Author's response to reviews

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

Authors:

Shin-ichi Toyabe (toyabe@med.niigata-u.ac.jp)

Version: 2 Date: 19 September 2012

Author's response to reviews: see over
Editor,
BMC Health Services Research

September 19, 2012

Dear Editor,

Thank you very much for giving us an opportunity to revise our manuscript entitled “Detecting inpatient falls by using natural language processing of electronic medical records”, which was previously submitted to BMC Health Services Research.

We have revised the manuscript in accordance with the comments by the referees and have also provided point-by-point replies to questions from the referees in this letter. All of the corrections in the manuscript have been indicated by underlines.

Sincerely yours,

Shin-ichi Toyabe, MD
Professor, Niigata University Crisis Management Office,
Niigata University Medical and Dental Hospital,
Asahimachi-dori 1-754, Chuo-ku, Niigata 951-8520, Japan
E-mail: toyabe@med.niigata-u.ac.jp
Responses to Reviewer 1

Major Compulsory Revision 1: The explanation of a categorical decision rule may be not enough. Each decision rule seems to be a set of morphemes and it may be decided to be true if a text contains all the morphemes in the set, but it is not clearly explained. Showing some examples of decision rules are helpful for readers.
Response: I appreciate your comments and advice. Each decision rule is not only a set of morphemes but also includes relationship between the morphemes. It is decided to be true when a text contains all of the morphemes in the set and the relationship between the morphemes is the same as the decision rule. According to your advice, I added an explanation of decision rules and some examples of decision rules in Methods.

Major Compulsory Revision 2: Did the author use semantic technologies in decision rules? The author should describe how the semantic technology is used, or remove a representation of "semantic".
Response: Thank you for your expert advice. Certainly, I did not use semantic software technologies. According to your advice, I removed the representation “semantic”.

Major Compulsory Revision 3: In addition, how did the author decide that "170" is the best number of rules?
Response: I apologize that I might have misled you about how I selected the decision rules. I statistically compared proportions of text that satisfy each decision rule between two sets of data from fall-related and fall-unrelated incident reports. As a result, the proportions were significantly different between the two datasets in 170 decision rules. I have corrected the description about the number of decision rules in Methods and Results.

Minor Essential Revision 1: Two types of gold standards seem to be used in this paper. To compute the sensitivity, a gold standard is obtained by the chart review. However, to compute PPV, manual checked data of the possible fall events is used as a gold standard. The author should explain why two deferent types of gold standards are used.
Response: I apologize for the confusion. There is only one gold standard obtained from a manual chart review in this study. I analyzed which type of data source was the most
suitable for detecting falls from two aspects. One is how many falls were recorded in each type of data source, and the other is how effectively the category decision rules can detect true falls when each type of data source is used. The former is calculated by true falls that were recorded in a data source by dividing the total number of true falls found by the chart review. I think that the term “sensitivity” might mislead readers, and I have corrected the term “sensitivity” to “data-containing rate”. The latter was calculated by PPV according to the usual definition. However, I have changed the index of performance from PPV to F-measure. I have corrected the description about gold standards as well as related explanations in Methods.

Minor Essential Revision 2: To evaluate the method proposed, the sensitivity and PPV are mainly employed. The sensitivity is usually shown together with the specificity and PPV is usually shown together with NPV. Why the two measures of the sensitivity and PPV are selected?

Response: As mentioned above, I apologize for the confusion in the term “sensitivity”. Since the term “sensitivity” used in the section differs from usual definition, I have corrected the term “sensitivity” to “data-containing rate”. PPV was selected because it indicates how effectively the decision rules could detect true falls in each type of data source. However, another referee recommends using F-measure instead of PPF.

Discretionary Revision: The sensitivity of the method is very low for texts from image order entries. Isn’t it a problem for practical use?

Response: Certainly, the method using image order entries is not suitable for everyday detection of falls with no or mild injuries because of its low sensitivity. However, the method using image order entries enables prompt detection of injurious falls, which seems to be a major weak point of the incident reporting system.
Responses to Reviewer 2

Major Compulsory Revision 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.

Response: Thank you for your comments and advice. Certainly, the motivation and aims of this study might be not clear for readers of this paper. According to your advice, I added description of the aim and purpose of this study to Abstract and Background sections.

Major Compulsory Revision 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).

Response: I appreciate your thoughtful advice. According to your advice, I provided a general overview from occurrence of events to submission of data to the hospital information system. A description of the overview was added to the Methods section.

Major Compulsory Revision 3: Regarding the collection of fall data from the various sources, little was described on how each report was categorized 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.

Response: Thank you for your detailed checks. Incident reports contained information on categories of incidents and adverse events. Therefore, it is easy to divide incident reports into fall-related and fall-unrelated. Since disagreement between reviewers in a retrospective chart review is known as a significant problem, the manual review was performed by single reviewer who was a physician. I added this information to the Methods section.

Major Compulsory Revision 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.

Response: The categorical decision rules were developed by using a dataset of incident
reports and were evaluated by using another dataset of incident reports. All other data were not used for this evaluation process of the rules.

Major Compulsory Revision 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.

Response: Thank you for your expert advice. According to your advice, I added discussion on comparison to other systems such as Bayes and SVM algorithm to the Discussion section.

Major Compulsory Revision 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.

Response: Thank you for your excellent advice. According to your advice, I calculated F-measure and revised Figure 1 and the Results section.

Major Compulsory Revision 7: Another issue regarding performance metrics is that the abstract and the paper emphasized 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.
Response: The analyses in this study consisted of three stages. The first stage was construction of the decision rules, and the second stage was analysis of which data are the most suitable for text mining. The third stage was comparison of incident reports and image order entries. As you mentioned, a manual chart review was performed in the second stage but not in the third stage, which was stated in the Methods section. I added PPV and F-measure to the first stage analysis as I did in the third stage analysis.

Major Compulsory Revision 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.

Response: I specifically compared the performance of NLP to detect fall events when using incident reports and image order entries as data sources, because performance (F-measure) was excellent when using incident reports and image order entries. Progress notes seem unsuitable for daily use because analysis of a large number of false-positive cases is cumbersome. Discharge summaries contain little information on fall events. I added this information to the Methods and Results sections.

Minor Essential Revision 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.

Response: Exactly. I corrected the sentence.

Minor Essential Revision 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.

Response: I added brief information on other studies on detection of inpatient falls by using the NLP method.
Minor Essential Revision 3: Methods (Settings; paragraph 2): Mention that reports were in Japanese from the outset.
Response: I added description to the Methods section.

Minor Essential Revision 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.
Response: Certainly, the decision rules were sets of morphemes and the relationship between the morphemes that were significantly more frequently observed in fall-related texts than in fall-unrelated texts. I added this information to the Methods section.

Minor Essential Revision 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.
Response: I apologize for the misleading statements. I statistically compared proportions of text that satisfied each decision rule between fall-related reports and fall-unrelated reports. As a result, the proportions were significantly different between the two datasets in 170 rules. As for “semantic”, I did not use semantic software technologies as you mentioned, and I removed the representation “semantic” from my paper.

Minor Essential Revision 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 report 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.

Reply: Thank you for your advice. According to your advice, I reported both the number of patients and number of patient-days. I added the number of incident reports to the Methods section.

Minor Essential Revision 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...”.

Reply: I deleted the sentence.

Minor Essential Revision 8: Method (Comparison of lag-time...): Explicitly mention what data was used when referring to “Distribution of continuous data...”

Reply: I revised the sentence.

Minor Essential Revisions 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.

Reply: Thank you for your advice. According to your advice, I added some examples of decision rules in Methods.

Minor Essential Revisions 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).

Reply: I apologize for my mistakes. I corrected these figures.

Minor Essential Revision 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?

Reply: Since a manual chart review was not performed in this stage of analysis as mentioned above, the true number of falls in this dataset was uncertain. Therefore, the performance index such as F-measure was also uncertain for this dataset.

Minor Essential Revision 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.

Reply: Thank you for your advice. There is a field to input the degree of harm of the event in the incident reporting system, and the entered information on degree of harm is checked and corrected by the safety manager when it is incorrect. As for the degree of harm, I added information on number of events with each degree of harm together with their data sources.

Minor Essential Revision 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.

Reply: Thank you again for your advice about general overview of the data flows. I added a description about it in the Methods section and revised the statement in the Discussion section.

Minor Essential Revision 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.

Reply: Physicians give an urgent x-ray or CT image order when a patient is suspected of having severe injuries after falls such as bone fractures or intracranial hemorrhage. I added this information to the general overview of the data flow.

Minor Essential Revision 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).

Reply: I added more detailed information on the basis to the Discussion section.

Minor Essential Revision 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”.

Reply: I corrected the sentence according to your advice.

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

Reply: Thanks. I corrected the mistake.

Discretionary Revision 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.
Reply: Thank you for your valuable advice. As you mentioned, I also think information on fall events obtained from progress notes has an advantage that it can collect all events. However, low PPV is a barrier for daily use and it seems to be necessary to develop more sophisticated text-mining algorithms.

*Discretionary Revision 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.*

Reply: Thank you for your kind advice. I agree with your opinion. In order to improve performance, it is necessary to incorporate machine-learning technology and cross-validation methods into our method. I think further studies are necessary.