Author's response to reviews

Title: Automated identification of pneumonia in chest radiograph reports in critically ill patients

Authors:

Vincent Liu (vincent.x.liu@kp.org)
Mark P Clark (Clarkmark@rocketmail.com)
Mark Mendoza (mark.x.Mendoza@kp.org)
Ramin Saket (Ramin.R.saket@kp.org)
Marla N Gardner (Marla.N.Gardner@kp.org)
Benjamin R Turk (Benjamin.R.Turk@kp.org)
Gabriel J Escobar (Gabriel.Escobar@kp.org)

Version: 2 Date: 13 June 2013

Author's response to reviews: see over
Response to reviewers – MS: 7718537749436367
‘Automated identification of pneumonia in chest radiograph reports in critically ill patients’

(Reviewer comments are listed and condensed for clarity).

Reviewer 1: Wendy Chapman

1. The authors should cite and describe more related work including that related to hedge detection in text (e.g., the Bioscope corpus). Because uncertainty identification is one of the listed strengths of the manuscript, it would be important to help readers understand the state of the art in this area. I think the modeling of uncertainty makes the manuscript a novel contribution to the literature.

We thank the reviewer for their helpful comments and agree that additional discussion related to prior NLP work in uncertainty and hedge detection is necessary. We have made several edits throughout the manuscript to address this issue including adding three citations (e.g., the suggested reference detailing the development of the BioScope corpus by Vincze et al in BMC Bioinformatics 2008). Additionally, we have extensively edited the Discussion by including the following new paragraph:

Prior NLP studies have also evaluated the role of uncertainty in accurately interpreting biomedical reports.[28-30] For example, Vincze and others describe the development of the BioScope corpus which is annotated for a wide range of negations and linguistic speculations.[28] Many of the uncertainty profiles we captured in our lexicon are also described by the BioScope investigators including syntactic structures that connote ambiguity through auxiliaries, adjectives, or adverbs that are associated with keywords of interest. While the BioScope corpus contains free text from a wide variety of sources, including medical texts, biologic manuscripts, and abstracts, our corpus is drawn from a relatively proscribed source with a set of common and well-defined terms and phrases. As a result, the uncertainty profiles used in our NLP queries may have limited applicability to other free text sources. For example, common phrases in CXR reports like ‘cannot exclude infiltrates’ may be infrequently used in scholarly manuscripts or medical texts.

2 and 3. I do not think that the evaluation reflects the focus on uncertainty since it only evaluates the ability to identify positive and negative reports, seeming to leave out the 35% of uncertain reports. I suggest reporting statistics on identification of reports that were classified as uncertain since for practical reasons it may be necessary to assign reports to three bins: positive (no human review necessary), negative (no human review necessary), and possible (needs human review). A similar technique was used in a recent paper (Dublin et al: Natural language processing to identify pneumonia from radiology reports. Pharmacoepidemiol Drug Saf 2013). Table 3 could show results for each bin and the overall performance of the system. It could also be interesting to look at performance if uncertain classifications were lumped with negative or positive.

We have followed the reviewer’s suggestions to include data regarding the radiograph reports with uncertain interpretations. For example, we have updated Table 3 by including data regarding the test characteristics of our electronic algorithm in the 30-40% of reports that were deemed ‘possible’. We have also clarified Table 3 headings to indicate that test characteristics account for combining studies together—for example,
when evaluating the test characteristics for the ‘possible’ category, those in either the ‘negative’ or ‘positive’ categories are combined.

We have also added the citation suggested above (Dublin et al, 2013) and have extensively revised the Discussion as follows:

A recent study by Dublin and others evaluated the performance of an open-source NLP system (ONYX) to assist with differentiating electronic CXR reports that required further manual review from those that could be conclusively labeled as ‘consistent’ or ‘inconsistent’ with pneumonia.[26] Out of 5,000 reports, between 12% and 25% were determined as requiring additional manual review—a lower, but still substantial, number of reports compared with our study. In their study, some criteria used to determine which reports required manual review were similar to those in our study (e.g., the presence of both atelectasis and pneumonia). In the remaining reports, their NLP system demonstrated excellent test characteristics similar to, or better than, those reported in prior NLP CXR report studies.[6, 8, 9, 26, 27] However, it is important to note the substantial differences in the patient populations from which the CXR reports were obtained. In the Dublin study, for example, 92% of reports were from outpatients—a population in whom radiographic image quality is expected to be higher and features like atelectasis or infiltrates are expected to be less prevalent.[26]

4. I suggest a more extensive discussion of both the processing of uncertain classification (the 27 uncertainty profiles) and their performance. I think the supplementary table needs to be included. I agree that the proprietary nature of the system limits the ability of others to test the system. The manuscript should give a general description of how the algorithm/system uses the lexicon to identify pneumonia-related findings.

We have followed the suggestion to move e-Table 1 into the main manuscript (now Table 1). We also provide an overview of the 20 steps in the electronic algorithm in the Supplement (e-Table 1). We have also extensively modified the Methods sections of the manuscript to improve our description of the system. For example, we have rewritten the following paragraphs:

Using the gold-standard physician interpretations in the development and derivation sets, we then developed an electronic algorithm for assigning interpretations to CXR reports. The algorithm included twenty steps where each step incorporated rules- or probability-based strategies to analyze combinations of NLP query hits (e-Table 1). For example, a CXR report that included a ‘blanket normal’ statement (e.g., ‘no acute cardiopulmonary findings’) without any other pneumonia terms would be assigned a ‘negative’ interpretation. A report that included only pneumonia terms within high uncertainty profiles (‘infiltrate versus atelectasis’) would be assigned a ‘possible’ interpretation.

Because many reports included hits from several query elements that precluded simple rules-based interpretation, we also incorporated a set of predicted probabilities in selected algorithm steps. Using the development and derivation sets, we generated three logistic regression models to assign predicted probabilities that each report would have a ‘positive’, ‘possible’, or ‘negative’ interpretation. These probabilities were generated using backward stepwise logistic regression where NLP query hits associated with the binary outcome (e.g., for the ‘negative only’ outcome, negative=1 and positive or possible=0) with a p-value <0.2 were retained in the final model. The beta-coefficients, based on the derivation sample, were then used to calculate probabilities in the validation sample (e-Appendix 1). These probabilities were then used in concert with NLP query profiles to assign interpretations to reports that could not be classified simply with rules-based
approaches. For example, after removing reports interpreted in the prior 11 steps, step 12 deemed a report ‘negative’ if its ‘negative’ predicted probability was >30%, its ‘possible’ probability was <30%, and its ‘positive’ probability was <10%.

5. It would be very interesting to see how this algorithm performs when using an open-source system like pyConText, however, I don’t believe this is necessary for publication.

We agree with reviewer that it would be ideal to test the algorithm using an open-source NLP system and we are highly interested in pursuing this in a future research study. With our existing set of cases, we could compare the performance of an open-source adaptation to the existing I2E system.


As we have attempted to describe above (answer to question 4), few of the ‘uncertainty’ phrases directly map to final classifications through simple rules-based mechanisms. We provide an example of a simple rules-based step in which a pneumonia-related term in a ‘high uncertainty’ phrase (e.g., ‘infiltrate versus atelectasis’) in isolation (without other normal or low uncertainty pneumonia-related terms present) would map directly to a ‘possible’ interpretation. However, many of the terms and uncertainty phrases are used to calculate the predicted probabilities employed in the later algorithm steps.

7. The manuscript needs more detail about physician interpretation. Did the physicians read and classify each report independently or did they come to consensus. If it was by 2 physicians, the authors could report agreement statistics.

We have clarified our Methods as follows:

For each report, two physicians experienced with interpreting ICU CXR reports reached a consensus on whether the report was ‘positive’, ‘possible’, or ‘negative’ for pneumonia in a presumed scenario where CXRs were performed in patients whose clinical differential diagnosis included pneumonia (e.g., a patient with dyspnea).

8. It wasn’t clear to me what the ‘other’ category represents.

We have clarified the Methods as follows:

Two physicians experienced with critical care reviewed >1,000 CXR reports to empirically develop a lexicon focused on categorizing features associated with pneumonia (e-Table 1) within three broad categories: (1) terms and term groups; (2) uncertainty profiles; and (3) ‘other’ features. .... The lexicon also encoded ‘other’ features relevant to interpreting radiograph reports including those related to disease progression (‘worsening of infiltrates’), anatomic location (‘bilateral opacities’), and stability (‘unchanged from prior’).

Reviewer 2: COSMIN A BEJAN

10. To assess the validity of the system proposed in this paper, the authors need to compare it against a baseline system. The baseline could be a previously proposed system or one based on simple rules derived from the pneumonia lexicon. For example, a baseline system could classify a CXR report as positive based on at least one
pneumonia-related term or term group.

In the current study, we validated our system against the standard of physician interpretation rather than against an alternative NLP interpretation system since there is no existing automated system that matches the quality of manual review.

However, in preparation for using our NLP system for an ongoing quality improvement effort evaluating pneumonia in hospitalized patients, we did compare the performance of our system to manual review in a set of different CXR reports. Trained medical chart abstractors evaluated serial CXR reports in patients with a hospital diagnosis of pneumonia to evaluate the non-negated presence of a collection of terms that could be seen in pneumonia (e.g., pneumonia, infiltrate, air bronchogram, consolidation, opacity, etc). In a sample of 6,031 CXR reports, over 90% of the CXRs deemed ‘negative’ by reviewers were also assigned a ‘negative’ interpretation by our system. More than 94% of the CXRs deemed ‘positive’ by reviewers were assigned a ‘possible’ or ‘positive’ interpretation by our system. These results were presented in an abstract at the American Thoracic Society 2013 research conference and we plan to publish them in a future research manuscript.

11. Please describe more details about how the query strategies are applied to determine the CXR report category. What are the exact steps from e-Figure 1 for categorizing that sample as pneumonia? Are the queries prioritized? If the decision is ‘positive’, is that because the majority of hits associated with the report correspond to ‘positive’ queries?

Please see the response to question 4 and 6 above for additional clarification.

The algorithm steps are prioritized so that reports that can be interpreted in the simpler rules-based steps are removed leaving a typically more complex set of reports for the rules- and probability-based steps. The sample demonstrated in e-Figure 1 contains a variety of query hits precluding a simple rules-based interpretation. For example, it contains multiple elements that denote ‘high uncertainty’ for pneumonia-related terms (e.g., infiltrates/consolidation, correlate clinically, exclude infiltrates), no uncertainty for other pneumonia-related terms (e.g., opacity), and no uncertainty for non-pneumonia processes (e.g., pulmonary edema/CHF/effusions). It also includes a variety of other terms (e.g., interval development). In this case, the algorithm uses the set of predicted probabilities to determine an interpretation with special emphasis on the potential for ‘infiltrates’ (as a sign of pneumonia) to denote a ‘possible’ report. We have outlined an overview of each of the 20 steps in the Algorithm in our updated Supplemental material.

12. How was the entire set of 194,615 reports used to solve the classification task. Please explain why it is necessary to index all 194,615 reports. Would the results be different if they weren’t all indexed?

We thank the reviewer for helping us clarify the manuscript. It was necessary to index all 194,615 CXR reports so that a sample could be randomly drawn from this full sample. However, the results of the algorithm and NLP queries would not differ if only the smaller sample of reports were indexed. We included all the CXR reports to provide a broad scope of how many ‘hits’ would be generated in a large sample of reports.

13. The authors should report results for each of the assertion categories and not only
for positive/negative categories.

We have updated Table 3 as suggested.

14. The main goal of accurately classifying the positive reports should be emphasized a bit more. It would be helpful if you can discuss why your approach did not perform so well in identifying this type of report.

In the sample of almost CXR reports manually reviewed by practicing clinicians, we found that very few reports conveyed conclusive language for ‘positive’ pneumonia. As a result, we believe that our system had difficulty evaluating all the nuances conveyed by radiologists in their reports.

We have clarified the Discussion as follows:

In this study, we evaluated a large sample of chest radiograph reports from critically ill patients. Among nearly 2,500 reports categorized by manual review and physician consensus, 42% could not be classified as either ‘negative’ or ‘positive’. In many cases, these ‘possible’ reports included language from interpreting radiologists that conveyed frank uncertainty about whether the findings represented pneumonia or another condition with an appearance similar to pneumonia. In these cases, interpreting physicians felt that additional clinical information, beyond the CXR report, were necessary to determine whether a pneumonia was present or absent. Only a minority of reports (6.5%) included language that was deemed conclusive for, or highly likely to be, pneumonia.

15. Some of the counts do not add up correctly in your paper. For example, in Table 3, the total should be 2,466. Also, the ‘possible’ count should be 1,029. From e-Table 3 there were 21 steps instead of 20 steps.

We have clarified these inconsistencies.

16. What do you mean by ‘empirically’ developing a lexicon from the initial CXR reports?

We reviewed >1,000 CXR reports and recorded all of the pneumonia terms, non-pneumonia terms, and uncertainty phrasing used by interpreting radiologists to inform the development of the lexicon. We use the term ‘empiric’ to denote that we attempted to build the lexicon based on the natural language we discovered in these reports.

Reviewer 3: Michael Klompas

17. The major thing missing from the study is a discussion of the potential applications for this work. While at first blush it appears obvious—it could be used for surveillance or for studies on risk factors and treatments for pneumonia or for clinical decision support—but on further reflection I am not so sure. As the authors indicate, a substantial number of CXRs in critically ill patients are ambiguous. Where will this work find its place clinically?

We thank the reviewer for this clarification. We believe there are several potential applications for this type of system, especially in quality improvement efforts and clinical decision support. For example, in our healthcare system, we are planning to deploy this system to evaluate all CXR reports among hospitalized patients to determine the
baseline incidence of hospital acquired pneumonia to inform our quality improvement efforts. Other systems also use NLP interpretation to trigger pneumonia clinical decision support (e.g., electronic tools deployed by the Intermountain Health group in Utah). We have revised our Discussion as follows:

This set of tools could be useful in a variety of healthcare domains. For example, in our healthcare system, quality improvement efforts are aimed at reducing the frequency of healthcare- or ventilator-associated pneumonia; however, these efforts are limited by the resource strain of reviewing CXR reports among all hospitalized patients to identify relevant cases.[2, 31] Our tool could be used to automatically evaluate all CXR reports in hospitalized patients and flag those whose cases require further detailed review. This tool could also be used in conjunction with electronic decision support tools that aid clinicians in correctly triaging patients and choosing appropriate antibiotics.[11, 25, 31, 32]

18. Please clarify how ‘validation/derivation’ differs from ‘overall’. I presume it’s the addition of the development CXRs. If so, please include their performance characteristics in this Table as well in order to make the “overall” classification more transparent.

We have revised Table 3 in order to include the ‘Possible’ category in addition to the ‘Negative’ and ‘Positive’ categories. Thus, we have eliminated the ‘validation/derivation’ row to streamline the table.