J Digit Imaging (2012) 25:30–36
DOI 10.1007/s10278-011-9426-6

Automated Detection of Critical Results in Radiology Reports

Paras Lakhani · Woojin Kim · Curtis P. Langlotz

Published online: 25 October 2011

Abstract The goal of this study was to develop and validate text-mining algorithms to automatically identify radiology reports containing critical results including tension or increasing/new large pneumothorax, acute pulmonary embolism, acute cholecystitis, acute appendicitis, ectopic pregnancy, scrotal torsion, unexplained free intraperitoneal air, new or increasing intracranial hemorrhage, and malpositioned tubes and lines. The algorithms were developed using rule-based approaches and designed to search for common words and phrases in radiology reports that indicate critical results. Certain text-mining features were utilized such as wildcards, stemming, negation detection, proximity matching, and expanded searches with applicable synonyms. To further improve accuracy, the algorithms utilized modality and exam-specific queries, searched under the "Impression" field of the radiology report, and excluded reports with a low level of diagnostic certainty. Algorithm accuracy was determined using precision, recall, and *F*-measure using human review as the reference standard. The overall accuracy (*F*-measure) of the algorithms ranged from 81% to 100%, with a mean precision and recall of 96% and 91%, respectively. These algorithms can be applied to radiology report databases for quality assurance and accreditation, integrated with existing dashboards for display and monitoring, and ported to other institutions for their own use.

Keywords Algorithms · Communication · Critical Results Reporting · Data Mining · Natural Language Processing · Quality Assurance · Quality Control

P. Lakhani (✉) · W. Kim · C. P. Langlotz
Department of Radiology,
Hospital of the University of Pennsylvania,
3400 Spruce Street,
Philadelphia, PA 19106, USA
e-mail: Paras.lakhani@jefferson.edu

Introduction

In 2002, the Joint Commission established its National Patient Safety Goals program, which required the timely reporting of critical results including those rendered by diagnostic imaging services [1]. Likewise, in their Standard Practice of Communications, the American College of Radiology (ACR) emphasized the timely reporting of critical results and the documentation of critical results communications in the radiology report [2].

Like many other institutions, our radiology practice follows the ACR guidelines and requires the documentation of such communications in the body of the radiology report. We do not use an automated notification system for relaying critical findings, and therefore manual review of a large number of radiology reports is required to demonstrate compliance and for Joint Commission accreditation, which is time-consuming, inexact, and prone to sampling error.

Thus, an automated system that could detect and track the communication of critical findings would be of tremendous value in this regard. In a previous published experiment, we developed an automated system with an overall accuracy of 98% for detecting radiology reports that indicate communication of results to a healthcare provider [3]. However, no automated methods exist for detecting radiology reports that contain critical results themselves.

Consequently, the purpose of this study is to develop text-mining algorithms that could detect the presence of certain critical results in radiology reports. Our hypothesis is that rule-based algorithms using standard languages can achieve high accuracy for identifying radiology reports with critical findings. In this study, we chose results that most practices would consider

critical including tension or increasing/new large pneumothorax, acute pulmonary embolism, acute cholecystitis, acute appendicitis, ectopic pregnancy, scrotal torsion, unexplained free intraperitoneal air, new or increasing intracranial hemorrhage, and malpositioned tubes and lines.

Materials and methods

This study was approved by the Institutional Review Board and was compliant with the Health Insurance Portability and Accountability Act, utilizing a preexisting de-identified database of radiology reports.

Database

Initial testing was performed on a stand-alone radiology reports database consisting of approximately 2.3 million diagnostic radiology procedures performed at The Hospital of the University of Pennsylvania from 1997 to 2005. Subsequent testing was performed on a database of approximately 10 million radiology reports performed at our institution and regional affiliates from 1988 to 2011. The radiology reports were transferred from our radiology information system (RIS; IDXrad v9.6, IDX, Burlington VT) onto a secondary research relational database management system (RDBMS) using Oracle 10 g Enterprise Edition as the database server (Oracle, Redwood Shores, CA), with full-text data-mining options enabled and accessible via structured query language (SQL). After initial testing, the queries were tested for compatibility with another popular RDBMS, MySQL (Sun Microsystems, Santa Clara CA), with the Sphinx Search Engine (Sphinx Technologies, http://www.sphinxsearch.com) enabled, which allowed more rapid indexing and robust search capabilities. The databases were accessed locally on a computer for development and testing, which was equipped with the Intel i7 core processor (Intel, Santa Clara, CA) using 4 GB of RAM, and configured with a dual-boot system that could run either Windows 7 (Microsoft, Redwood, WA) or Ubuntu Linux 11.04 (Canonical/Ubuntu Foundation, London, UK. http://www.ubuntu.com).

Critical results

Nine commonly encountered critical results were chosen from an established list maintained by our radiology department, which included acute pulmonary embolism, acute cholecystitis, acute appendicitis, ectopic pregnancy, scrotal torsion, tension or new/increasing large pneumothorax, unexplained free intraperitoneal air, increasing or new intracranial hemorrhage, and malpositioned nasogastric, feeding, and endotracheal tubes.

Algorithm development

Query algorithms were developed using SQL, and designed to search for common words and phrases in radiology reports that indicate critical results.

Whenever applicable, synonyms were utilized to expand the search. For example, "ectopic pregnancy" and "extrauterine pregnancy" were considered equivalent. In addition, the algorithms were limited to search relevant modalities and study types, so that a brain magnetic resonance imaging (MRI) would have been excluded in the algorithm for acute cholecystitis.

Proximity searching was also utilized to further improve algorithm accuracy. For example, the algorithm for acute pulmonary embolism detects reports where the word "embolism" is within a certain word distance from the word "pulmonary." To further improve accuracy, negation detection was employed to exclude reports where critical findings were preceded by negative modifiers such as "no" or "without."

The algorithm also contained wildcards to expand the list of searchable words with a common stem such as "embol%", which would search for "embolism", "embolic", "emboli," and "embolus." Finally, phrases that conveyed a low-level of diagnostic certainty were excluded from the algorithms such as "ectopic pregnancy is unlikely" (Fig. 1).

Impression parser

Since many critical results are contained within the "Impression" section of the radiology report, a PHP-based parsing script based on regular expressions was developed that parsed the "Impression" field from every radiology report. The contents of the "Impression" field were then duplicated onto a separate searchable indexed column in the database. Some algorithms exclusively searched the "Impression" field to improve precision. Synonyms, certain misspellings, and pleural forms of "Impression" were also parsed such as "Summary," "Diagnosis," "Opinion," "Conclusion," and "Pression" (Fig. 2). However, the algorithm for tension or increasing large pneumothorax searched the entire radiology report and not exclusively the "Impression" section.

Algorithm validation

The algorithms underwent a process of iterative refinement until their accuracies did not improve significantly between iterations and validated using human review as the reference standard. To explain further, 50–100 reports were

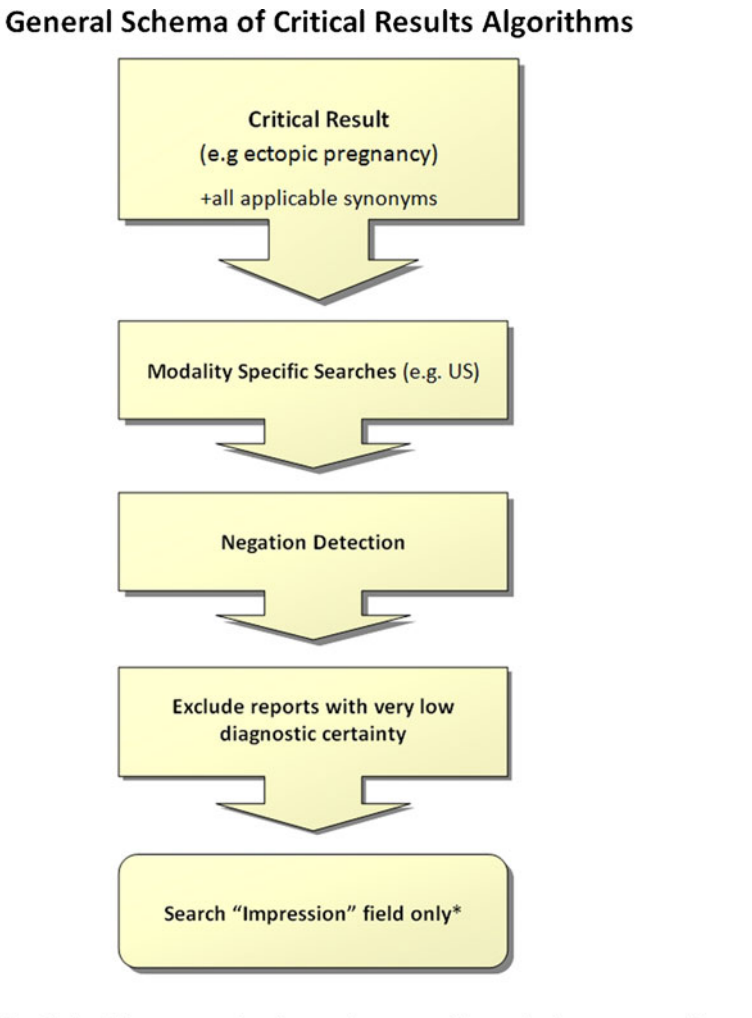

Fig. 1 This is a general schema of the critical results algorithms consisting of the concepts outlined above

initially selected by the algorithms, and these reports were scored as containing or not containing the appropriate critical results. Subsequently, the algorithms were modified accordingly, and a new batch of 50–100 reports were then selected. Only new reports were selected to prevent selection bias. These steps were repeated until precision and recall did not improve significantly between iterations. For the final algorithms, algorithm accuracy was determined using precision, recall, and *F*-measure according to previously published methods [3–6], by analyzing 100 and 500–2000 new random reports for each algorithm, respectively. A minimum number of reports was sampled to achieve an estimated 95% confidence interval of approximately ±10%.

Precision represented the percentage of radiology reports selected by the algorithm that actually contained the critical result in question. Recall represented the percentage of radiology reports selected by the algorithm of all possible reports in the database positive for that critical value. This number was estimated by selecting only reports from relevant modalities and exam types for the critical value in question. As an example, for new or worsening intracerebral hemorrhage, only head CT or brain MRI examinations were selected. Of these studies, a subset of reports excluded by the query algorithm was selected, and the frequency of critical findings in this subset was determined. From this number, the estimated recall rate and 95% confidence intervals were calculated [3, 4]. The *F*-measure was then determined for each critical value, which represented a weighted harmonic mean of precision and recall, using equal weighting between the two values [6].

Results

The algorithms were functional with Oracle and MySQL databases and compatible with Windows and Linux operating systems.

For accuracy of the query algorithms, a summary of the results is provided in Table 1. For acute appendicitis, the

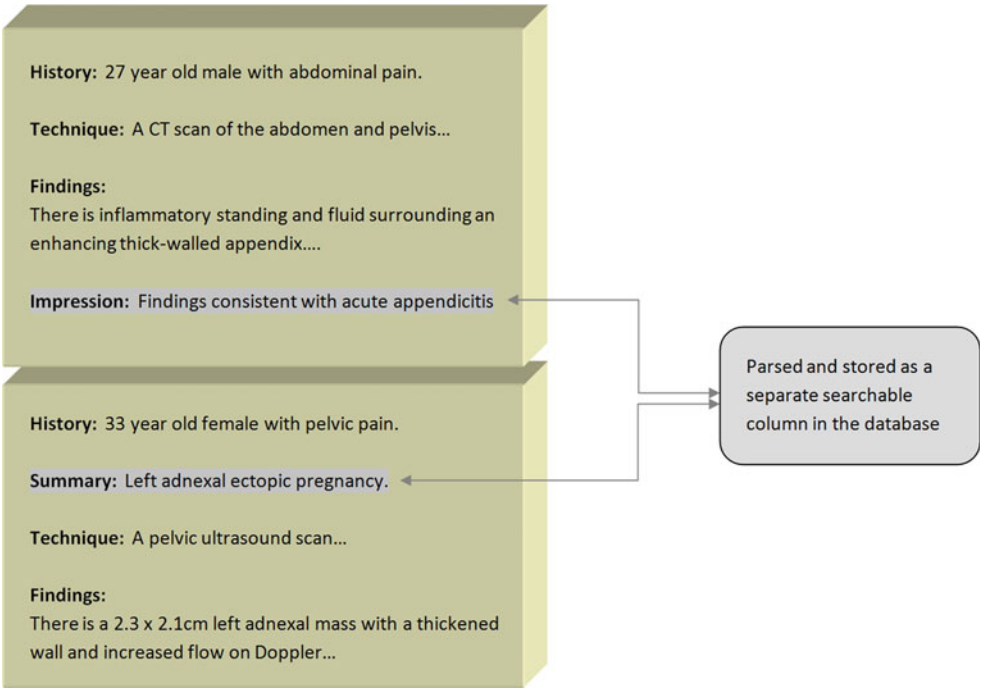

Fig. 2 Impression parser: The "Impression" field of every report in the database was parsed using a PHP script based on regular expressions. The script was able to parse the "Impression" section no matter its location in the report. The parser could also handle synonyms and misspellings of "Impression"

precision, recall, and *F*-measure of the algorithm were 99.0% (CI: 97.1–100.0%), 89.0% (CI: 83.0–95.8%), and 94%, respectively; for acute cholecystitis, 96.0% (CI: 92.2–100.0%), 88.6% (CI: 82.5–95.7%), and 92.3%, respectively; for ectopic pregnancy, 98.0% (CI: 94.1–100.0%), 94.8% (CI: 86.1–100.0%), and 96.4%, respectively; for unexpected free intraperitoneal air, 87.0% (CI: 80.4–93.6%), 93.7% (CI: 80.3–100.0%), and 90.4%, respectively; for new or increasing intracranial hemorrhage, 94.0% (CI: 87.3–97.5%), 68.0% (CI: 59.7–79.0%), and 81.0%, respectively; for large or tension or new/increasing large pneumothorax, 96.0% (CI: 92.2–99.8%), 84.2% (CI: 75.3–95.4%), and 90.1%, respectively; for acute pulmonary embolism, 99.0%(CI: 97.1–100.0%), 97.8% (CI: 81.4–100.0%), and 98.9%, respectively; for acute scrotal torsion, 96.0% (CI: 88.3–100.0%), 100.0% (CI: 96.8–100.0%), and 98.0%; for malpositioned nasogastric, feeding, and endotracheal tubes, 98.0% (CI: 95.3–100.0%), 100% (CI: 92.2–100.0%), and 99.0%, respectively.

Discussion

In this study, we developed and validated multiple rule-based algorithms for identifying radiology reports that contain certain critical results. This research was driven by a need to create automated, precise query methodologies that could probe radiology report databases to determine frequency of such findings for quality control purposes and to help satisfy Joint Commission requirements.

While this work fits under the realm of natural language processing (NLP), which has been used to classify radiology reports [7–11], much of the work performed here involved direct mining of unstructured text. Rather than employ statistical or machine-learning methods to classify data, which is used in many modern NLP systems [12–14], the methods used in this study are more similar to traditional rule-based approaches to text classification [15, 16]. Some authors in the information retrieval field believe that rule-based or knowledge-engineering approaches are the most accurate and reliable methods for text classification [17].

The precision, recall, and overall accuracy (*F*-measure) of these algorithms compare favorably to that published elsewhere in the literature, with average values of 95.9%, 90.7%, and 93.3%, respectively [5, 8–10]. We utilized many text-mining features to improve accuracy including wildcards, proximity matching, search expansion with synonyms, negation detection, modality and exam-specific queries, and searching only the "Impression" field, which was separately parsed and indexed (Fig. 2).

In addition, the algorithms required robust negation detection ability, which was uniquely tailored for each critical result in question (the algorithms had specific negation dictionaries and word proximity rules). Thus, the algorithms were able to eliminate reports containing phrases such as "no evidence of acute cholecystitis," for example, which occur far more frequently than reports that actually indicate "acute cholecystitis."

Except for one, all of the algorithms had overall accuracies (*F*-measure) greater than 90%, which is considered highly accurate for classifying text [12]. The algorithms for acute pulmonary embolism and malpositioned tubes were the most accurate with overall accuracies greater

Table 1 Accuracy of critical results algorithms

Critical result	Precision	Recall
Malpositioned tubes*	98.0% (CI: 95.3–100.0%)	100% (CI: 92.2–100.0%)
Acute pulmonary embolism	99.0% (CI: 97.1–100%)	97.8% (CI: 81.4–100.0%)
Tension or new/increasing large pneumothorax	96.0% (CI: 92.2–100%)	84.2% (CI: 75.3–95.4%)
Acute cholecystitis	96.0% (CI: 92.2–100.0%)	88.6% (CI: 82.5–95.7%)
Ectopic pregnancy	98.0% (CI: 94.1–100.0%)	94.8% (CI: 86.1–100.0%)
Acute appendicitis	99.0% (CI: 97.1–100%)	89.0% (CI: 83.0–95.8%)
Unexplained free intraperitoneal air	87.0% (CI: 80.4–93.6%)	93.7% (CI: 80.3–100.0%)
New or increasing intracranial hemorrhage	94.0% (CI: 87.3–97.5%)	68.0% (CI: 59.7–79.0%)
Acute scrotal torsion	96.0% (CI: 88.3–100.0%)	100.0% (CI: 96.8–100.0%)

*(ET, NG, feeding tubes only)

than 98%. The algorithm for new or increasing intracranial hemorrhage was the least accurate with an *F*-measure of 81% due to the recall rate of 68% (Table 1). In this case, the algorithm excluded a significant number of reports that were actually positive for new or increasing intracranial hemorrhage. One of the main reasons was because the negation detection component of the new/increasing intracranial hemorrhage algorithm was designed to exclude reports containing negative modifiers within a certain distance of "hemorrhage" or its synonym. However, there were many instances in which a negative modifier, such as "no," was near the critical value in question, but actually negated another entity. An example of such a report is provided in Fig. 3. One method to remedy this problem would include sentence-specific queries, which is possible using newer indexing technologies.

While the algorithms shared certain similarities, the text-mining algorithms were individually customized for each critical result in question. The complexity and length of the algorithms had varied, with some being relatively simple, and others being rather complex. One of the simpler algorithms was that for acute appendicitis, and one of the more complex algorithms was that for unexpected free intraperitoneal air, which consisted of over 20 general and well over 100 specific rules (Figs. 4 and 5).

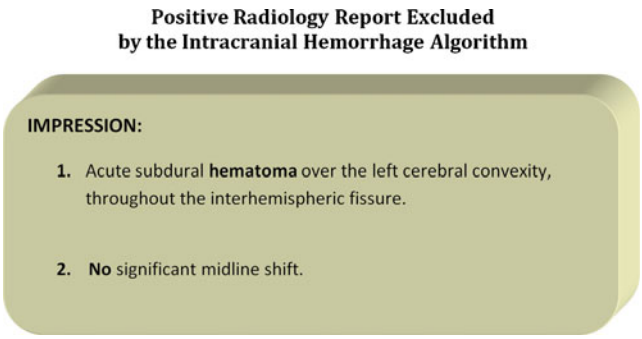

Fig. 3 Explanation: The negative modifier "no" is within close proximity to "hematoma." The algorithm is currently designed to exclude these reports. However, in this example, "no" is used to negate "midline shift" and not "hematoma." Future efforts are underway to remedy this problem using sentence-specific searches

In designing these algorithms, there was occasionally a tradeoff between precision and recall. That is, the more radiology reports recalled by the algorithms, the less precise they were by selecting some reports that did not contain critical findings. In such situations, the algorithms were preferentially tailored to have greater precision. That way, the reports selected by the algorithm were more likely to contain critical findings and therefore to be relevant for quality improvement efforts in this area. Nonetheless, the recall rates were relatively high, with an average recall rate of 90.7%.

All of the algorithms were designed to search the "Impression" section or equivalent, except that for acute scrotal torsion and tension or new/increasing large pneumothorax, which searched the entire report text. By doing this, the precision of the algorithms improved significantly, with little drop in recall, since the majority of critical findings were noted in the Impression section. If the radiology report had no Impression section or equivalent, the algorithms then searched the entire report.

Future efforts are underway to further improve the accuracy of these algorithms by refining the methods described above. In addition, we plan to develop algorithms to detect other critical results. Ultimately, we plan to use these algorithms to probe a larger up-to-date radiology reports database encompassing all reports at our institution and affiliates from 1997 to 2010 for performance monitoring and accreditation. These and other similar algorithms could also prove useful for retrospective research purposes and bio-surveillance initiatives, where critical results could be tracked and monitored over time. Because the algorithms were built using PHP and SQL, they can also be incorporated into dashboards for real-time monitoring and can also be used with automated notification systems that could run periodically throughout the day or at report sign-off. We also plan to use these algorithms in conjunction with a prior algorithm that

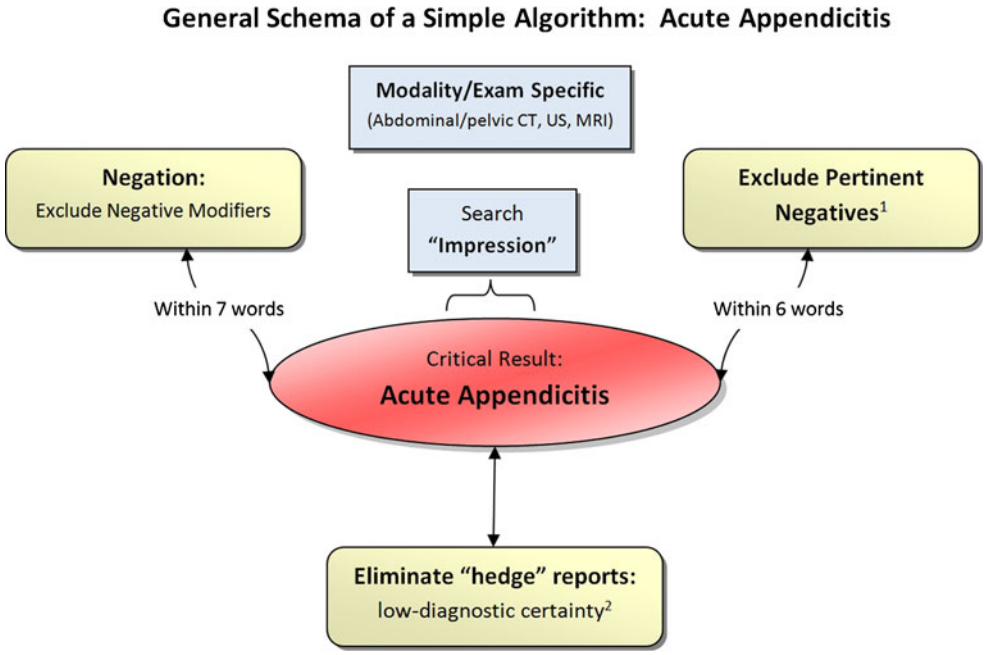

Fig. 4 This is a general schema of the algorithm for acute appendicitis. The algorithm consisted of modality and exam specific searches under the "Impression" field of the radiology report. A negation detection algorithm was employed. Reports with low diagnostic certainty and with certain pertinent negatives were also excluded. (1) For example, the algorithm excluded radiology reports containing phrases such as "No peri-appendicular inflammatory stranding." (2) The algorithm excluded reports such as "The appendix is not seen and therefore appendicitis cannot be excluded"

identifies reports containing documentation of communications [3, 18]). This, in turn, can be used to determine the frequency of reports with critical results that also have documentation of communications.

There are some limitations to this work. The algorithms were developed using a database of radiology reports from one tertiary care hospital and four regional affiliates. Thus, the accuracies may be reduced when applied to radiology

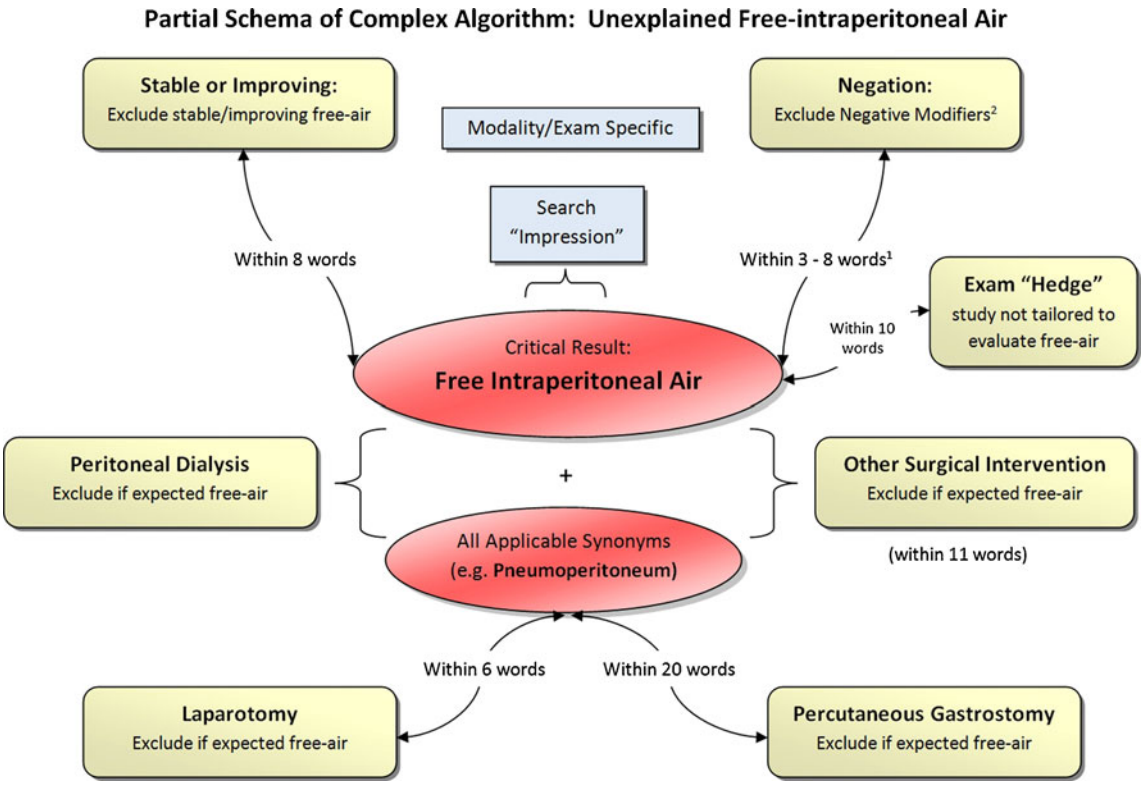

Fig. 5 This is a partial generalized schema of the algorithm for Unexplained Free Intraperitoneal Air. The entire algorithm had over 20 general and 100 specific rules. (1) Negation distance varied depending on the context and the negation term in question. (2) Certain statements containing negation terms were permitted such as "the pneumoperitoneum that was *not* clearly seen on yesterday's study has increased in size"

report databases from other institutions. However, given that the database spanned a long time frame and consisted of reports and created by over 100 attending, fellow, and resident radiologists, many of whom also worked and trained at other institutions, we feel that the algorithms are likely generalizable to all radiology practices. Also, many thousands of reports were manually reviewed and scored in developing these algorithms, and therefore, errors due to fatigue are possible. To minimize this effect, small batches of reports were reviewed over a long period of time.

Conclusions

Pattern- and rule-based approaches can achieve high accuracy for classifying critical results in unstructured radiology reports. The overall accuracies of the validated algorithms range from 81% to 100% including tension or increasing/new large pneumothorax, acute pulmonary embolism, acute cholecystitis, acute appendicitis, ectopic pregnancy, scrotal torsion, unexplained free intraperitoneal air, new or increasing intracranial hemorrhage and malpositioned tubes/lines. These algorithms can be applied to radiology report databases for quality assurance and accreditation, integrated with existing dashboards for display and real-time monitoring, and ported to other institutions for their own use.

Acknowledgements This research was funded by the Society for Imaging Informatics (SIIM) Research Grant.

References

1. Patient safety requirement 2 C (standard NPSG.2a). In: The 2007 Comprehensive Accreditation Manual for Hospitals: The Official Handbook. Oakbrook Terrace, IL: Joint Commission Resources, 2007, pp NPSG-3–NPSG-4
2. American College of Radiology: ACR Practice Guideline for Communication of Diagnostic Imaging Findings. In: Practice Guidelines & Technical Standards 2005. Reston, VA: American College of Radiology, 2005
3. Lakhani P, Langlotz CP: Automated detection of radiology reports that document non-routine communication of critical or significant radiology results. J Digital Imaging 23:647–57, 2010. Dec (Epub 2008 Oct)
4. Hersh WR, Detmer WM, Frisse ME: Information-retrieval systems. In: Medical Informatics. Springer, New York, NY, 2001, pp 539–572
5. 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. 34:301–310, 2001
6. Hripcsak G, Rothschild AS: Agreement, the *F*-measure, and reliability in information retrieval. J Am Med Inform Assoc. 12:296–298, 2005
7. Lacson R, Khorasani R: Natural language processing: the basics (part 1). J Am Coll Radiol. 8(6):436–7, 2011
8. Goldstein I, Arzrumtsyan A, Uzuner O: Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. AMIA Annu Symp Proc 11:279–283, 2007
9. Imai T, Aramaki E, Kajino M, Miyo K, Onogi Y, Ohe K: Finding malignant findings from radiological reports using medical attributes and syntactic information. Stud Health Technol Inform 129:540–544, 2007
10. Mamlin BW, Heinze DT, McDonald CJ: Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc 420–424, 2003
11. Hripcsak G, Austin JH, Alderson PO, Friedman C: Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 224: 157–163, 2002
12. Taira RK, Soderland SG: A statistical natural language processor for medical reports. Proc AMIA Symp 970–974, 1999
13. 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 234:323–329, 2005
14. Cheng LT, Zheng J, Savova GK, Erickson BJ: Discerning tumor status from unstructured MRI reports—completeness of information in existing reports and utility of automated natural language processing. J Digit Imaging 23(2):119–132, 2010. Apr (Epub 2009 May 30)
15. Hayes-Roth F: Rule-based systems. Commun ACM 28:921–932, 1985
16. Baud R, Lovis C, Rassinoux AM, Michel PA, Scherrer JR: Automatic extraction of linguistic knowledge from an international classification. Stud Health Technol Inform 52(Pt 1):581–585, 1998
17. Wilcox AB, Hripcsak G: The role of domain knowledge in automating medical text report classification. J Am Med Inform Assoc 10(4):330–338, 2003
18. Lakhani P, Langlotz CP: Documentation of nonroutine communications of critical or significant radiology results: a multiyear experience at a tertiary hospital. J Am Coll Radiol 7(10):782–790, 2010