Reviewer's report

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

Version: 1 Date: 18 April 2013

Reviewer: COSMIN A BEJAN

Reviewer's report:

This paper describes an approach for automatically classifying CXR reports of critically ill patients as 'positive', 'possible', or 'negative' for pneumonia. To perform the classification task, this approach employs manually crafted rules integrated into a text retrieval platform.

Major Compulsory Revisions

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 for this type of classification or a system based on simple rules derived from the pneumonia lexicon. For instance, a baseline system for classifying a CXR report as positive for pneumonia could be a system that identifies in the report at least one pneumonia related term (or term group) which has a positive profile (e.g., without being preceded by words such as 'no' or 'possible').

Please describe into more details how the query strategies are applied in order to determine the assertion category of a CXR report. I think it would be helpful if you can use an example to elaborate on this methodology. For instance, what are the exact steps for categorizing the example from e-Figure 1 as possible for pneumonia. Do you prioritize your queries? In the event a report contains elements which are indicative for all three assertion categories, how do you decide on the most probable category? If you decide on, e.g., 'positive' is it because the majority of hits associated with this report correspond to 'positive' queries?

Minor Essential Revisions

It is unclear how the entire set of 194,615 reports was used to solve the classification task. Also, please explain why it is necessary to index the entire collection of 194,615 reports. Will the results be different if you would index only the sampled set of 2,465 or 2,466 (see below!) reports?

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

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

Discretionary Revisions

Some of the counts do not add up correctly in your paper. For example, in Table 3 the total count should be 2,466 ( = 739 + 950 + 777) instead of 2,645. Also in Table 3, the count for 'possible' should be 1,029 instead of 1,028. Similarly, from e-Table 3, I counted 21 steps instead of 20 steps how is mentioned in the paper. The first 3 groups are represented by 11 steps and not 10 how is specified in the paper. Please carefully review all these numbers.

"... we empirically developed a lexicon to categorize ...", "... 1,000 CXR reports to empirically develop a lexicon focused ..." This is a bit confusing to me. What do you mean by "empirically"? My initial understanding was that the lexicon was manually created by physicians as it is suggested in this passage: "In the development and derivation sets, the physicians who created the lexicon and queries ...". Please explain.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests