Author's response to reviews

**Title:** Automated Systems to Identify Relevant Documents in Product Risk Management

**Authors:**

Xue Ting Wee (wee_xue_ting@yahoo.com)
Yvonne Koh (Yvonne_KOH@hsa.gov.sg)
Chun Wei Yap (phayapc@nus.edu.sg)

**Version:** 2  **Date:** 21 November 2011

**Author's response to reviews:** see over
Dear Editor,

We are resubmitting a manuscript entitled “Automated Systems to Identify Relevant Documents in Product Risk Management” for consideration for publication in BMC Medical Informatics and Decision Making.

The following is our response to the reviewers’ comments.

Yours Faithfully,

YAP Chun Wei, Ph.D.
Assistant Professor
Reviewer: Martijn Schuemie

Major Compulsory Revisions
1. First and foremost, the description of the field of text-mining as included in the introduction is not at all representative. This is a field with many publications over the years, and the authors seem to have picked a few at random. I suggest referring to some review articles instead (a simple search in Pubmed already found http://www.ncbi.nlm.nih.gov/pubmed/21245076, but many others do exist), and there are several articles that are especially relevant for the work described here, such as http://www.ncbi.nlm.nih.gov/pubmed/19376821, where also PubMed records are classified. Also, the authors mix up the broad field of text-mining with the more narrow sub-field of literature based knowledge discovery (starting with the Swanson paper cited by the authors). The work described in this manuscript falls in the category of text-mining, but has nothing to do with literature based knowledge discovery.

Response: Thank you for your comments. We agree that the introduction is not comprehensive due to the broadness of the field. Thus, our objective for the introduction in text mining is to introduce to the readers the various possible applications of text mining, without going into the specifics of each sub-field. Thank you for highlighting to us some of the works such as PMID: 21245076 and PMID: 19376821. We have added these publications into the introduction.

2. The authors compare their method with a random ranking of the papers retrieved by a query, but this is like comparing a new drug to placebo when alternative therapies are available. The baseline should be a standard information retrieval system, for instance one that ranks documents based on TFIDF in relation to the search query. This could be a standard instance of Lucene, for example.

Response: Thank you for your comment and suggestion to compare our method with Lucene. We have deliberated on this and have decided not to perform such comparison. This is because the main focus of this article is to highlight the potential of text mining for product risk assessment work and not compare among different text mining technologies. Whether our current method is better or worse than Lucene is not important as our results already suggests that text mining methodologies is potentially suitable for product risk management and it is important to consider specific automated systems rather than a general system for all drugs.

Minor Essential Revisions
p4. "such as Therapeutic Goods Administration in Australia, Food and Drug Administration in the United States of America and European Medicines Agency in Europe." should be replaced with "such as the Therapeutic Goods Administration in Australia, the Food and Drug Administration in the United States of America and the European Medicines Agency in Europe."

Response: The sentence has been replaced.
p8, figure 1: please include the 'generalizability set' in figure 1, so figure 1 depicts all data sets used
Response: Figure 1 has been updated to include the generalizability set.

p9. Please explain why synonyms from Omniviz were used instead of more obvious sources such as UMLS
Response: The synonyms from Omniviz were more readily available in our laboratory than UMLS. As mentioned above, the main focus of this work is to demonstrate the potential of text mining methodologies for product risk assessment work.

p19. I don't understand why spelling errors would be a bigger problem in translated abstracts (translators are often very good in English). Have the authors observed these problems, or is it just speculation?
Response: We have not done a thorough study into the spelling error problem. We observed that some of the abstracts contained more spelling errors than others and we postulated that it may be due to the translation of the abstract into English by the authors themselves, who may not be native English speakers. However, since this postulation is based only on limited observations, we have removed this paragraph.

Tables 1 and 3: please specify in the figure caption whether the generic or the specific features were used.
Response: We have added the information to the caption.

Discretionary Revisions
To be honest, I'm surprised that including IDF improves performance. The IDF is a constant value applied to a feature over all documents. Could the authors give an explanation why this increases the ability to distinguish between documents?
Response: We are unable to give a good explanation for this observation.
Reviewer: Courtney Corley

1. The author's present a methodology to automatically classify health product literature into "useful" or "not useful" in determining risk. Two systems are developed, one based on general predictors the other on product specific ones (TNF-alpha). Generally, the hypothesis is understandable; however, the problem statement is not well formed. For example, have what are the specific health risks and how is literature determined to be useful. Granted, the authors do present sufficient motivation in the background section, literature review, and provide a complex algorithm of how usefulness is decided; though the specific problem to be solved is not adequately described in the text. The author's vocabulary is challenging to parse and understand. Product risk management is a broad field that encompasses more than health products, what is "primary literature?"

Response: Thank you for your comments. The problem is that with the large number of articles, it is time-consuming and inefficient for evaluators to manually sieve through this large number of articles and identify those that are valuable to product risk management. Thus, the goal of this study is to identify potential solutions for this. The literature is determined to be useful by evaluators who are working in the field. Primary literature is original research results that are usually published in research journals.

2. There are significant weaknesses and not novel methods in the research. Term normalization (morphological reduction), removing stem words, key-term extraction, term weighting schemes are not novel and should be given minimal treatment in the text, similarly, Figure 3 is not needed - also be careful not to confuse word occurrence with term phrase occurrence. TF-IDF and binary schemes are not strictly frequency measures, they are a measure of term weight. Technically, the terms are not predictors but features to the machine learning algorithm used to build a model. Why do you provide the general terms but not the specific terms? Do you extract general terms from the general data set or both? What about specific terms, do they come from the specific data? If not how do you determine they are relevant to TNF-alpha risk assessment. 730 general terms are reported, however, that seems too low for a corpus of nearly 4200 abstracts. Only the term weights of the predictor terms were chosen as features in building the supervised model. There are many other features that might be useful to this task and could significantly improve performance of both the general and specific systems. Syntactic (e.g., parse trees, part-of-speech), semantic (e.g., UMLS), or other rich features. What is the motivation for using the Gaussian kernel, oftentimes radial kernels produce the best results on text analysis. The explanation of SVM is very good, if you choose to explain this classifier - then describe each of the techniques list in the tables. A short two or three sentence description is sufficient.

Response: We agree that no novel methods were used in this study. Thus, we have already given them minimal treatment in the text. We have removed Figure 3. The term predictors and features are largely interchangeable. In the medical field, predictor is more frequently used and thus we adopt this term to enhance understanding for the readers. The general terms are provided because it is not easily reproducible by the readers as it entails removing specific terms. Specific terms are not provided because these are just terms identified by the software from the abstracts and thus can be easily obtained by the
readers. There are no general datasets or specific datasets. There is only one dataset from which terms are extracted from. The terms are then classified into general terms and specific terms. The terms were not classified by whether they are relevant to TNF-alpha risk assessment. They are classified by whether the terms are ‘specific’ to TNF-alpha or ‘general’ terms. We disagree that 730 general terms for 4200 abstracts are considered to be low. We agree that other features might be useful and we acknowledge this as a limitation of our research. Thus, we have highlighted this in the discussion section and have suggested potential improvements to future study. The Gaussian kernel is frequently used and we have used Gaussian kernel successfully in many of our studies. All the modeling methods used in this study has been thoroughly described in the literature and thus we do not think it is necessary to describe them in this paper, especially since they do not compare as well to SVM.

3. The data collection method is not sound, it is extremely subjective and not clear which data annotation are automated and which manually. How is the usefulness algorithm developed, who developed it, what are the underlying assumptions, has it been validated, is it specific to TNF-alpha? Please report how many annotators reviewed the articles, what was their agreement (e.g., Kappa statistic). The number of abstracts in the specific data-set is an order of magnitude larger than in the general data-set. The variance between them is too large, specific and general is Because of the small “N” of the generalizability data set, I suggest increasing the generalizable data set. Other comments on the method appear in remarks above.

Response: All the data annotation were done manually. We agree that it is a subjective process but all information retrieval are in a sense subjective since it is the end-user who decides whether an article is useful or not. In this study, one of the authors (YK) is an experienced evaluator in product risk management and she came up with the algorithm to determine whether an article is useful or not. She trained the first author (XTW) on the data annotation exercise and helped to resolve any ambiguity that may arise. The reviewer may have gotten confused by the TNF-alpha dataset and the generalizability dataset. These two datasets do not represent the general and specific terms. The generalizability dataset is used to determine the generalizability potential of the general automated system and thus it is smaller in size that the TNF-alpha dataset. Based on our results, it is not necessary to increase the generalizability dataset as our results already show that the general automated system may not work as well as the specific automated system.

4. The manuscript adheres to the relevant standards for reporting and data deposition.
Response: Thank you.

5. The authors should compare the various weighting schemes across each of the supervised methods. That is, TF, IDF, TF-IDF for logistic, k-means, etc. too. It is difficult to assess the system performance in the general automated system. What is meant by the case where the general system was not used? The terminology becomes confusing, do
you use the general terms or specific terms for the generalized system? After reading the article many times, I do not understand what is meant by the first 10% or 20% of articles presented to reviewers. The section "specific automated system" in results, does not make sense, the document selection technique is confusing. Why is the training set size only two for 1:1 useful/non-useful articles? The ROC AUC results are interpretable with such a small "N." What is the justification for such small N? Performance statistics are listed throughout the article for the general system but not the specific system.

**Response:** We disagree on the comparison of various weighting schemes across each supervised methods since this is not the main objective of the study and doing so will only serve to confuse the readers more. The performance of the general automated system is already presented in the usual AUC measure and we have already written clearly on the interpretation of figures 4 and 5. Thus we believe that readers will have no problems with assessing the performance of the general automated system. The case where the general system was not used is the current case where the evaluation is done in an unranked manner by the evaluators. The general system, as its name implies, uses only general terms. The final system will rank the articles and present them to the evaluators in decreasing rank. Thus the first 10% or 20% of articles presented to reviewers are the top 10% or 20% ranked articles by the system. For the development of the specific system, we used different training set size. For a training set size of two, there must be 1 useful article and 1 non-useful article. Otherwise the modeling method will not be able to learn a model. Thus the ratio is 1:1. The AUC is determined from the validation set which has 1190 abstracts and thus N is not small. The performance statistics for the specific system is listed in Table 3.

6. The author's do an OK job presenting the potential limits of the work. However, there are many other limitations, based on this reviewer's previous comments.

**Response:** We have already highlighted the more important limitations of our study. It is impractical to list all possible limitations in a study.

7. The authors clearly acknowledge any work upon which they are building, both published and unpublished.

**Response:** Thank you.

8. The title accurately convey what has been found. However, the abstract is not representative of the work. See comments below.

**Response:** Abstracts are intended to give a broad overview of a study and pertinent results from the study. They are not intended to provide specific details, which are already present in the main article. Thus we have written our abstract concisely and we disagree that it is not representative of the work.

9. The manuscript needs to be edited and reviewed by a professional technical editor to compensate for English not being the first language of the authors.
Response: We agree that some of the sentence structure could be improved and we have attempted to do so. For the information of the reviewer, English is the first language of the authors.

Abstract
10. What is the problem statement/hypothesis, what is the basic premise?
Response: The problem statement is given in the last two sentences of the background section of the abstract.

11. What are health products, old and newer? provide an example of each, and the types of risks this type of system would mitigate.
Response: This information is too lengthy to be provided in an abstract and the intended target audience, drug evaluators, know this very well.

12. the sentence ending "useful for risk assessment work." is awkward and not clear.
Response: We have added in the word product in front of risk assessment work to make it clearer.

13. What is meant by work efficiency?
Response: Work efficiency measures the amount of work that can be done in a certain amount of time.

14. How do you define useful? provide a sentence or two on what is considered useful
Response: The definition of useful has been provided in the main text and will not be appropriate in an abstract.

15. How many specific automated systems were created? The text only describes one, for TNF-alpha. What are the four other drug classes? This is not captured in the text.
Response: Specific automated systems were created only for TNF-alpha. They were not created for the four other drug classes. The four other drug classes were only used as generalizability dataset to determine the generalizability performance of the general automated system.

16. Are evaluator's people or automated systems
Response: Evaluators are people.

17. what is 10% and 20% of articles, does not make sense why this would happen
**Response:** The final system will rank the articles and present them to the evaluators in decreasing rank. Thus the first 10% or 20% of articles presented to reviewers are the top 10% or 20% ranked articles by the system.

18. Provide the performance results for the specific systems too, (not just general).
**Response:** We have added in the AUC for the specific system into the abstract.

19. How was the specific system better? it is a strong statement not backed up by data.
**Response:** We have added the AUC for the specific system. This will make it clearer to the readers that the specific system is better as they can see that its AUC is higher than the general system.

Background

20. What are other types of product risk management, and have similar systems been developed?
**Response:** Different fields have different product risk management. In this study, we focused on health product risk management. No similar system has been developed in this area.

21. How were the 700 abstracts containing valuable information determined? Who determined it?
**Response:** The 700 useful abstracts were determined by manually classifying the 3966 abstracts using the classification algorithm given in appendix 1. Both XTW and YK did the manual classification.

Methods

22. How many annotators reviewed the text and what was the Kappa statistic?
**Response:** The annotations were performed predominantly by XTW and YK, an experienced evaluator, help to remove any ambiguity if present. Thus no Kappa statistics were calculated.

23. Who developed the algorithm, is it reported elsewhere since it is in the appendix? What are the underlying assumptions.
**Response:** The algorithm was developed by YK. It is not reported elsewhere since it is an algorithm used in actual practice and not purely for academic purposes. As mentioned in the main text, this classification is subjective and key advantage of the specific automated systems found in this study is that they are well-suited for such subjective data.
24. were the 693 articles determined to be useful, automatically or by a human annotator?
**Response:** The articles were determined manually by a human annotator.

Text Mining
25. How were the specific and general predictors selected? By human or automatically, is the algorithm similar to choosing useful abstracts?
**Response:** The general predictors were selected manually by a human. A different set of algorithm is used. Basically, if the term is associated with TNF-alpha blockers or diseases treated by TNF-alpha blockers, it is classified as specific. Otherwise, it is classified as general.

Specific automated system
26. The last two sentences do not make sense, need to re-write for clarity TNF-α blockers or diseases treated by TNF-α blockers
**Response:** We have re-written the sentence to make it clearer.

Measuring prediction performance.
27. Need to better describe that AUC is from a ROC curve
**Response:** We have clarify that AUC is from a ROC curve in the text.

Results
28. The description of Figure 5 is not clear, what is the case where general automated system was not used. What is the line graph, confidence, how was confidence calculated,
**Response:** The case where the general system was not used is the current case where the evaluation is done in an unranked manner by the evaluators. The confidence values were calculated by the SVM automatically and is the result of the SVM equation.

Discussion
29. The first paragraph of the "general automated system" is not clear what you mean. the terms general and specific and generalized are confusing.
**Response:** We have used a consistent set of terminology in this paper. Specific terms are terms associated with TNF-alpha blockers or diseases treated by TNF-alpha blockers. General terms are any other terms. Generalizability (generalized was not used in the text at all) is the ability of a model to predict the classification of abstracts which are sufficient different from the those in the training set.

30. PG 17 (first full paragraph). How and why is this done,
**Response:** The specific automated system can be developed by having an evaluator evaluated 20 articles first. The specific automated system will then be trained and used to
automatically classify and rank the remaining articles for the evaluator. This is done so that the system will learn from the evaluator what constitutes useful and non-useful articles and thus will be able to present the appropriate articles to the evaluator.

31. the subsection title "potential application of automated..." is not representative of its content.
   **Response:** We disagree that the title is not representative as the description shows how the general and specific automated system can be used together during routine risk assessment work.

- Minor Essential Revisions
32. PG 20. typo - text is superscript after "[39]"
   **Response:** We have corrected this error.