Journal of Digital Imaging

Evaluation of Negation and Uncertainty Detection and its Impact on Precision and Recall in Search

Andrew S. Wu,[1] Bao H. Do,[2] Jinsuh Kim,[1] and Daniel L. Rubin[3]

Radiology reports contain information that can be mined using a search engine for teaching, research, and quality assurance purposes. Current search engines look for exact matches to the search term, but they do not differentiate between reports in which the search term appears in a positive context (i.e., being present) from those in which the search term appears in the context of negation and uncertainty. We describe RadReportMiner, a context-aware search engine, and compare its retrieval performance with a generic search engine, Google Desktop. We created a corpus of 464 radiology reports which described at least one of five findings (appendicitis, hydronephrosis, fracture, optic neuritis, and pneumonia). Each report was classified by a radiologist as positive (finding described to be present) or negative (finding described to be absent or uncertain). The same reports were then classified by RadReportMiner and Google Desktop. RadReportMiner achieved a higher precision (81%), compared with Google Desktop (27%; $p<0.0001$). RadReportMiner had a lower recall (72%) compared with Google Desktop (87%; $p=0.006$). We conclude that adding negation and uncertainty identification to a word-based radiology report search engine improves the precision of search results over a search engine that does not take this information into account. Our approach may be useful to adopt into current report retrieval systems to help radiologists to more accurately search for radiology reports.

KEY WORDS: Data extraction, data mining, databases, efficiency, natural language processing, reporting, negation

INTRODUCTION

Radiology reports are the primary work product of radiologists, and they contain a wealth of information that could be mined for teaching, research, administrative, and quality assurance purposes. Radiologists need to search radiology reports to locate specific findings and diagnoses, and there is a growing body of work to create methods to search radiology reports and to extract information from them.[1–5] A good report search engine could enable for radiology what Google has done for the World Wide Web—rapid and relevant retrieval of documents that users seek.

Several search engines for radiology reports have been described that retrieve reports based on word content.[1,3,6] These search engines look for keywords that match terms in the user's search. For example, search engines can look through collections of radiology reports and return a list of cases containing the word "appendicitis."

A challenge limiting the success of current radiology search engines is their inability to differentiate *positive* findings from those mentioned in context of *uncertainty* and *negation*. For example, three reports with impressions "No evidence of appendicitis," "Cannot completely exclude appendicitis," and "Acute appendicitis" would be retrieved in a user search for "appendicitis" even though only one of them actually refers

[1]From the Department of Radiology, University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, IA, 52242, USA.

[2]From the Department of Radiology, Body Imaging, Stanford University, 300 Pasteur Drive, Room H1307, Stanford, CA, 94305, USA.

[3]From the Department of Radiology, Stanford University, Richard M. Lucas Center, 1201 Welch Road, Office P285, Stanford, CA, 94305, USA.

Correspondence to: Daniel L. Rubin, Department of Radiology, Stanford University, Richard M. Lucas Center, 1201 Welch Road, Office P285, Stanford, CA, 94305, USA; tel: +1-650-7255693; fax: +1-650-7235795; e-mail: dlrubin@stanford.edu

Online publication 10 November 2009
doi: 10.1007/s10278-009-9250-4

to a case of appendicitis. Thus, only one of these three retrieved reports would be relevant to the radiologist's search; the other two types of reports are false-positive documents. These false positives clutter the search results and reduce efficiency of search. To circumvent this challenge, radiology report search engines would need to detect the *context* of findings in reports—positive, negated, or uncertain findings.

Considerable work has been done on assertion classification,[7–10] some of which used subsets of radiology reports to test their systems,[2,7] but few have focused specifically on negations in unstructured radiology reports.[11,12] In addition, little work has been done to date in applying such methods to improve the precision of searches of radiology reports.

Our goal is to develop a method to improve the precision of searches of radiology reports by determining the *context* of terms in the text. Specifically, our goal is to develop a system that detects negated and uncertain findings and uses this information to exclude the corresponding reports from the search results. We describe RadReportMiner, a context-aware search engine, and compare its performance with a currently available keyword-based method of search.

METHODS

Following institutional review board approval, 238,952 radiology reports for clinical exams performed between January 1, 2006 and December 31, 2006 were acquired from the radiology information system and imported into a MySQL database (MySQL version 14.12 distribution 5.0.45 for Win32, running on Windows XP and Apache HTTP web server 2.2.6). The front-end user interface was created using PHP 5.2.5 (Hypertext Preprocessor).

To create RadReportMiner, we built PHP scripts that applied a modified version of NegEx,[13] an algorithm developed to detect negation in medical (nonradiology) reports. NegEx is a rule-based system for detecting negation, adopting regular expressions to detect signals in the texts that commonly indicate negation. We expanded NegEx's functionality by: (1) adding to its rule base a set of negation phrases specific to the radiology domain, such as "no evidence of," "no longer visualized," and "has healed" (Appendix A). We also implemented a module to detect uncertainty associated with findings by adding a separate category of "uncertainty" terms to NegEx, such as "cannot exclude" and "not ruled out" (Appendix B).

RadReportMiner parses free-text radiology reports into their individual sections, recognizing the main headings such as history, findings, and impression. Demarcating the report sections is important since RadReportMiner gives the user the opportunity to exclude partiuclar report sections that often contain search terms that are confounded by negation or uncertainty, particularly the history section. For this study, every report was processed in its entirety without excluding any section. RadReportMiner identifies the start of the history section by looking for any of the phrases "history," "indication," or "clinical finding" near the beginning of the report. The history section ends when "findings" or a synonym of "impression" is encountered. Synonyms of impression include "impression," "conclusion," "opinion," and "interpretation." The findings and impression sections are segmented in a similar manner.

The RadReportMiner algorithm uses regular expressions to search sentences for negations preceding and following the keyword, up to six words apart (inclusive), as with NegEx.[13] RadReportMiner includes regular expressions to detect statements of uncertainty in a similar fashion as to detecting negation. When a match is found, RadReportMiner computes a relevance score using the rules listed in Table 1. The relevance score is used to classify each report into one of two categories: positive or negative, indicating whether the report contains the search term in a positive context (the term is stated to be present) or a negative context (term stated to be absent or uncertainty is expressed about the term). When RadReportMiner returns multiple search results, it sorts them by decreasing relevance to the radiologist.

The scoring system classifies a report as negated when the search term is associated with a negation phrase in RadReportMiner's regular expression database. When a search term is found to be associated with an uncertainty phrase, the report is classified as uncertain. Both negated and uncertain reports are considered negative or not containing the finding of interest. Reports containing the search term without any associated negation or

Table 1. Relevance Scoring and Classification in RadReportMiner

Condition	Score
Pre-conditional phrase	+1
Post-conditional phrase	+1
Pre-negation phrase	+10
Post-negation phrase	+10
History section not identifiable	+1000
Keyword in "history" but not in "findings/impression"	+10000
Pre-uncertainty phrase	+100000
Post-uncertainty phrase	+100000
Negation absent	-10
Score Range	**Classification**
score <= 0	Positive
0 < score < 100000	Negative (Negated)
score >= 100000	Negative (Uncertain)

This table lists the heuristic clues in text used to detect positive, negative, and uncertainty phrases in radiology reports and the score assigned to each when the corresponding clue is found to be associated with a finding. The scores are summed for each finding and the total score is used to classify the report as positive or negative according the score range shown in the table

uncertainty phrases are classified as positive. If the history section of a report cannot be separated reliably or if the search term is found only in the history section, the report is classified as negative.

We created a set of gold standard report classifications by asking a radiologist to perform five searches of single terms in our radiology reports database. The reports were retrieved using an exact-match keyword-based search engine, which has 100% precision and recall in identifying reports containing the keyword/key phrase of interest. The search terms were "appendicitis," "optic neuritis," "pneumonia," "hydronephrosis," and "fracture." After each search term was entered, the radiologist manually categorized the first 100 results (or fewer, if fewer were found) as "positive" or "negative" for the finding or diagnosis in question, thus establishing the ground truth. If more than 100 reports contained the keyword of interest, only the first 100 reports were used. A positive report was defined as one that gave the radiologist reasonable certainty that the desired finding or diagnosis would be present on the image(s) being described. All other reports were classified as negative.

To evaluate RadReportMiner, the same set of reports for which the radiologist provided the gold standard classification was then classified automatically by RadReportMiner (as positive or negative).

The reports were also written to individual files for processing by Google Desktop, a common word-based indexing engine, to compare its search results with those obtained using RadReportMiner. Google Desktop version 5.7.0806 was configured to search the radiology reports. After Google Desktop completed indexing the reports, the same five searches used to evaluate the RadReportMiner system were performed with Google Desktop. The "by relevance" link at the top right corner of the Google Desktop results was selected to sort the results using Google Desktop's internal algorithm.

Precision and recall were calculated for RadReportMiner and Google Desktop, and the results were compared against the radiologist's ground truth determination for each report. Precision, the fraction of relevant documents retrieved by the search engine, was calculated by dividing the number of true-positive reports (reports marked positive by the search engine that were also marked positive by the radiologist) by the total number of reports retrieved (classified as positive) by the search engine. Recall, the fraction of all existing relevant documents retrieved, was calculated by dividing the number of true-positive reports by the total number of reports marked positive by the radiologist within the 100-reports-or-fewer collection.

Mean precision and recall were computed in two different ways: per-term and per-document. Per-term means were calculated by weighing each search term equally, regardless of the number of reports retrieved (denominator is the number of search terms). Per-document means were calculated by weighing each retrieved report equally (denominator is the number of reports retrieved). Fisher's exact test was used to determine statistical significance of the differences between the two search methods for each search term. Wilcoxon signed ranks test was used to detect differences in overall precision and recall between the two search methods.

RESULTS

A total of 464 reports were returned for the five search terms (four terms each yielded the maximum 100 reports and one yielded 64 reports). The radiologist classified 119 (26%) of these as positive and 345 (74%) as negative. These 464 reports were authored by 40 different physicians.

Of the 464 reports, RadReportMiner classified 106 as positive. Of these, 86 (81%) were true positives (marked as positive by the radiologist); thus, RadReportMiner achieved an overall per-document precision of 81% (86 of 106 reports) and a recall of 72% (86 of 119 reports). Google Desktop classified 385 documents as positive, of which 104 (27%) were true positives. Thus, Goolgle Desktop achieved a precision of 27% (104 of 385 reports) and a recall of 87% (104 of 119 reports). Per-term and per-document calculations produced the same numbers except for RadReportMiner's recall statistics, which were 76% per-term and 72% per-document.The overall precision for RadReportMinerwas significantly higher than that of Google Desktop whether weighing each report equally (per-document; $p<0.0001$) or weighing each search term equally (per-term; $p=0.042$; median difference of 48% with 95% confidence interval of 34–75%). Google Desktop generally achieved higher recall than RadReportMiner; the difference was marginally significant per-document ($p=0.0057$) but not statistically significant on a mean basis per-term ($p=0.273$); for two individual terms, the differences were statistically significant. Table 2 lists the results by search term.

As shown in Table 2, a different number of reports were retrieved by RadReportMiner and Google Desktop for the search terms "fracture" and "pneumonia," for which Google Desktop returned 83 and 38 reports, compared with 100 and 100 reports for RadReportMiner, respectively. For the three other search terms, both search engines returned the same number of reports.

The majority of false positives from RadReportMiner were due to unrecognized uncertainties and word distance (the number of words from the negation phrase to the search term, including both) greater than six words. For example, with "no" as the negation term and "appendicitis" as the search term, the following phrase which has a word distance of seven would not be recognized by RadReportMiner as a negated search term: "**no** *periappendiceal fat stranding to suggest* **appendicitis**." Table 3 lists the causes and examples of false positives. The majority of false negatives from RadReportMiner were due to absence of keyword in the findings/impression sections and negation of some but not all instances of the keyword. Table 4 lists the causes and examples of false negatives.

DISCUSSION

The ideal radiology search engine is one that has high "recall" (retrieving all the relevant reports pertinent to a user's interest from the entire database of reports) and precision (retrieving mostly relevant reports among all reports retrieved). For searching radiology reports, precision is often more important than recall—when radiologists use a search engine to find reports containing a specific finding or diagnosis, they want high-precision searches to minimize the number of false-positive reports; irrelevant reports require the radiologist to read and manually exclude them, wasting precious time and lowering efficiency. High recall is important in situations where the radiologist is interested in searching for rare findings and diagnoses; one would not want to

Table 2. Search Performance Comparison between RadReportMiner and Google Desktop

	RadReportMiner		Google Desktop		p value*
# Results Returned	#		#		
Appendicitis	100		100		
Fracture	100		83		
Hydronephrosis	100		100		
Optic neuritis	64		64		
Pneumonia	100		38		
Total	*464*		*385*		
Precision	#	%	#	%	
Appendicitis	13/20	65%	14/100	14%	**< .0001**
Fracture	16/18	89%	29/83	35%	**< .0001**
Hydronephrosis	21/22	96%	31/100	31%	**< .0001**
Optic neuritis	21/27	78%	21/64	33%	**0.0002**
Pneumonia	15/19	79%	9/38	24%	**0.0001**
*mean (per-term)^*		*81%*		*27%*	**0.042**$^{+}$
*mean (per-document)^*	*86/106*	*81%*	*104/385*	*27%*	**< .0001**
Recall	#	%	#	%	
Appendicitis	13/14	93%	14/14	100%	1
Fracture	16/32	50%	29/32	91%	**0.0008**
Hydronephrosis	21/31	68%	31/31	100%	**0.0008**
Optic neuritis	21/21	100%	21/21	100%	1
Pneumonia	15/21	71%	9/21	43%	0.118
*mean (per-term)^*		*76%*		*87%*	0.273^{+}
*mean (per-document)^*	*86/119*	*72%*	*104/119*	*87%*	**0.0057**

^ per-term means are calculated weighing each search term equally; per-document means are calculated weighing each retrieved report equally

* calculated using Fisher's Exact Test

$^{+}$ calculated using Wilcoxon Signed Ranks Test

The top third of the table lists the terms searched and the number of results retrieved by each search engine. The middle third shows the precision values associated with each search term, calculated by dividing the number of true-positive reports by the total number of reports retrieved, with the percentage to the side. The bottom third shows recall values, calculated by dividing the number of true-positive reports by the total number of positive reports. The right-most column lists the *p* values

miss any reports of such rare cases, and the number of false positives would be small since the search term is rare. Our interest is to improve search in circumstances of search for common diseases and diagnoses—to reduce the number of false-postive search results.

There is generally a trade-off between precision and recall. The current word-based search engines perform superbly in terms of recall, but often lack precision. While high recall is certainly beneficial when the radiologist is searching for uncommon entities, it can be highly problematic when search-

Table 3. RadReportMiner False Positives

Reason	#	%
Unrecognized uncertainties	11	55%
Word distance > 6 words	5	25%
Typographical / transcription error	2	10%
Uncommon phrasing	1	5%
Failure to recognize "history of" as a negation equivalent	1	5%
Examples		
Unrecognized uncertainties	Mesenteric inflammation with abscess may be secondary to appendicitis or diverticulitis	
Word distance > 6 words	There is no fluid or inflammatory change about the cecal tip to suggest appendicitis (word distance = 12 words)	
Typographical error	Hallux **rigidusNo** acute fracture identified. (no space)	
Uncommon phrasing	Fracture ***is*** healed (instead of ***has)***	
Failure to recognize "history of" as negation equivalent	Given patient's history of optic neuritis, these findings most likely represent demyelination.	

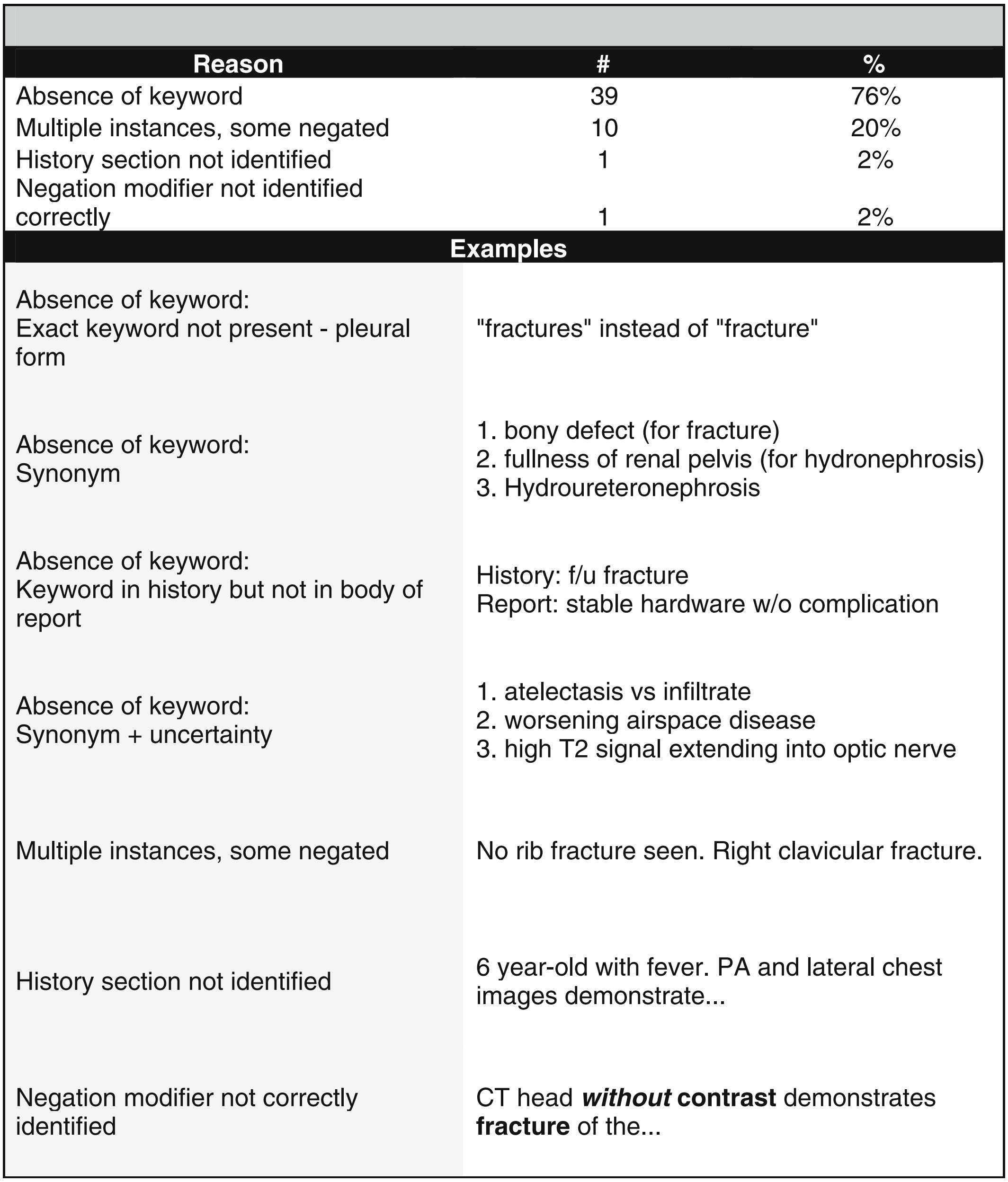
The majority of RadReportMiner's false positives (e.g., nonfracture classified as fracture) were due to unrecognized uncertainties and word distance greater than six words. Word distance is defined as the number of words from the negation phrase to the search term, including both. For example, with "no" as the negation term and "appendicitis" as the search term, the following phrase has a word distance of seven: "**no** *periappendiceal fat stranding to suggest* **appendicitis**"

ing for common diseases and findings. Consider, for example, a search engine with a precision of 27%, as we found with Google Desktop. A search using this engine that returns 200 reports contains only 54 that are true positives—the radiologist would need to wade through and discard 146 false-positive reports to discover the positives. On the other hand, the RadReportMiner system, having a precision of 81%, would return only 38 false-positive hits. Given the high precision of RadReportMiner, it is not surprising that it had a lower recall than Google Desktop, though the difference was only marginally statistically significant.

Our goal in this study was to improve precision by recognizing negations and uncertainties. Our results demonstrate improved precision of our system over the word-based Google Desktop search engine. We consider our results beneficial to radiologists wanting efficient methods to search reports with higher precision than current keyword-based methods provide. In fact, based on our results, for each search term, one would need to go through an average of 36.8 more results returned by Google Desktop to find the same number of positive reports returned by RadReportMiner, amounting to reading 137% additional reports.

There were several limitations of this study. First, there were a small number of terms selected for the searches used in our evaluation. In addition, a limited number of reports was analyzed. We are currently undertaking a study on a larger sample of

Table 4. RadReportMiner False Negatives



Reason	#	%
Absence of keyword	39	76%
Multiple instances, some negated	10	20%
History section not identified	1	2%
Negation modifier not identified correctly	1	2%

Examples	
Absence of keyword: Exact keyword not present - pleural form	"fractures" instead of "fracture"
Absence of keyword: Synonym	1. bony defect (for fracture) 2. fullness of renal pelvis (for hydronephrosis) 3. Hydroureteronephrosis
Absence of keyword: Keyword in history but not in body of report	History: f/u fracture Report: stable hardware w/o complication
Absence of keyword: Synonym + uncertainty	1. atelectasis vs infiltrate 2. worsening airspace disease 3. high T2 signal extending into optic nerve
Multiple instances, some negated	No rib fracture seen. Right clavicular fracture.
History section not identified	6 year-old with fever. PA and lateral chest images demonstrate...
Negation modifier not correctly identified	CT head ***without*** **contrast** demonstrates **fracture** of the...

The majority of RadReportMiner's false negatives (e.g., true fracture classified as negative) were due to an absence of the keyword in the findings/impression sections. Another 20% were due to negation of certain instances of the keyword but not others. The algorithm classifies a report as negative once a single instance of negation of that keyword is found

reports with more search terms. Another limitation of our work is that it does not account for multiple keywords and synonyms. For example, we are unable to search for "pneumonia" or "infiltrate" near "infectious." We are also unable to search for both the singular and plural forms of "fracture." Having these capabilities can potentially improve recall. Adding a module to recognize multiple forms of a word and to handle multiple keywords would address several of these limitations. We will be refining our algorithms to account for these issues.

There are several benefits of our methods. All the software used in this project, with the exception of the operating system (Windows XP), is in the public domain and, thus, is affordable to any institution wishing to implement such a search engine. We based our algorithm on a previously researched and published method of negation detection, which shortens the development time and reduces coding complexity. Finally, our system runs quickly on a large corpus of radiology reports. We ultimately aim to deploy our methods in radiology report search systems to help radiologists to find information in vast report archives more effectively.

CONCLUSION

We developed a method to identify negation and uncertainty of findings in radiology reports. Adding negation and uncertainty identification to a word-based radiology report search engine improves the precision of search results over a search engine that does not take this information into account. Our approach may be useful to adopt into current report retrieval systems to help radiologists to more accurately search for radiology reports.

APPENDIX A

Additional negation phrases incorporated into the RadReportMiner algorithm:

1. Clear of
2. Healed
3. No evidence of
4. No evidence for
5. Resolution of
6. No longer seen
7. No longer present
8. No longer appreciated
9. No longer visualized
10. Has resolved
11. Have resolved
12. Has healed
13. Have healed

Phrases 1–5 were added to the prenegation list. Phrases 6–13 were appended to the postnegation list.

APPENDIX B

Uncertainty terms utilized in the RadReportMiner algorithm:

1. Cannot rule out
2. Can not rule out
3. Cannot completely rule out
4. Can not completely rule out
5. Cannot absolutely rule out
6. Can not absolutely rule out
7. Cannot exclude
8. Can not exclude
9. Could represent
10. May represent
11. Equivocal
12. Cannot be ruled out
13. Can not be ruled out
14. Cannot be excluded
15. Can not be excluded
16. Not ruled out
17. Not excluded
18. Cannot be completely excluded
19. Not completely ruled out
20. Should also be considered

Phrases 1–11 are preuncertainties that occur before the search term. Phrases 12–20 are postuncertainties that follow the search term.

REFERENCES

1. Desjardins B, Hamilton RC: A practical approach for inexpensive searches of radiology report databases. Acad Radiol 14(6):749–756, 2007

2. Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, Halpern EF, Thrall JH: Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 234 (2):323–329, 2005

3. Erinjeri JP, Picus D, Prior FW, Rubin DA, Koppel P: Development of a Google-based search engine for data mining radiology reports. J Digit Imaging 22(4):348–356, 2008

4. Rubin DL, Desser TS: A data warehouse for integrating radiologic and pathologic data. J Am Coll Radiol 5(3):210–217, 2008

5. Wong ST, Hoo Jr, KS, Cao X, Tjandra D, Fu JC, Dillon WP: A neuroinformatics database system for disease-oriented neuroimaging research. Acad Radiol 11(3):345–358, 2004

6. Ramaswamy MR, Patterson DS, Yin L, Goodacre BW: MoSearch: a radiologist-friendly tool for finding-based diagnostic report and image retrieval. Radiographics 16(4):923–933, 1996

7. Uzuner O, Zhang X, Sibanda T: Machine learning and rule-based approaches to assertion classification. J Am Med Inform Assoc 16(1):109–115, 2009

8. South BR, Phansalkar S, Swaminathan AD, Delisle S, Perl T, Samore MH: Adaptation of the NegEx algorithm to Veterans Affairs electronic text notes for detection of influenza-like illness (ILI). AMIA Annu Symp Proc 1118, 2007

9. Mitchell KJ, Becich MJ, Berman JJ, Chapman WW, Gilbertson J, Gupta D, Harrison J, Legowski E, Crowley RS: Implementation and evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports. Stud Health Technol Inform 107(Pt 1):663–667, 2004

10. Meystre SM, Haug PJ: Comparing natural language processing tools to extract medical problems from narrative text. AMIA Annu Symp Proc 525–529, 2005

11. Yang H, Lowe HJ: A grammar-based classification of negations in clinical radiology reports. AMIA Annu Symp Proc 988, 2005

12. Yang H, Lowe HJ: A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc 14(3):304–311, 2007

13. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 34(5):301–310, 2001