Reviewer's report

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

Version: 1 Date: 20 August 2012

Reviewer: Anthony Nguyen

Reviewer's report:

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

Major Compulsory Revisions

1. The paper aims to motivate the use of NLP to fast track fall incidents using incident reports as well as other data sources such as image order entries, progress notes and discharge summaries. The author notes that current practices reveal that there is under/non-reporting and delay of submission of fall data, which prevent daily use or real-time detection of fall events and propose to overcome these through detection of fall events from EMRs. With respect to the delay of submission issue, one would assume that adverse events are attended immediately as it happens and not retrospectively after the adverse event has been recorded in an EMR. In addition, image order entries was proposed to lessen the lag time, however, this only accounted for 10 out of the possible 80 fall events. It is also not clear from the paper whether the study addressed the under/non-reporting issue when image order entries detected an additional 3 fall events on top of the 52 fall events detected by the incident reports; still there’s another 25 fall events still not accounted for. The “most useful source of information” and “degree of harm” was also studied; however, it was not clear how this relates to the aim of the study. As a result, it was unclear what the motivation and aims of the paper were, and whether it was achieved through the use of NLP from EMR.

2. It would help understand the context of the work better if a general overview was provided for the events and data flow that occur from an inpatient fall through to the possible submission to the “voluntary” incident reporting system as well as the hospital information system. This affects the number and type of fall incidents found from each data source. For example, one would suspect that image order entries are likely to be requested if the “degree of harm” is moderate to high, thus causing the bias that the author reports as being statistically significant. On the other hand, Incident reports (although voluntary) can be reported for any incident and degree of harm. Describing such an overview will also help explain and understand some of the findings that the author reports. Furthermore, it would benefit from a description of the possible reasons for the lag that occurs from a fall incident to the submission to the hospital information system (e.g. paper based process, manual vs. automatic entry into hospital
information system, regular updates, etc).

3. Regarding the collection of fall data from the various sources, little was described on how each report was categorised as fall-related or fall-unrelated. For example, was there a field in the incident reporting system flagging falls, thus generating the gold standard for comparisons against the NLP system? If not, how was the gold standard generated? How many experts were involved with the manual review? Were all reports reviewed by experts? What was the agreement between the experts? These questions also need to be addressed for the other data sources collected for the study. In particular from the Results (Fall events detected from each data source), it looks suspicious that only those flagged as falls by the NLP system were reviewed for correctness. The data collection and gold standard generation methodology is required to be made explicit to address the above potential misunderstandings.

4. The NLP rules were developed from a development set of incident reports, and was subsequently used for evaluation on all other data sources. Was there a reason for pursuing such a methodology? One would assume that the writing style and contents of an incident report would be vastly different to any of the data sources studied.

5. The results reported require an expectation baseline to compare against. For example, results with respect to reported performances by other systems used for the same task, either for fall detection (even for non-Japanese NLP systems) or other subspecialties need to be presented in order to present some kind of an expectation baseline. Simple machine learning approaches such as SVMs can be readily applied and have proven to work well for such classification tasks and can be used as an alternative baseline for comparisons.

6. The author reports on the NLP detection of falls results for various data sources in terms of sensitivity and PPV. As one can see, high recall data sources have low PPV, while high PPV data sources have low sensitivity. A trade-off is expected in terms of sensitivity and PPV, however, the results are at one extreme operating point, where one metric performs very well, while the other is rather poor. A metric that combines both sensitivity (recall) and PPV (precision) is F-measure. This measure is commonly used for classification to compute a single metric for the system. The author is encouraged to use this metric to report on the overall system performance.

7. Another issue regarding performance metrics is that the abstract and the paper emphasised on the use of sensitivity and PPV, however, within the “Validation of the performance of category decision rules,” specificity was used instead. What was the justification for using different metrics? It is suspected that specificity was not used in latter analyses (of non-incident reports) due to the need to review many cases of fall-unrelated data sources, which again raises the question about whether all data was manually reviewed to determine fall and non-falls incidents.

8. Furthermore, the motivation for selecting incident reports and image order
entries for subsequent degree of harm and lag analysis is not convincingly justified. The PPV might be high but a large number of falls are being missed by the system. Image order entries detected 10 out of the 80 fall incidents but gets 100% PPV because there are no false positive falls detected. It needs to be made clear why such choices were made and how it aims to address the goal of the paper i.e. addressing under/non-reporting and delayed submission of fall incidents.

Minor Essential Revisions

1. Abstract (paragraph 3): The author describes that the “Sensitivity to detect falls was excellent when using progress notes (100%), incident reports (65.0%) and image order entries (12.5%).” The 65% or 12.5% reported is not, for any metric, an excellent result.

2. Background (paragraph 2): Other NLP systems for detecting inpatient falls should be briefly described and also address explicitly what the current study offers that previous methods didn’t.

3. Methods (Settings; paragraph 2): Mention that reports were in Japanese from the outset.

4. Methods (NLP of free-text...): Decision rules were “whether the text data contained sets of morphemes that were unique to fall-related reports but not to fall-unrelated reports.” This approach seems to be targeted towards a high precision task. However, it seems like that this statement might not be entirely true given that some morphemes may indeed be contained in fall-unrelated reports through the analysis of morpheme frequencies.

5. Methods (NLP of free-text...): Regarding the selection of the top 170 sets of morphemes as rules, why was this threshold chosen and what p-value did this correspond to? The author also mentions “semantics” in the section title, however semantics wasn’t referred to when discussing how the rules were generated. Likewise within the section Results (Category decision rules ...) it was unclear what the “semantic views” were referring to.

6. Method (Various data sources to detect falls): The number of “patient days” was continually used as being the value for “number of patients.” This was initially confusing until the metric of number of falls per patient days was reported later on in the paper. Regardless, to a reader the number of patient days is not the same as number of patients and this should be made clear within the text. I would have like to see both the number of patients and number of patient days to be reported. In addition, the study was performed using reports, and no mentioned of the number of reports for each data source were mentioned.

7. Method (Various data sources to detect falls): The following sentence seemed redundant given that the data sources were mentioned within the chart review sentence: “We also obtained text data...”.

8. Method (Comparison of lag-time...): Explicitly mention what data was used
9. Results (Category decision rules ...): Examples of some of the rules would help understand how they were related to the “four major groups” of rules.

10. Results (Performance of the category decision rules): Ensure that all reported results are correct. There was a total of 2,231 fall-unrelated incident reports (not 2,236), and the specificity should be 97.7% (not 97.5%; i.e. 2179 TN / (52 FP + 2179 TN) where TN is true negative and FP is false positive). Also within the Results (Comparison between incident reports and image order entries...) section, 67 falls detected from the image order entries should be 57 falls (i.e. 15 image order only + 42 overlap from 2 sources).

11. Results (Comparison between incident reports and image order entries...): How was the number of falls detected in this dataset? If by the NLP system, what were its sensitivity, specificity and PPV? It also appears to be the case that the incident reports were used as the baseline where additional falls detected from image order entries would provide the increase in detection rate. Is this baseline assumption appropriate for the study?

12. Results (Comparison between incident reports and image order entries...): How was the degree of harm classified? Was this a manual process or a field within the hospital information system?

It would be good to know the number of incident reports and image order entries that overlap by degree of harm. Such analysis will clarify the results in the following section on Results (Falls with moderate to severe injuries) where the degree of harm classes for the 3 additional falls from image order entries that were not found in incident reports, and also for the 3 incident reports not found by image order entries could be further discussed about. This will prompt discussions around why there is a potential under/non-reporting from each data source.

13. Discussion (paragraph 1): The statement “Although falls with no or mild injuries could not be detected by this method...” needs to be placed in context. If the comparison is done between incident reports and image order entries, then it is unlikely that all incidents from the incident report require an image order. One would assume that only the more serious falls would require an image order. Depending on the type of fall and how these reports play their role in a hospital setting is important when interpreting the results. Likewise for the lag time statement that image orders are significantly shorter; how are the incident reports and image order entries submitted to the hospital information system? Is it because the reports are manually/automatically submitted to the hospital system at regular time intervals? The context is thus very important and the reason why a general overview of the data flows from a fall through to the submission of the report into the hospital information system was advised to be included in the paper.

14. Discussion (paragraph 2): When describing the types of injuries after falls, describe how these injuries affect the data flow i.e. which reports are likely to
contain this information, etc and is the study using such data sources to detect the more serious falls.

15. Discussion (paragraph 3): How was 97% and 34 fold computed? There’s not enough information within the paper to determine this (e.g. knowledge of the total number of reports).

16. Discussion (paragraph 3): The statement that the NLP system “maintaining quality of information on adverse events” is misleading given that you will either have high sensitivity and low PPV, or high PPV and low sensitivity, but not both. This wouldn’t be considered “quality”.

17. References: Reference 4 contains “Bmj” which should be corrected.

Discretionary Revisions

1. The author focused on reports that gave 100% PPV (with potentially low sensitivity), however, progress notes was able to achieve 100% sensitivity. Reports from EMR seem to have complementary information which could be used to the author’s advantage to improve fall detection performance.

2. Developing rules from only one data source and evaluating it on other data sources can be quite problematic as can be seen from the results. Work on using a combination of data sources for development would hopefully improve results across the board. Machine learning has also been shown very promising for such classifications tasks and could be used in conjunction with the rule based system to further improve results.

**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