Annotation of Critical Findings from Radiology Reports: Towards Automated Communication Through the Electronic Health Record

Neel Patel, University of Utah; Evan Narasimhan; Danielle L. Mowery, PhD; Wendy W. Chapman, PhD; Brian E. Chapman, PhD; Marta Heilbrun, MD; Brian E. Chapman, PhD (Presenter)

Background

The communication of critical findings observed in radiology imaging studies between the radiologist and the ordering provider is a key factor in providing efficacious patient care [1]. 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 may be tedious, slow, and inefficient.

Our long-term goal is to improve communication by developing a real-time natural language processing (NLP) system that leverages the pyConText algorithm [2] to initiate the automated communication of critical findings through the electronic health record. In the short-term, our goal is to develop a schema representing critical findings as well as their contexts in text and the urgency for communication to the care team.

Evaluation

The goal of this study was to develop an annotation schema as well as train and test reader agreement between trained readers in the detection of critical findings from the impression section of radiology reports.

Using a development set of 180 randomly sampled radiology impression sections, we developed a schema representing 41 unique critical findings. For each impression, the annotation tasks included recording what critical finding was reported in the impression (e.g., Finding = pneumothorax) and its context values including the existence of the finding (Existence: scale ranging from definite existence to definite negated existence), the historicity of the finding (Historicity: acute, chronic, historical), and the urgency with which of the findings should be communicated (Acuity: communicate within days, hours, or minutes) [3] (Figure 1).

Two trained readers (senior medical students interested in a career in radiology) performed the annotation of 60 randomly sampled radiology reports. We assessed inter-annotator agreement (IAA) based on two types of match criteria: 1) if both readers identified the same critical finding e.g., both readers identify Finding = aneurysm in the same impression and 2) for
each the three context types, if both readers identified the same critical finding and the same context value e.g., both readers identify Finding = aneurysm with Existence = Definite Existence in the same impression. For this study, we present the schema and report IAA using recall, precision, and F1-score. This study is Institutional Review Board-approved.

Discussion

High IAA agreement (greater than 80%) for matching critical findings was found (Table 1). The agreement dropped, but was still reasonable (greater than 70%) for matching critical findings with each context for all metrics (recall, precision, and F-score). Recall and precision were comparable. This suggests that disagreements were equally likely due to missed critical findings (and contexts) as they were due to spuriously marked critical findings (and contexts) by each student reader.

Table 1

<table>
<thead>
<tr>
<th>Match criteria</th>
<th>Example</th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Finding</td>
<td>&quot;ectopic pregnancy&quot;</td>
<td>82%</td>
<td>81%</td>
<td>81%</td>
</tr>
<tr>
<td>Critical Finding + Existence</td>
<td>&quot;Likely aneurysm&quot; Probable Existence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical Finding + Historicity</td>
<td>&quot;Known fracture&quot; Historical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical Finding + Acuity</td>
<td>&quot;Acute appendicitis&quot; Communicate within hours</td>
<td>73%</td>
<td>72%</td>
<td>72%</td>
</tr>
</tbody>
</table>

Table 1. Examples and inter-annotator agreement for critical findings and each context (existence, historicity, and acuity).

An example of a disagreement is the following: the impression sentence is, “No pneumothorax on today’s examination.” Reader 1 scored this as Finding = No Critical Finding while Reader 2 stated that while Finding = Pneumothorax, Existence = Definitely Negated Existence. As such, both are reasonable interpretations of the information regarding criticality and findings, but represent a difference in relation to the granularity of the annotation. The action following both interpretations is that no direct physician-to-physician verbal communication is necessitated. Thus, it is not clear that this represents a clinically significant disagreement. In other cases, the disagreement may be related to ambiguity in the impression and differential diagnostic considerations such that there is some semantic similarity. For example, the readers disagreed in the annotations for this impression, “Findings are concerning for abscesses. Pyosalpinx is considered less likely. The appendix is normal... Given the patient's age malignancy is thought to be less likely, however, not entirely excluded...” In this case, Reader 1 annotated Finding = appendicitis, Existence = Definite Negated Existence. Reader 2 annotated that Finding = cancer, Existence = Probably Negated Existence. Again, both readers annotated information that is present in this impression. Both readers also agreed in the more critical of the annotations, e.g., Finding = Inflammation, Existence = Definite Existence.

Conclusion

We conclude that critical findings can reliably be annotated from the impressions of radiology reports. Currently, we are leveraging these 60 reports to develop a knowledge base of lexical terms and phrases representing each critical finding and contexts for the pyConText algorithm. We are also actively annotating a larger corpus of 1,000 reports for further development and evaluation of the algorithm.


Keywords

Critical Findings, Annotation, Natural Language Processing, Reporting