms for enhancing RadLex, we applied our NLP method to 270,540 free-text mammography reports obtained from three academic institutions. Our method demonstrated precision/positive predictive value of 0.77 (159/206 terms) and a recall/sensitivity of 0.94 (159/170 terms). The overall accuracy of the system is 0.80 (235/293 terms). When we ran our system on the set of 270,540 reports, it found 31,800 unique noun phrases that are potential candidates for RadLex. Our data-driven approach to mining radiology reports can identify new candidate terms for expanding the breast imaging lexicon portion of RadLex and may be a useful approach for discovering new candidate terms from other radiology domains.

Keywords Breast imaging · Informatics · Mammography · Natural language processing · Ontology · RadLex

Introduction

There are many terminologies in the medical domain, but no single terminology serves all clinical needs, particularly in specialized domains such as radiology. RadLex (http://radlex.org) is a single unified source of radiology terms that is designed to fill this need for the radiology domain. Beginning in 2005, the Radiological Society of North America, an international society of radiologists, medical physicists, and other medical professionals convened experts in imaging informatics and radiological subspecialties to create this resource, which is freely available to the public.

RadLex is a structured radiology-specific ontology, which currently includes more than 75,000 terms [1]. Since its inception, RadLex has continued to expand as the radiology community has identified gaps and contributed new terms. [2, 3]. RadLex was built in a top-down manner, in which experts assembled the terminology based on their radiological knowledge and review of existing terminology and knowledge sources. However, this manual approach is resource-intensive and incomplete, especially for collecting synonyms or other alternative forms of preferred terms.

The American College of Radiology's Breast Imaging Reporting and Data System (ACR BI-RADS®), created in 1992, provided a mammography lexicon to describe lesion

Hakan Bulu and Dorothy A. Sippo contributed equally to this work.

✉ Dorothy A. Sippo
dsippo@mgh.harvard.edu

1 Department of Radiology and Department of Biomedical Data Science, Medical School Office Building (MSOB), Stanford University, 1265 Welch Road, X383, Stanford, CA 94305-5464, USA

2 Department of Radiology, Avon Comprehensive Breast Evaluation Center, Massachusetts General Hospital, Wang Ambulatory Care Building, Suite 240, 15 Parkman Street, Boston, MA 02114, USA

3 Department of Radiology, Seattle Cancer Care Alliance, University of Washington, 825 Eastlake Avenue East, Suite G2-600, Seattle, WA 98109, USA

4 Department of Radiology, E3/311 Clinical Science Center, University of Wisconsin School of Medicine and Public Health, 600 Highland Avenue, Madison, WI 53792-3252, USA

features and characteristics to be used in standardized reporting and communication about clinical management of breast lesions [4]. The ACR BI-RADS Atlas was widely adopted and is now in its fifth edition [5]. The initial mammography lexicon has been extended to include breast ultrasound and MRI. The ACR BI-RADS committee has an ongoing evidence-based process to evaluate and revise the breast imaging lexicon to add or eliminate terms [6].

RadLex focuses on compiling a comprehensive set of terms for radiology and making the relationships among terms explicit. Within RadLex, the well-established BI-RADS lexicon terms are mapped in an ontology, which has the potential to facilitate powerful tools to support clinical structured reporting and research efforts. Natural language processing (NLP) of imaging reports offers an automated approach to identifying new candidate terms for RadLex. The purpose of our study was to use a large corpus of multi-institutional mammography reports to develop, evaluate, and apply an NLP algorithm to identify candidate terms for RadLex. Our aim was to assemble as comprehensive a list as possible of terms used in clinical practice.

Material and Methods

The development and evaluation of our NLP algorithm was conducted with Institutional Review Board (IRB) approval of this HIPAA-compliant study. Informed consent was not required to access the mammography reports used with our algorithm. This was because there were no direct identifiers associated with this data, thereby minimizing any risk (specifically, the risk to patient confidentiality). We applied our NLP method to 270,540 free-text mammography reports obtained from three independent academic institutions to identify candidates for enhancing RadLex. Two of the institutions are in the Midwest and one is on the West Coast of the USA. We compiled the candidates into a list, ordered by frequency, for review by the RadLex curation staff, who maintain the ontology and oversee periodic updates to it, in coordination with radiologist members of the RadLex Breast Subcommittee. An overview of our NLP system for identifying candidate terms for RadLex by mining radiology reports is shown in Fig. 1.

1. Generating a Noun Phrase List from Free-Text Mammography Reports

GATE (General Architecture for Text Engineering) [7] is a Java suite of tools for NLP development. We use GATE's Noun Phrase Chunker [8] to extract the noun phrases and adjectives from free-text mammography reports and collect them in a list called the Initial Noun Phrase List. Figure 2 provides an example list of noun phrases for a given free-text mammography report. The challenge is to eliminate from this initial list those terms that are unlikely to be RadLex candidates ("noisy terms").

2. Pre-Processing

Removing All Non-Letter Characters

We assume that a RadLex term contains only letters (no numbers or punctuation). Thus, in this step, the system removes all non-letter characters from each noun phrase in the Initial Noun Phrase List. Rather than discarding the entire noun phrase, if one of its terms contains a non-letter character, the system continues searching within the noun phrase for smaller noun phrases that may be candidate terms. Example inputs and outputs from this pre-processing are shown in Table 1. For example, for the noun phrase "2 -," after this pre-processing step, no noun phrase remains, so that noun phrase is removed from the Initial Noun Phrase List.

Removing Too Short and Too Long Noun Phrases

After the system removes the non-letter characters from the Initial Noun Phrase List, the resulting noun phrase list contains noun phrases that are either too short or too long to be likely RadLex term candidates. We define two thresholds for the length of a RadLex term candidate: Total Character (TC) denotes a minimum allowed character count for an individual candidate noun phrase and Total Word (TW) denotes a maximum allowed term (word) count in an individual candidate noun phrase. We assume that if a noun phrase contains RadLex term(s), it cannot be shorter than two characters and cannot contain more than eight terms (based on reviewing current terms in RadLex). We set the TC equal to 2 and TW equal to 8 in our system. The threshold values can be iteratively adjusted, depending on the dataset and domain. Some example noun phrases are shown in Table 2.

Removing Noisy Terms Located at the Beginning of the Noun Phrase

The system removes noisy term(s) such as "a, the, and" located at the beginning of the noun phrase. These noisy terms are referred to as "stop words." Some example noun phrases are listed in Table 3. Details about how noisy terms are identified are given in the following Section 5, "Removing Noisy Terms."

Removing Existing RadLex Terms

Since the goal is to identify new RadLex candidate terms, the system removes all existing RadLex terms from our noun phrase list. In this process, the system only removes those noun phrases, which exactly match an existing RadLex term. It does not remove plural or singular versions of the noun phrases, since it is

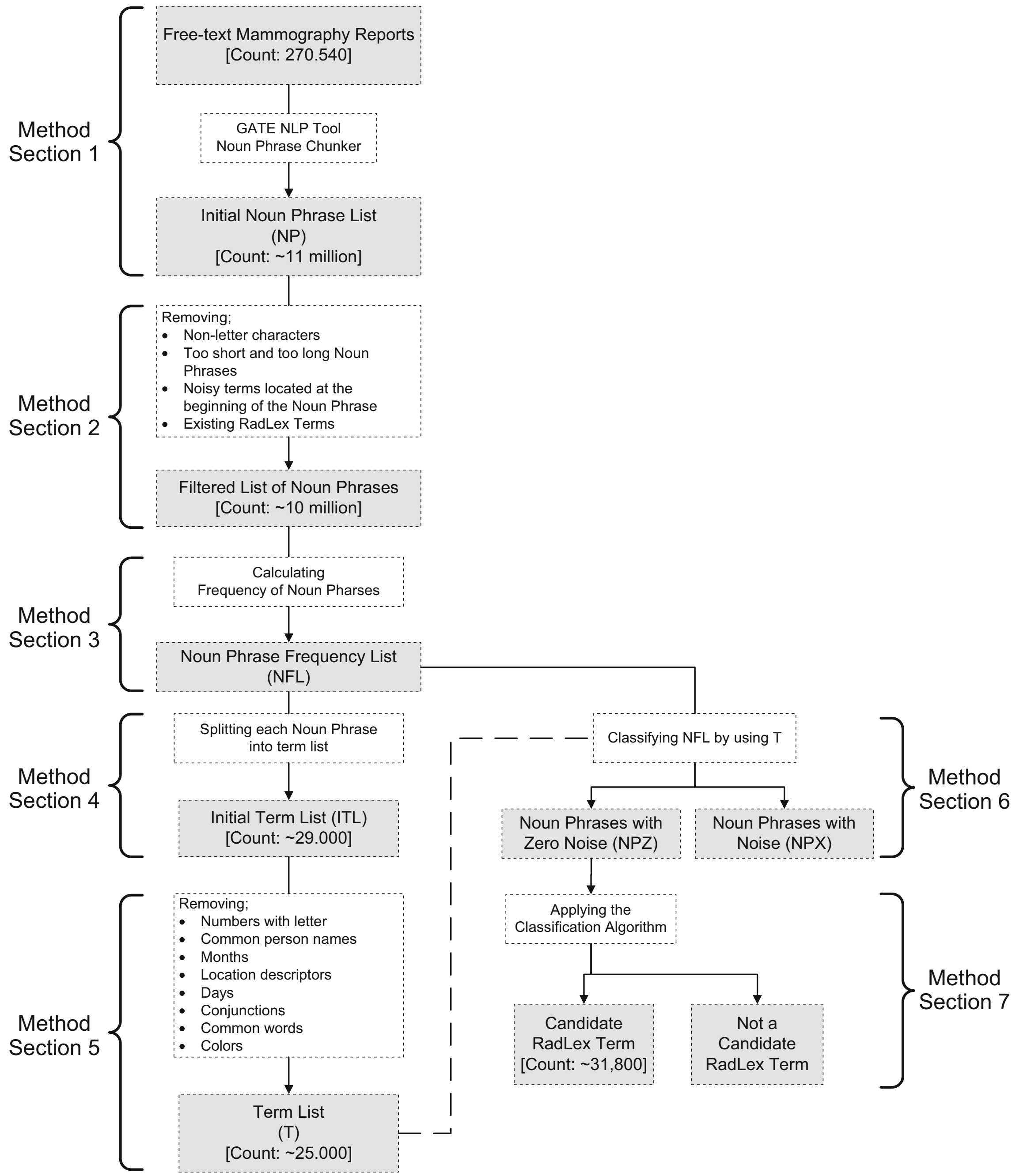

Fig. 1 Overview of the NLP system to identify candidate RadLex terms. (Method Section 6 uses the output of both Method Sections 3 and 5.)

possible that both singular and plural forms of a term may be appropriate for RadLex. For example, the term "calcification" is placed in the RadLex term "Skin calcification" (RID34252) as singular while the RadLex term "Intraductal calcifications" (RID49682) is plural. The system also does not remove the noun phrases that are a combination of existing RadLex terms, since a RadLex term can be a pre-coordination of more atomic RadLex terms. For example, "benign round mass" would not be removed from the Initial Noun Phrase List, even though RadLex contains "round," "mass," and "round mass," since "benign round mass" is not in RadLex. As another example, the RadLex term "Spiculated margin" (RID5713) consists of the RadLex terms "Spiculated" (RID34284) and "Margin" (RID5972).

3. Calculating Frequency of Noun Phrases

We assume that if a particular noun phrase is seen in the mammography reports more frequently than the others, it is more likely to be a candidate RadLex term. Thus, the system

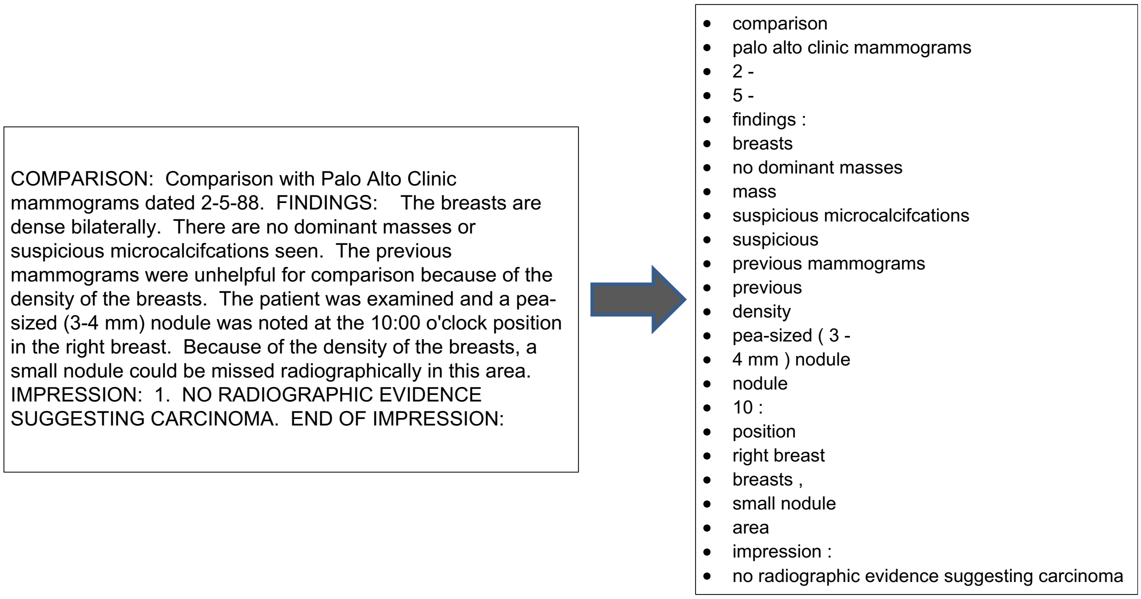

Fig. 2 Example initial noun phrase and adjective list for a given free-text mammography report. Many terms in this initial noun phrase list are not relevant to the identification of candidate RadLex terms

creates a list that contains noun phrases and their frequencies. The list is called the Noun Phrase Frequency List (NFL).

4. Building Term Frequency List

In this step, the system splits each noun phrase from the NFL into atomic terms (single words) and collects the atoms in a list with their frequencies. For example, the noun phrase "focal area" is split into atoms "focal" and "area." We call this list the Initial Term List (ITL). The high-frequency atoms from this list are shown in Table 4.

5. Removing Noisy Terms

The Initial Term List usually contains many terms that are not relevant in the context of RadLex. We call these terms "noisy terms" (examples shown in Table 4) and include stop words such as "and," location descriptors, the names of persons, and other common, non-domain terms. To eliminate names of persons, the system compares each term with a dictionary of approximately 5500 person names [9]. It also compares the list of candidate terms with 1000 common English words and 90 common conjunctions to remove additional non-domain terms. In addition, our system also considers the location descriptors as noisy. While the location descriptors such as "left" and "right" are RadLex Terms, they can be used to describe another RadLex Term, for example "nipple retraction in *left* breast." In this study, we do not consider a location descriptor to be a candidate RadLex Term, so we added all possible location descriptors into the noise-filtering list. A complete list of the noise control groups includes:

- Numbers; {one, two, hundred etc.}
- Common person names; {addison, adale, george etc.}
- Months; {january, february etc.}
- Days; {sunday, monday etc.}
- Location descriptors; {left, right, inner, outer etc.}
- Conjunctions; {after, also, for etc.}
- Colors; {black, white, yellow etc.}
- Common English words; {the, a, an, be, have etc.}
- Common medical words; {room, patient, doctor etc.}

The output list from this step is sorted from high to low frequency, and we refer to this final term list as "T."

Table 1 Removing non-letter characters

Before	After
2 -	REMOVED
pea-sized (3–	pea sized
4 mm) nodule	mm nodule

6. Selecting Noun Phrases Which Do Not Contain Noisy Terms

In this step, we select noun phrases from the NFL which only contain terms from "term list" (T). These noun phrases are put in a new list called "Noun Phrase List with Zero Noise" (NPZ). Noun phrases from the NFL that contain noisy terms are excluded in the list called "Noun Phrases with Noise" (NPX). Some example noun phrases from the NPZ and NPX lists are given in Table 5 with their frequencies. A classification algorithm is then performed on the NPZ list.

Table 2 Sample noun phrases that were removed for being too short or too long

Noun Phrase	Decision
M	Too short
Dx	Too short
shs breast invasive carcinoma summary site left breast specimen type wire localized lumpectomy invasive carcinoma type ductal histologic grade composite	Too long

7. Classification Algorithm

In this stage of processing, a module classifies noun phrases from the NPZ list as "Candidate RadLex Term" or "Not a Candidate RadLex Term." To accomplish this, we analyze the frequency of the noun phrases and the frequency of their constituent term(s). The classification algorithm takes two inputs; term list, T, and noun phrase list, NPZ (Fig. 1). Terms in these lists are sorted from high to low frequency. Furthermore, the algorithm applies thresholds to the frequencies of the terms and the noun phrases to consider a noun phrase to be a candidate term for RadLex. If the frequency of the terms or noun phrase is lower than the corresponding threshold, that noun phrase is discarded by the algorithm. Depending on the size of the dataset and/or the domain of interest, the value of these thresholds can be changed. In other words, for larger and more diverse sets of radiology reports, the frequency thresholds can be set higher to employ a more stringent selection process. The thresholds are denoted as tF and tNP, respectively.

For each T term, our algorithm first checks its frequency. If the frequency of the term is higher than the tF threshold, the algorithm searches for a noun phrase in the NPZ list that contains that term. The search operation is performed in sequential order from highest to lowest frequency noun phrases. The noun phrase with the highest frequency that contains the T term is then analyzed. If the frequency for that noun phrase is higher than the tNP threshold, that noun phrase is moved from the NPZ list to the Candidate RadLex Term True list and the T term is removed from the T term list. To be selected as a Candidate RadLex Term, all the noun phrase's constituent T terms must have frequencies higher than the tF threshold. For example, if the T term is "breast," the algorithm searches the NPZ list for this term and finds the noun phrase "breast tissue." If the frequencies for breast, tissue, and breast tissue are higher than the respective tF and tNP thresholds, then breast tissue is added to the True list. Breast tissue is removed from the NPZ list and the terms breast and tissue are removed from the T term list.

Table 3 Sample stop words that were removed from the beginning of noun phrases

Before	After
a duct	duct
a focal area	focal area
a new large area	large area
the palpable mass	palpable mass

8. Evaluation

To evaluate the performance of the NLP algorithm, we randomly selected 100 mammography reports from our report database. We developed an online evaluation tool to record a gold standard for these reports (Fig. 3). This tool presents the text report and a list of all noun phrases identified by the NLP algorithm in the report. An expert radiologist, fellowship trained in breast imaging, reviewed the report and recorded the noun phrases that are candidate terms for RadLex. New terms could be added that were not extracted by the NLP algorithm. The annotations created by the expert using this tool established the gold standard from which we calculated the True Positive, True Negative, False Positive, and False Negative rates by comparing results from our algorithm to this gold standard.

9. Statistical and Analytics Methods

We tested the performance of our algorithm using 100 free-text mammography reports in which an expert radiologist determined the true positive RadLex candidate noun phrases ("Has Candidate RadLex Term") and the true negative noun phrases ("Does Not Have Candidate RadLex Term"). We calculated precision and recall, which are also known as positive predictive value and sensitivity, to evaluate our system's performance. Precision/positive predictive value was defined as true positive noun phrases identified by the NLP algorithm divided by all positive noun phrases identified by the

Table 4 Terms with their frequencies

Term	Frequency	Noisy term?
breast	42,078	No
and	38,695	Yes
mass	25,891	No
right	24,659	Yes
biopsy	22,937	No
left	21,910	Yes
ultrasound	15,072	No
benign	12,915	No
calcifications	12,402	No
mammogram	11,401	No
bilateral	10,572	Yes
enhancement	10,407	No
tissue	10,162	No

Table 5 Examples noise-free noun phrases and examples of noun phrases with noise (noisy terms in italics)

Noise-free noun phrases	Frequency	Noun phrases with a noise	Frequency
malignancy	154,989	malignancy *on*	128
mass	141,919	*no* masses	8698
images	109,802	*separate* images	2537
calcifications	108,565	*these* calcifications	3363
benign	102,240	benign *end*	925
architectural distortion	96,707	*new* architectural distortion	370
palpable abnormality	8905	*any* palpable abnormality	172
round mass	195	*cm* round mass	105

algorithm (both true and false positives). Recall/sensitivity was defined as true positive noun phrases identified by the algorithm divided by all positive noun phrases identified by the expert radiologist (true positives and false negatives of the algorithm). Overall accuracy was also determined.

Results

Evaluation Results

Within the 100 mammography reports used for our evaluation, our algorithm identified approximately 3000 noun phrases. Our algorithm correctly retrieved RadLex candidates identified by the expert radiologist with a precision 0.77 (159/206) and a recall 0.94 (159/170). The overall accuracy of our system was 0.80 (235/293).

Term Discovery for RadLex

The system identified more than 11 million initial noun phrases. It removed about one million noisy noun phrases (9.1%) from the Initial Noun Phrase List (NP) and approximately 4000 noisy terms (13.7%) from the Initial Term List (ITL) in the noise reduction steps. These steps are described in "Materials and Methods" sections "Pre-processing" and "Removing noisy terms" and outlined in Fig. 1. The final system output was approximately 31,800 unique noun phrases that are candidates for RadLex. These were submitted to the RadLex curation staff for review and possible incorporation into RadLex.

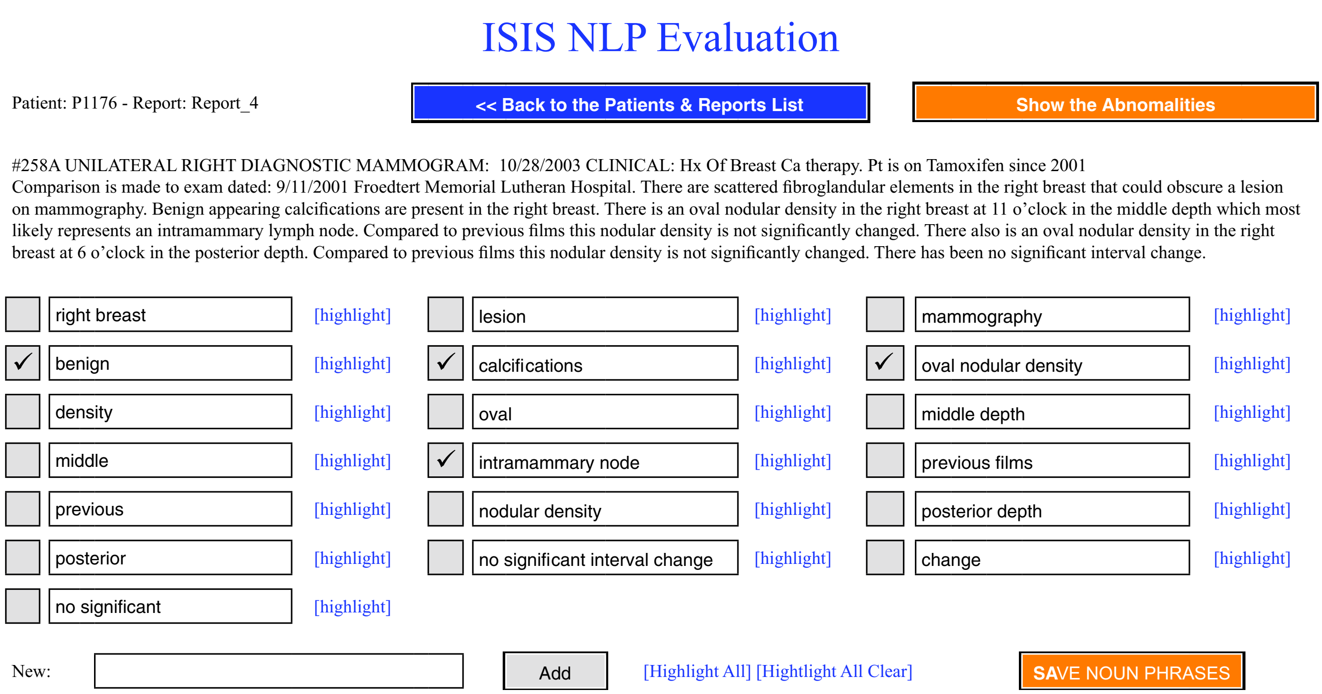

Fig. 3 Tool for establishing the gold standard for our study. The tool presents the report text, the candidate RadLex terms it discovered, and the expert puts a check mark next to those terms that are actual RadLex candidates (true positives). The expert can also record terms not found by the tool (false negatives) by entering them in the text box at the bottom of the screen

Table 6 Example candidate RadLex terms identified by our NLP algorithm

Frequency	Count	Example
More than 5000	11	tissue, density, palpable, marker, etc.
Between 3000 and 5000	30	region, abnormality, stereotactic, lumpectomy, etc.
Between 1000 and 3000	110	dominant, discrete, asymmetric, clip, etc.
Between 500 and 1000	137	lump, echogenic, discharge, post lumpectomy, etc.
Between 100 and 500	840	surgical excision, fibroadenomas, debris, nodularity, etc.
Less than 100	30,698	peri areolar, swelling, postradiation, lucency, etc.

Because of the large number of noun phrases, we sorted them from high frequency to low frequency. We believe the frequency values will help with the determination of whether a noun phrase should be added to RadLex. Some example candidate RadLex terms from our system are given in Table 6 with their frequency values.

Discussion

In this study, we present an automated approach using NLP to identify candidate terms to expand RadLex. This approach enabled analysis of more than 11 million noun phrases obtained from 270,540 screening and diagnostic mammography reports to identify new candidate terms. RadLex is a comprehensive terminology covering all domains in radiology, created via a "top-down" approach with experts compiling the terminology based on their knowledge and other terminology sources. RadLex is designed to be a comprehensive knowledge resource to enable informatics applications in radiology. As such, it is important for RadLex to include robust coverage of terms frequently used in practice. Within the breast imaging domain, good sources for these terms are reports created in clinical practice that may not yet be included in ACR BI-RADS® and RadLex.

Studies have shown that RadLex does not provide complete coverage of radiology terminology. In 2008, Marwede et al. validated RadLex against terms found in thoracic CT reports by analyzing 200 thoracic CT reports [10]. They extracted 363 distinct terms from the reports and found that 59 (16%) of these terms were not listed in RadLex. In 2012, Hong et al. extracted 6489 reporting elements from 70 radiology reporting templates from the RSNA Reporting Template Library. One third (832) of these reporting terms were not included in RadLex [11]. In 2013, Woods and Eng conducted a study to estimate the completeness of RadLex in the chest radiography domain and analyzed 100 chest radiograph reports. They showed that despite the large number of terms in RadLex, terms are still absent and complexities in the definitions of terms exist [3]. These prior approaches were based on review of existing data sources and they are difficult to scale. In 2011, Hazen et al. developed an automated system to extract image observations and observation characteristics from 1128 journal articles for possible inclusion in RadLex [12]. We similarly employed an automated process, NLP, to search a large collection of clinical reports to discover terms that may fill gaps in the RadLex terminology.

Our approach is based on analyzing the frequency of the noun phrases in mammography reports. Mammography reports contain many common noun phrases, some of which may be suitable for being candidate RadLex terms, and term frequency could be useful for identifying clinically meaningful candidate RadLex terms. The performance of the training algorithm is related to the values of the pre-defined threshold values. These values could be tailored in the future to optimize the trade-off between sensitivity and specificity for the needs of specific NLP applications. The recall of the algorithm was much greater than its precision. Ideally, higher precision would be preferable. However, we developed the algorithm to optimize recall with the primary goal of identifying potential candidate terms. Candidate terms not suitable for inclusion in RadLex could be identified by the RadLex Breast Subcommittee and RadLex curators reviewing the candidate terms. As a next step, the algorithm's performance could be further optimized by using a larger training set of radiology reports.

The limitations of our NLP approach include the relatively large number of noun phrases (approximately 31,800) that were identified as candidates for RadLex. The frequencies of noun phrases could be used to prioritize review to more commonly used phrases. Still, the RadLex curation staff and radiologist members of the RadLex Breast Subcommittee could find vetting all the terms a daunting task. There is the potential that NLP could be further applied to the list of candidate RadLex terms to consolidate redundant noun phrases or organize them. This could help expedite the human-review process. Another limitation is that our NLP algorithm was developed for mammography reports which tend to follow a relatively standardized structure. NLP may prove less successful if applied to other radiology domains where reports are not as consistently structured. Our results may not be as readily applicable to those domains.

Though we have focused this work on mammography, we believe our approach may be extensible to other domains in radiology. The number and domains of ACR Reporting and Data Systems are growing [13]. Consideration will be given to the incorporating terminology from these Systems into RadLex. As automated methods, such as NLP, mature for identifying terms commonly used in clinical practice, these approaches may also help inform the further development of Reporting and Data Systems. Partnership between the Radiological Society of North America and the American College of Radiology has enabled harmonization across RadLex and the BI-RADS lexicon [6]. This successful collaboration also provides a template for other Reporting and Data Systems to merge structured reporting and ontologies to provide powerful, reusable tools that support accurate and standardized application of imaging terminology. Establishing a process for new candidate terms identified in RadLex to be further evaluated for inclusion in the clinical BI-RADS lexicon is an important step to further strengthen and support clinical practice and research initiatives.

Conclusion

We developed a data-driven approach to identify candidate terms for expanding RadLex in the breast imaging domain by applying NLP methods to mine free-text mammography reports. Our system performed well with high recall and reasonably high precision, making it potentially useful for curating RadLex. By applying it to a large corpus of reports, we have already identified approximately 31,800 new potential candidate terms, which will now be reviewed by the RadLex Breast Subcommittee and RadLex curators. Our methodology could help to improve RadLex not only in mammography but in other radiology domains as well.

Funding The Radiological Society of North America provided grant support for this study.

Compliance with Ethical Standards

The development and evaluation of our NLP algorithm was conducted with Institutional Review Board (IRB) approval of this HIPAA-compliant study.

Disclosures Dorothy A. Sippo, MD, MPH, is the recipient of a research grant from GE Healthcare. Janie M. Lee, MD, MSc, is the recipient of a research grant from GE Healthcare and has a consulting agreement with GE Healthcare. Other authors have nothing to disclose.

References

1. RadLex. Available at: https://www.rsna.org/RadLex.aspx. Accessed 1 November 2017
2. Beitia AO, Kuperman G, Delman BN, Shapiro JS: Assessing the performance of LOINC® and RadLex for coverage of CT scans across three sites in a health information exchange. AMIA Annu Symp Proc 94–102, Nov 16, 2013
3. Woods RW, Eng J: Evaluating the completeness of RadLex in the chest radiography domain. Acad Radiol 20(11):1329–1333, 2013
4. American College of Radiology: Breast Imaging Reporting and Data System (BI-RADS). Reston, VA: American College of Radiology, 1992
5. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA et al.: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. American College of Radiology: Reston, VA, 2013
6. Burnside ES, Sickles EA, Bassett LW, Rubin DL et al.: The ACR BI-RADS experience: learning from history. J Am Coll Radiol 6(12):851–860, Dec 2009
7. Cunningham H: GATE, a general architecture for text engineering. Computers and the Humanities 36:223–254, 2002
8. Ramshaw L, Marcus M: Text Chunking Using Transformation-Based Learning. In Proceedings of the Third ACL Workshop on Very Large Corpora, 1995
9. Etymology and History of First Names. Available at: http://www.behindthename.com. Accessed 1 November 2017
10. Marwede D, Schulz T, Kahn T: Indexing thoracic CT reports using a preliminary version of a standardized radiological lexicon (RadLex). J Digit Imaging 21(4):363–370, 2008
11. 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
12. Hazen R, Van Esbroeck AP, Mongkolwat P, Channin DS: Automatic Extraction of Concepts to Extend RadLex. J Digit Imaging 24(1):165–169, 2011
13. Reporting and Data Systems (RADS). Available at: https://www.acr.org/Quality-Safety/RADS. Accessed 1 November 2017