Characterizing the Progression of Pulmonary Edema Severity: Can Pairwise Comparisons in Radiology Reports Help?

Stephanie M Hu[1,2], Steven Horng[2], Seth J Berkowitz[2], Ruizhi Liao[1], Rahul G Krishnan[4], Li-wei H Lehman[1], Roger G Mark[1]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]UCSF School of Medicine, San Francisco, CA, USA
[3]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[4]University of Toronto, Toronto, ON, Canada

Abstract

Accurately assessing pulmonary edema severity is critical for making treatment decisions in congestive heart failure patients. However, the current scale for quantifying pulmonary edema based on chest radiographs does not have well-characterized severity levels, with substantial inter-radiologist disagreement. In this study, we investigate whether comparisons documented in radiology reports can accurately characterize pulmonary edema progression. We propose a rules-based natural language processing approach to assess the change in a patient's pulmonary edema status (better, worse, no change) by performing pairwise comparisons of consecutive radiology reports, using regular expressions and heuristics derived from clinical knowledge. Evaluated against ground-truth labels from radiology experts, our labeler extracts comparisons describing the progression of pulmonary edema with 0.875 precision and 0.891 recall. We also demonstrate the potential utility of comparison labels in providing additional fine-grained information over noisier labels produced by models that directly estimate severity level.

1. Introduction

Pulmonary edema is a condition in which fluid accumulates in the lungs, making it difficult to breathe and ultimately leading to respiratory failure if treated improperly [1]. It is often diagnosed using chest radiographs, which are interpreted by radiologists in a radiology report. Rather than the mere presence or absence of pulmonary edema, radiologists assess the severity of the condition, which allows clinicians to make better-informed treatment decisions based on quantitative phenotyping of patient status [2]. This is particularly important for congestive heart failure (CHF) patients, who demonstrate heterogeneous responses to treatment [3].

Accurate grading of pulmonary edema is a challenging task [4] . The underlying physiology of the condition is a continuous variable, but due to the innate difficulty of estimating continuous values, standard clinical practices use a discrete scale to rank the severity instead. The severity scale ranges from 0 to 3, with 0 indicating no pulmonary edema and 3 indicating the most severe level of pulmonary edema [5, 6]. However, the boundaries between these bins are not well defined in practice and do not generalize effectively across patients, resulting in substantial overlap between different radiologist interpretations.

Over the years, there has been increasing interest in using machine learning to improve the speed and accuracy of radiograph interpretation and eliminate the subjectivity of human judgment. Since most radiological data is unlabeled, researchers have investigated the use of natural language processing (NLP) techniques on radiology reports to extract labels for the associated radiographs [7, 8]. Liao et. al. developed a program that employs keyword-matching to automatically extract pulmonary edema severity labels from radiology reports. These severity labels were used as "ground truth" for training a computer vision model to predict the severity of pulmonary edema from chest radiographs [7]. However, this keyword-matching approach makes the assumption that the radiologist who interpreted the associated radiograph correctly quantified the status of pulmonary edema in the image, a task that has been proven difficult even for experts [4].

While severity scores may not be accurate, other pieces of information documented in radiology reports may be more reliable. Often, when interpreting radiographs, radiologists reference previous radiographs in the series in order to describe how radiologic features such as fluid status have changed over time. Since making comparisons is easier than estimating values on a scale, we propose using comparison labels to extract more granular information about pulmonary edema status. In this study, we

present a rules-based NLP approach for automatically assigning comparison labels to radiology reports that document changes in pulmonary edema status. The comparison labels are used to derive comparisons between pairs of consecutive reports, which can be used for a number of applications. We hope the results can assist researchers in developing more accurately labeled datasets for modeling, better inform radiologists trying to understand the characteristics of the different pulmonary edema severity levels, and help clinicians develop more reliable tools for clinical decision-making.

2. Dataset

In this study, we used radiology reports from the MIMIC-CXR database [9]. Since the same keywords can imply different clinical findings depending on disease context, we limited our cohort to CHF patients to reduce keyword confounding as in Horng et al. [10]. The average number of chest radiographs taken per CHF patient during a single hospital stay was 13.78 (compared to 5.43 for a non-CHF patient), making it possible to generate multiple pairwise comparisons for a given patient.

We further filtered our dataset to include only consecutive radiology reports. Two radiology reports were defined as *consecutive* if the associated radiographs were acquired within 48 hours of each other, and if no other radiographs were performed in between. In total, we used 7,141 radiology reports comprising 4,896 pairs across 1,114 patients in our study. Reports that were both preceded and followed by another report were included in two distinct pairs.

Given a pair of consecutive radiology reports r and r' written at time t and t', respectively, where $t' > t$, we define a *comparison* to be any description of change on the patient's pulmonary edema state between time t and t'. This change is captured in the text of r', so the comparison label identified for the pair (r, r') is simply the comparison label extracted from the document r'. Comparisons are either *worse*, *better*, or *no change*. We also used a *no comparison* label for documents that contained no explicit comparisons about pulmonary edema in the text. In our study, we assumed that any comparisons described in a given radiology report were made relative to the report dated immediately before it.

2.1. Training Set

Our labeler was developed using a training set of 257 radiology reports across 34 patients, where each radiology report was involved in one or two consecutive pairs. Board-certified radiologists provided sentence-level comparison labels for sentences extracted from the "Findings" and "Impressions" sections of the radiology reports. These sentences were first identified as being relevant or not relevant to describing pulmonary edema status, and the relevant sentences were further assigned a comparison label capturing the change in severity of the condition. The 272 sentences that were considered to be relevant to pulmonary edema status yielded the following distribution of comparison classes: 36 worse, 40 better, 88 no change, and 108 no comparison.

2.2. Testing Set

To evaluate the performance of our labeler in constructing pairwise comparisons at the document level, we randomly selected 101 pairs of consecutive radiology reports that had been labeled as one of *worse*, *better*, or *no change* by the labeler. None of the documents in this set overlapped with the training set. A board-certified radiologist, blinded from the results of the labeler, provided manual comparison labels for these pairs at the document level only. Of the 101 pairs, the radiologist provided the following distribution of comparison labels: 24 worse, 26 better, 49 no change, and 2 no comparison. The 2 pairs of reports with the *no comparison* label were excluded from analysis, yielding a final testing set size of 99.

3. Pairwise Comparison Labeler

An automatic rules-based labeler[1] was developed for assigning pairwise comparison labels to consecutive radiology reports written in free text. The labeler comprised three stages: 1) identifying sentences relevant to pulmonary edema, 2) identifying comparisons in individual sentences, and 3) constructing document-level pairwise comparison labels from sentence-level labels (Figure 1).

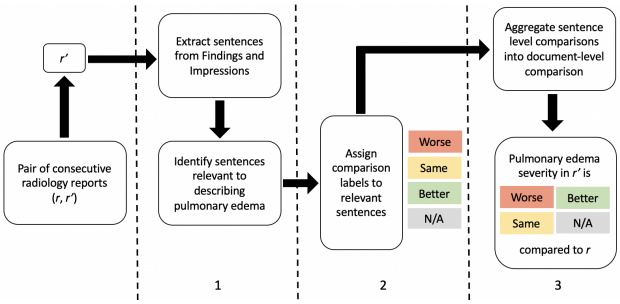

Figure 1. Labeler pipeline in three main stages.

3.1. Identifying Relevance

Any sentence containing either a positive or negative mention of pulmonary edema was considered relevant. We used CheXpert and a curated set of *regular expressions* (character sequences that define a search pattern) to identify these mentions. CheXpert is a rule-based labeler that extracts 14 different observations from the free text of radiology reports, including pulmonary edema [8]. Expanding

[1] Project code available at https://github.com/shu98/pulmonary-edema-project.

on the keywords curated by Liao et al. [7] and Irvin et al. [8], and with expert guidance from clinicians, we additionally constructed a more comprehensive list of regular expressions capturing (a) the various ways that pulmonary edema can be described in a radiology report and (b) radiologic findings that are related to, but not definitive for, the presence of pulmonary edema [5, 6]. Sentences mentioning related radiograph findings were only considered relevant if they did not contain mentions of other findings that could represent a differential diagnosis.

3.2. Identifying Comparisons

Only sentences that were considered relevant to pulmonary edema were assigned a comparison label. We used a rules-based approach to assign one of four classes (*better*, *worse*, *no change*, or *no comparison*), using a set of regular expressions developed in our study to capture the directional change for each of the categories. The presence of a comparison phrase from one of the *better*, *worse* and *no change* categories was used to assign the appropriate comparison label. Once comparison phrases were identified, we used the Negex library to determine whether the comparison was a positive or negative mention. Absence of any comparisons resulted in a *no comparison* label.

3.3. Constructing Pairwise Comparisons

Given the sentence-level comparison labels for a radiology report, we developed a set of rules to generate a single document-level comparison label. If all relevant comparison-containing sentences in the report were assigned the same class, then the radiology report was also assigned to that class. In the case of discrepancies between sentence-level labels, we employed a "last-first" approach, in which the last sentence describing a change in pulmonary edema status was given the highest precedent in determining the label of the entire radiology report. In a given pair of consecutive radiology reports (r, r'), the pairwise comparison label was taken from r'.

4. Results

We evaluated the performance of the first and second stages of our labeler against our training dataset, which contains radiology reports labeled at the sentence level. All 1,492 sentences in this set of radiologist reports were provided a relevance label by a radiologist, and the 272 sentences that were considered relevant to pulmonary edema were provided a comparison label.

4.1. Identifying Relevance

The goal of this first stage was to identify sentences relevant to characterizing pulmonary edema status in radiology reports. Because these sentences would later be aggregated to form document-level labels, we wanted to capture as many sentences as possible in this step. It was therefore more important to prioritize recall over precision. Our labeler achieved an accuracy of 98.8% on this task, with a precision of 0.96 and a recall of 0.98.

4.2. Identifying Comparisons

We restricted our experiments and evaluation of this stage to the 272 sentences that were identified by a radiologist as relevant in determining pulmonary edema status. Metrics for individual classes were computed using a one-vs-all approach and are presented with the overall performance across classes in Table 1.

Table 1. Identifying sentence-level comparisons.

Label	Accuracy	Precision	Recall
Worse	0.974	0.837	1.00
No change	0.971	0.955	0.955
Better	0.992	1.000	0.950
Overall	0.967	0.949	0.945

The performance of our labeler in assigning pairwise comparisons to pairs of consecutive radiology reports was evaluated against a testing set of reports that were not used in the development of any stage of the labeling pipeline. The results are displayed in Table 2.

Table 2. Identifying document-level comparisons.

Label	Accuracy	Precision	Recall
Worse	0.916	0.846	0.917
No Change	0.959	0.904	0.959
Better	0.808	0.913	0.808
Overall	0.909	0.875	0.891

4.3. Common Errors

While our labeler correctly labeled the majority of radiology reports, there were some cases in which it failed due to the presence of more complex sentences. For example, when a sentence contained multiple observations close together, all with different modifiers, the labeler occasionally selected the wrong modifier to associate with the change in pulmonary edema status (e.g. "Exam is otherwise remarkable for *improving* asymmetrical pulmonary edema and apparent *increase* in size of bilateral effusions", which contains the comparison words "improving" and "increase" in close proximity to the pulmonary edema mention.) Additionally, our labeler also struggled with sentences in which the same modifier applied to multiple observations and the modifier lay too far away from the pulmonary edema observations (e.g. "*Unchanged* evidence of moderate car-

diomegaly, atelectasis, and overall mild-to-moderate *pulmonary edema*").

4.4. Comparison to Severity Labeler

To illustrate the advantage of extracting comparison labels from radiology reports over directly identifying the severity of pulmonary edema, we compared the output of our comparison labeler to the output of the keyword-based severity labeler developed by Liao et al. [7]. Because we wanted to see how pairwise severities compared to pairwise comparison in our evaluation, it was only meaningful to consider pairs of reports in which both reports had been assigned a severity label and in which the overall pair had been assigned a comparison label. In total, there were 243 such pairs. Table 3 summarizes the discrepancies in output labels between the two methods.

Table 3. Distribution of class labels assigned by comparison versus severity labelers.

		Severity Labeler		
		Worse	No change	Better
Comparison Labeler	Worse	27	35	7
	No change	15	92	24
	Better	12	27	10

4.5. Comparison to Computer Vision Model

We also compared the comparison labeler with the computer vision model developed by Liao et al., which outputs severity labels for pulmonary edema based on radiograph images [7]. Again, we only considered pairwise radiograph studies in which both images had a severity label and the pair of reports had a comparison label, and we analyzed the document-level comparison label from the comparison labeler and the signed difference between severities from the computer vision model. Out of the 2,404 pairs that were considered, 47% of them showed discrepancies between the two labelers (Table 4). Interestingly, of all the pairs classified as *better* and *worse* by the comparison labeler, 43% of those pairs were indicated to have *no change* by the computer vision model, supporting the hypothesis that the computer vision model may not have taken into account more fine-grained differences between pulmonary edema status within the same severity class.

Table 4. Distribution of class labels assigned by comparison labeler versus computer vision model.

		Severity Labeler		
		Worse	No change	Better
Comparison Labeler	Worse	248	243	71
	No change	249	729	275
	Better	48	254	287

5. Conclusion

In this study, we presented a rules-based approach that achieves high performance on assessing directional change in pulmonary edema severity between consecutive radiology reports. This work may help clinicians gain more comprehensive insights into the pulmonary edema severity spectrum and better characterize the radiographic features that define each severity level. To the best of our knowledge, our work is the first attempt to automatically characterize pulmonary edema progression from radiology reports. In the future, we would be interested in exploring more advanced algorithms to capture the complex cases that our current labeler misses, improving computer vision models by incorporating comparison labels during training, or constructing approximate rankings of pulmonary edema severity from pairwise comparisons.

Acknowledgments

This research was in part funded by NIH grants R01 EB030362 and R01 EB017205. L. Lehman was in part funded by the MIT-IBM Watson AI Lab.

References

[1] Mahdyoon H, Klein R, Eyler W, Lakier JB, Chakko S, Gheorghiade M. Radiographic pulmonary congestion in end-stage congestive heart failure. Am J Card. 1989 Mar 1;63(9):625–7.

[2] Chakko S, Woska D, Martinez H, et al. Clinical, radiographic, and hemodynamic correlations in chronic congestive heart failure: conflicting results may lead to inappropriate care. Am J Med. 1991 Mar 1;90(1):353–9.

[3] Francis GS, Cogswell R, Thenappan T. The heterogeneity of heart failure: will enhanced phenotyping be necessary for future clinical trial success? J Am Coll Cardiol. 2014 Oct 21;64(17):1775-6.

[4] Hammon M, Danker P, Voit-Höhne HL, Sandmair M, Kammerer FJ, Uder M, Janka R. Improving diagnostic accuracy in assessing pulmonary edema on bedside chest radiographs using a standardized scoring approach. BMC Anesthesiol. 2014 Oct 18;14:94. PubMed PMID: 25364301.

[5] Cremers S, Bradshaw J, Herfkens F, editors. The Radiology Assistant: Heart Failure [Internet]. Radiological Society of the Netherlands; 2010 [cited 2020 Mar 22]. Available from: https://radiologyassistant.nl/chest/chest-x-ray/heart-failure.

[6] Gluecker T, Capasso P, Schnyder P, et al. Clinical and Radiologic Features of Pulmonary Edema. Radiographics. 1999 Nov 1;19(6):1507–31.

[7] Liao R, Rubin J, Lam G, et al. Semi-supervised learning for quantification of pulmonary edema in chest X-ray images. arXiv: arXiv:1902.10785 [Preprint]. 2019 [cited 2020 May 12]: [12 p.]. Available from: https://arxiv.org/abs/1902.10785.

[8] Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Jul;33:590-7.

[9] Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019 Dec 12;6(317):1-8.

[10] Horng S, Liao R, Wang X, Dalal S, Golland P, Berkowitz SJ. Deep learning to quantify pulmonary edema in chest radiographs. Radiology: Artificial Intelligence [Internet]. 2021 Jan 6 [cited 2021 Mar 5]: e190228. Available from: https://pubs.rsna.org/doi/abs/10.1148/ryai.2021190228.