Assessing the Feasibility of an Automated Suggestion System for Communicating Critical Findings from Chest Radiology Reports to Referring Physicians

Brian E. Chapman[1], **Danielle L. Mowery**[2], **Evan Narasimhan**[1], **Neel Patel**[1], **Wendy W. Chapman**[2], **Marta E. Heilbrun**[1]
[1] University of Utah, Radiology, Salt Lake City, UT
[2] University of Utah, Biomedical Informatics, Salt Lake City, UT
firstname.lastname@utah.edu

Abstract

Time-sensitive communication of critical imaging findings like **pneumothorax** or **pulmonary embolism** to referring physicians is important for patient safety. However, radiology findings are recorded in free-text format, relying on verbal communication that is not always successful. Natural language processing can provide automated suggestions to radiologists that new critical findings be added to a follow-up list. We present a pilot assessment of the feasibility of an automated critical finding suggestion system for radiology reporting by assessing suggestions made by the pyConTextNLP algorithm. Our evaluation focused on the false alarm rate to determine feasibility of deployment without increasing alert fatigue. pyConTextNLP identified 77 critical findings from 1,370 chest exams. Review of the suggested findings demonstrated a 7.8% false alarm rate. We discuss the errors, which would be challenging to address, and compare pyConTextNLP's false alarm rate to false alarm rates of similar systems from the literature.

1 Introduction

The communication of critical imaging findings from the radiologist to the referring physician is a key factor in providing efficacious patient care (Lakhani et al., 2012). Currently, the most common form of communication is a physician-to-physician telephone conversation, initiated by the radiologist at the time of image interpretation. This process is tedious, inefficient, and error prone. A missed communication can result in progressed disease, hospital readmission, and even death. In the United States, the American College of Radiology suggests three hallmarks of effective methods of communication: a) supporting the ordering provider in providing optimal patient care, b) using methods that are tailored to satisfy the need for timeliness, and c) implementing methods to minimize risk of communication errors (American College of Radiology, 2014). Critical findings may result in death or severe morbidity and require urgent or emergent attention (Larson et al., 2014). These critical test results are often documented in free-text imaging notes. Natural language processing (NLP) can automatically extract, track, and report these findings in a timely manner to support patient safety efforts.

2 Related Work

Machine learning and ruled-based NLP techniques have been used to detect critical information from radiology reports to support timely communication.

Yetisgen-Yildiz et al. created a machine-learning based text-processing pipeline that leverages a maximum entropy model to identify and classify sentences conveying clinically important follow-up recommendations concerning unexpected findings (Yetisgen-Yildiz et al., 2013). Pham et al. developed an NLP pipeline to detect and classify mentions of **thromboembolic disease** from angiography and venography reports. They used naive Bayes' feature selection then support vector machines and maximum entropy for classification (Pham et al., 2014). Esuli et al. developed two novel methods for extracting radiological findings from reports: a cascaded, two-stage ensemble of taggers generated by linear-chain conditional random fields (LC-CRFs) and a confidence-weighted ensemble method combining standard LC-CRFs and the two-stage method

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 181–185,
Berlin, Germany, August 12, 2016.

(Esuli et al., 2013).

Rule-based approaches have also been used to address critical finding detection. Lafourcade et al. created a linguistic-based algorithm to detect semantic relations between radiological findings (Lafourcade and Ramadier, 2016). Lakhani et al. developed an algorithm with finding-specific negation dictionaries to identify nine critical findings in impression sections and demonstrated a mean false alarm rate of 4% (Lakhani et al., 2012). Lacson et al. adapted and compared performance of two NLP systems—A Nearly New Information Extraction system (ANNIE) and Information for Searching Content with an Ontology-Utilizing Toolkit (iSCOUT)—to identify **pulmonary nodules**, **pneumothorax**, and **pulmonary embolus** with overall false alarm rates of 4% (ANNIE) and 10% (iSCOUT) (Lacson et al., 2012).

In this study, we applied a simple NLP system that leverages regular expressions by extending the lexicon for critical findings. We addressed a larger set of critical findings than in prior studies and evaluated not only the accuracy of the NLP system, but the appropriateness of generating a suggestion for critical finding communication.

Our long-term goal is to develop a communication system that identifies a variety of critical findings from radiology exams, facilitates appropriate communication of the findings to referring clinicians, and supports radiologist follow-up regarding the communicated findings. The short-term goal of this study is to build upon prior work by 1) adapting an NLP algorithm to automatically identify critical findings in radiology reports and suggest them to the radiologist for communication to referring physicians, 2) assessing the false alarm rate of the critical findings suggestion system, and 3) characterizing the errors generated by the system to determine feasibility of deploying the suggestion system in a radiology clinic. In this paper, we limit our analysis to imaging of the chest.

3 Methods and Materials

3.1 Data Set

In this IRB-approved study, we obtained all radiology reports from Oct-Dec 2013 generated by a large medical center in the United States. We excluded non-diagnostic exams (e.g., interventional procedures), as well as reports generated from services other than radiology, and reports with empty impressions sections, resulting in 54,459 exams. Mentions of critical findings in radiology reports are not common. Only 14,815 of the 54,479 reports (27%) contained critical finding expressions from our original knowledge base. Only a small portion of critical finding expressions would be expected to be observations of a new critical finding, because the majority are negated or chronic findings. For instance, from previous studies, we found that approximately 90% of pulmonary embolism mentions in radiology reports are negated.

From the 14,815 reports with critical finding expressions, we selected approximately half of the reports (7,176) for annotation. We built a Flask[1] web application for document-level annotation of the reports. Annotators used the tool to assign the following attributes to each finding mentioned in a report: *Existence* (definite negated existence, probable negated existence, ambivalent existence, probable existence, definite existence) and *Historicity* (new, chronic, historical) (Patel et al., 2016).

We annotated 39 critical findings occurring within abdomen/pelvis, chest, extremity, neuro, and spine exams. Eighteen of these critical findings were relevant to the chest: **aneursym, aortic dissection, cancer, ectasia, epiglottitis, fracture, free air, infarct, inflammation, mediastinal emphysema, pneumonia, pneumothorax, pulmonary embolism, retropharyngeal abscess, ruptured aneurysm, splenic infarct, tension pneumothorax,** and **thrombosis**.

With supervision by an attending radiologist (Author MH), two medical students (Authors NP, EN) independently annotated the impression sections of reports for any of 39 critical findings until acceptable agreement level between annotators was reached ($>$0.70). Each annotator then annotated reports independently, completing two-thirds of the 7,176 reports for a total 4,786 annotated reports.

From the full set of 7,176 reports, we sampled only reports from chest exams for this study, providing a development and test set of 1,538 chest exam reports. We randomly selected 168 annotated exams as a development set and further extended pyConTextNLP's knowledge base by reviewing pyConTextNLP's disagreements with the annotations. We then tested on the remaining blind set of 1,370 reports.

We split the impression section into sen-

[1] http://flask.pocoo.org/

tences using the Python text processing package TextBlob[2] then applied pyConTextNLP [3] to identify acute, positive critical findings.

3.2 Developing an automated critical findings suggestion system with pyConTextNLP

3.2.1 pyConTextNLP

We adapted an existing NLP algorithm, pyConTextNLP, to identify critical findings and their attributes from radiology imaging reports (Chapman et al., 2011). pyConTextNLP is an extension of the NegEx (Chapman et al., 2001) and ConText (Harkema et al., 2009) algorithms and relies on user-defined knowledge bases of targets (e.g., critical finding terms such as "pulmonary embolism"), modifiers (e.g., existence terms such as "may represent"), and lexical terms (e.g., "but") that terminate the scope of the modifiers.

3.2.2 Adapting and refining pyConTextNLP

The pyConTextNLP GitHub repository has a number of database files that have been created for previous projects[4]. We modified existing knowledge bases by comparing our automated classifications using pyConTextNLP against the annotator classifications. First, we reviewed false negative findings in the development set and added new terms to the knowledge base. The number of false negatives in the development set was small. To address potential alert fatigue from false alarms, we then focused our development on evaluating false positives in the development set. An acute, positive critical finding was defined as a mention of a critical finding with the following attributes: *Historicity*-new and *Existence*-probable or definite existence. If there was more than one mention of a given finding in a report, we assigned the report the same value as that of the most positive and most new mention. In reviewing the classifications, we examined the entire pyConTextNLP document for the report so that we could determine if the classification error occurred due to the knowledge base, classification rules, or algorithm implementation. Modifications to the code and knowledge bases were made iteratively to improve positive predictive performance compared to the annotations. Changes to pyConTextNLP primarily consisted of modifying synonyms and variants for critical findings and corresponding attributes.

[2] https://textblob.readthedocs.org/

[3] https://pypi.python.org/pypi/pyConTextNLP

[4] https://github.com/chapmanbe/pyConTextNLP/tree/master/KB

3.3 Evaluating pyConTextNLP

We ran pyConTextNLP over the test set and flagged documents with acute, positive critical findings for review. A radiologist (Author MH) was provided the flagged findings and their associated imaging report then asked the question, "Would you include this critical finding in a list of findings to communicate to the referring physician?" This question goes beyond analyzing accuracy of pyConTextNLP's annotations to the more stringent question of whether the finding should be communicated to another physician, which depends not only on accurate identification of the finding, but also on contextual information. We calculated precision (Eq. 1) and false alarm rate (Eq. 2) where FP (false positive) = rejected suggestion and TP (true positive) = accepted suggestion.

$$\text{precision} = \frac{TP}{(TP + FP)} \tag{1}$$

$$\text{false alarm rate} = 1 - \text{precision} \tag{2}$$

4 Results

Our primary goal was to assess the false alarm rate of the critical findings suggestion system and to characterize the errors generated by the system.

In total, we detected 77 findings requiring critical communication. These findings came from only five of our 18 categories. The most prevalent flagged findings were **pneumothorax** and **pneumonia** (Table 1).

Table 1: Distribution of flagged critical findings

critical finding	**count (%)**
pneumothorax	38 (49%)
pneumonia	29 (38%)
fracture	6 (8%)
cancer	3 (4%)
aneursym	1 (1%)
total	**77 (100%)**

Of the 77 observed critical findings, we observed 6 false positives, resulting in a false alarm rate of 7.8% and precision of 92.2%. Of the six false positives three were **cancer**, two were **pneumonia**, and one was an **aneurysm**.

5 Discussion and Conclusion

Our false alarm rate (7.8%) resides within the false alarm rates (4%-10%) reported by (Lacson et al., 2012), demonstrating promising results. Our false alarm analysis revealed several challenges for the task of critical finding identification.

For **cancer**, all three cases identified by pyConTextNLP were considered chronic by the radiologist. One report requires coreference resolution to determine that the tumor was not new: "Now with multiple nodular lesions within the bilateral lungs, demonstrating both enlarging of previously seen nodules, and development of new nodules. This is consistent with **metastatic melanoma to the lung parenchyma.**" The other two reports didn't contain explicit linguistic cues indicating chronicity: the radiologist inferred from context that the findings were chronic. One report described two lung lesions consistent with metastases then described another finding, "lytic t11 **lesion**," that should be correlated with a prior MRI. In the other report, a separate finding was described as unchanged: "a small right apical **pneumothorax** persists, and when allowing for differences in angulation, this is either unchanged or slightly increased." The mention of a previous exam, even though not directly in reference to the metastases or pneumothorax, implied that the findings were identified previously. pyConTextNLPs regular-expression-based algorithm cannot address coreference or inference.

For **aneursym**, error resolution would require either mapping different findings to different levels of severity or dropping the general synonym of "dilation" for an aneurysm: "mild **dilatation** of the main pulmonary artery, suggestive of pulmonary arterial hypertension." Identifying new cases of **pneumonia** poses similar challenges: two of 29 reports flagged with a new pneumonia were false positives. One was due to a missed negation, due to an implementation issue related to pruning targets: "*no* focal consolidation to suggest **pneumonia**." In the second, pneumonia was considered present, but as a side effect of cancer and not an infection that should be included in a critical finding follow-up list.

Limitations of this study include review by a single radiologist and only evaluating false alarms. Based on our iterative development, pyConTextNLP also missed valid critical findings, and a follow-up study will evaluate annotated reports to quantify and characterize false negatives.

Successful critical finding identification relies on negation detection, ignoring findings that are mentioned as the reason for exam, accurate differentiation of acute vs chronic findings, and modeling of uncertainty indicated by explicit cues (e.g., "may represent") as well as by linguistic variants used to describe the observation (e.g., "patchy opacity" vs. "pneumonia"). With a false positive rate of 7.8%, we believe pyConTextNLP could feasibly be deployed to suggest critical findings for communication to referring physicians without inducing alert fatigue or irritating radiologists with obvious errors. However, we will formally assess this hypothesis and determine how referring physicians would like to be presented with system recommendations in future user studies. Future work will also include assessment of false negatives, extension and evaluation of all 39 critical findings across all report types, and evaluation of execution speed and work flow integration.

Acknowledgments

We would like to thank the anonymous reviewers for valuable comments. This work was partly funded by the Department of Veteran Affairs (CRE 12-312) and University of Utah Healthcare System Hospital Project funds.

References

American College of Radiology. 2014. ACR practice parameter for communication of diagnostic imaging findings.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Brian E. Chapman, Sean Lee, Hyunseok Peter Kang, and Wendy Webber Chapman. 2011. Document-level classification of CT pulmonary angiography reports based on an extension of the context algorithm. *Journal of Biomedical Informatics*, 44(5):728–737.

Andrea Esuli, Diego Marcheggiani, and Fabrizio Sebastiani. 2013. An enhanced CRFs-based system for information extraction from radiology reports. *Journal of Biomedical Informatics*, 46(3):425 – 435.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851, Oct.

Ronilda Lacson, Nathanael Sugarbaker, Luciano M Prevedello, Ivan IP, Wendy Mar, Katherine P Andriole, and Ramin Khorasani. 2012. Retrieval of radiology reports citing critical findings with disease-specific customization. *The Open Medical Informatics Journal*, 6:28–35.

Mathieu Lafourcade and Lionel Ramadier. 2016. Semantic relation extraction with semantic patterns experiment on radiology reports. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Paras Lakhani, Woojin Kim, and Curtis P. Langlotz. 2012. Automated detection of critical results in radiology reports. *Journal of Digital Imaging*, 25(1):30–36.

Paul A. Larson, Lincoln L. Berland, Brent Griffith, Charles E. Kahn Jr, and Lawrence A. Liebscher. 2014. Actionable findings and the role of it support: Report of the acr actionable reporting work group. *Journal of the American College of Radiology*, 11(6):552–558, June.

Neel Patel, Evan Narasimhan, Danielle L. Mowery, Wendy W. Chapman, Brian E. Chapman, and Marta E. Heilbrun. 2016. Annotation of critical findings from radiology reports: towards automated communication through the electronic health record. Portland, OR. Society for Imaging Informatics in Medicine.

Anne-Dominique Pham, Aurélie Névéol, Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello, and Anita Burgun. 2014. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC Bioinformatics*, 15:266.

Meliha Yetisgen-Yildiz, Martin L. Gunn, Fei Xia, and Thomas H. Payne. 2013. A text processing pipeline to extract recommendations from radiology reports. *Journal of Biomedical Informatics*, 46(2):354–362, April.