Reviewer's report

Title: Detecting inpatient falls by using natural language processing of electronic medical records

Version: 2 Date: 16 October 2012

Reviewer: Anthony Nguyen

Reviewer's report:

Title: Detecting inpatient falls by using natural language processing of electronic medical records

Reviewer's report:

Major Compulsory Revisions

The paper reports the need for “quick identification of injurious falls” and “detecting adverse events in at timely and cost-effective way”, and proposes to “promptly detect severe injuries after falls by using the NLP method”. It is not clear what the purpose or benefits of quick identification is? Are you trying to detect falls immediately as it occurs, which makes sense in terms of having appropriate care for the patient. However this will not be the case with the proposed NLP system as the incident will have to be documented and subsequently scanned by an NLP system to detect the occurrence of falls. As a result, the author needs to clarify why the lag time is such an important factor e.g. why the current lag time exhibited in the study data is not suitable.

Thank you for including the overview of the workflow required to enter data into the hospital information system (HIS). It is however not explicit whether the image order entry and discharge summaries are part of the HIS or EMR from reading the description? The first sentence in the “Data collection and ethical consideration” section also aims to address this but did not successfully achieve it. However, from reading the remainder of the paper, it’s inferred that the progress notes and discharge summaries are from the EMR, while incident reports and image order entries are from the HIS. Assuming that this is the case, it seems like that progress notes and discharge notes will never be entered into the HIS, and as a result how would one derive the lag time and degree of harm for such data sources? The purpose of including data sources from EMRs in the study needs to be justified to make the paper coherent. These system and workflow constraints are important when discussing the implications of lag time and degree of harm in the paper.

Regarding the NLP rules development, the set of syntactic rules were developed from a development set of incident reports, but was then applied to all other data sources to detect falls. The writing style and contents of an incident report would be vastly different to any of the other data sources in the study. Was there a reason for pursuing such a methodology which would ultimately tailor the NLP
algorithm to incident reports?

The evaluation of the system still uses a varied set of measures between the experiments. Noteworthy is that specificity was omitted when analysing the NLP system against the various data sources (excluding incident reports). Can this measure be included for completeness? Otherwise, justification as to its omission would be beneficial to the reader.

Another issue with the results from “Fall events detected from each data source” is that there’s not enough information for the reader to analysis the results. The number of reports for each data source needs to be reported. Furthermore, some of the reported measures do not look right, for example, the image order entries had 10 fall events, however the sensitivity was 0.83 (i.e. you can only get 8 out of 10 recall, but not 8.3 out of 10). Showing the confusion matrix for the 4 data sources would overcome these issues and make it clearer.

The degree of injury and lag time between falls and submission of data to hospital information system can be determined regardless of the NLP application. As a result, I would have liked to see a profile of report types against the degree of harm and lag time. However, only incident reports and image order entries were used. How could this be done for other data sources? It just so happens that the selected reports for lag time and degree of harm analysis were both from the HIS. What if the NLP system performed better on other data sources?

The reporting of results for the “Comparison of incident reports and image order entries as text data sources” involved no gold standard comparison. We have no idea whether the detected falls were actually correct or the extent of the number of missed falls. The same methodology of analysis would have been better on the data used for “Fall events detected from each data source” where a gold standard did exist.

The lag time was found to be significantly shorter for image order entries than incident reports. Since incident reports contain numerous mild incidents, this could cause biases towards the reported results. It would be good to compare the lag time separately for each degree of harm. It could be the case that the lag times for severe cases are much shorter than those that are mild from incident reports.

It was reported in the Background section (1st paragraph) that the “judgement about adverse events in medical charts depends on the skills of reviewers”. Elaborate on this finding and determine the skill level of the single reviewer that was used in the paper. Justification based on the skill level of the single reviewer is required so that you have evidence to support the use of only one reviewer in the study. Otherwise, the evaluations of the NLP system may not be conclusive.

Minor Essential Revisions

Abstract (paragraph 3): “imager order entries” should be “image order entries”

Abstract (paragraph 3): Elaborate and clarify on the purpose of selecting incident
reports and image order entries as possible data sources. Was the purpose for reducing under-reporting, delay of submission, and/or for detecting serious falls?

Background (paragraph 2): What were the performances of the “sophisticated NLP algorithms”? This is worth noting to compare against the proposed “simpler” NLP algorithm.

Methods (NLP of free-text and construction of syntactic category rules): Elaborate on what an “antero-posterior relation” means? This is a medical term used in the context of an NLP relation between morphemes. In the second example, it was rather difficult to comprehend since it has two interpretations. Revise the example and maybe mention that “falling” is an example of “find a situation”.

Methods (Comparison of lag-time and degree of harm): How many image order entries were used? The degree of injury can be readily determined for the incident report from the HIS (see Methods – Data collection and ethical consideration), however, how was the degree of injury determined for the image order entries?

Results (Fall events detected from each data source): The number of patient days was used for number of patients. Detail both or be specific as to what the value refers to.

Results (Fall events detected from each data source): The location of Fig 1 reference was inserted when discussing the performance of the NLP system on image order entries. This should either be at the start or end of the results being reported.

Results (Comparison of incident reports and image order entries as text data sources): The number of mild events found from incident reports, order entries or both does not add up to the total of mild events. Also check the correctness of the number of degrees of injury in fall events found by incident reports and image order entries. This should be reviewed for its correctness.

Results (falls with moderate to severe injuries): Mention that the statement “image order entries increased the detection rate of falls” was for “moderate to severe injuries”.

Results (falls with moderate to severe injuries): It would be good to know the 3 injuries that were only from image order entries and the 3 that were from only the incident report. This will help identify the types of moderate to severe injuries that can be found in one but not the other, and consequently provide input towards the best data sources for detecting severe injuries.
Discussion (paragraph 3): Correct “NPL”; “loss <of> information”

Discussion (paragraph 3): Elaborate on what was meant by “positive results were obtained in only 3.0%”?

Discussion (paragraph 4): Revise “This report is probably the first…” to “To the best of our knowledge this is the first …”

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

Quality of written English: Needs some language corrections before being published

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