Author's response to reviews

Title: Studying the Potential Impact of Automated Document Classification on Scheduling a Systematic Review Update

Authors:

Aaron M Cohen (cohenaa@ohsu.edu)
Kyle H Ambert (ambertk@ohsu.edu)
Marian McDonagh (mcdonagh@ohsu.edu)

Version: 2 Date: 2 December 2011

Author's response to reviews: see over
Dear Editors:

Please find attached a revised draft and below our detailed responses to the reviewers comments for BMC manuscript MS:1192997773590152, our submission entitled “Studying the Potential Impact of Automated Document Classification on Scheduling a Systematic Review Update”.

Throughout our responses, the original reviewer comments are shown in **bold font** and our replies as plain text. Where appropriate, we have quoted the modified manuscript text following our responses to the individual comments.

Please extend our thank you to the reviewers for all of the helpful comments and suggestions that have enabled us to greatly improve this paper.

Sincerely,

Aaron M. Cohen, MD MS  
Associate Professor  
Department of Medical Informatics and Clinical Epidemiology  
Oregon Health & Science University  
Portland, OR USA

Comments from reviewer: Lana Yeganova

**…While the work is still preliminary, it raises a number of interesting questions.**

First, it is interesting to know why the algorithm does not perform well on AEDs and ProtonPumpInhibitors topics, as compared to other topics. Second, it would be very interesting to design the classifier targeted at identifying those “motivating” publications. In general, can one define how are these “motivating” publications different from other positive publications included in the SR. What features could one use to identify these “motivating” publications?

We of course agree that these are interesting questions deserving of future investigation. While the F measure performance of the current algorithm does vary quite a bit across topics, the overall correct and incorrect alert rates are still very reasonable and in the same range as the other topics. The number of missed important publications is largest for the ProtonPumpInhibitors topic. This may be due to the fact that this topic has a large number of excluded samples as well as a large proportion of included samples that are deemed important. This highlights the need to develop classifiers specifically for identifying important motivating publications. The annotation schema provides some clues to how this could be done. Identifying
concepts within publications that fit the schema categories and are relatively rare in other topic publications is an approach that we plan to investigate.

**Minor Essential Revisions**

**Abstract does not clearly explain what authors are trying to accomplish in this study. I would suggest expanding the methods section of the abstract.**

We have expanded the methods section of the abstract as follows:

“In this work we study the potential impact of a machine learning-based automated system for providing alerts when new publications become available within an SR topic. Some of these new publications are especially important, as they are more likely to initiate a review update. To this end, we have designed a classification algorithm to identify articles likely to be included in an SR update, along with an annotation scheme designed to identify the most important publications in a topic area. Using an SR database containing over 70,000 articles, we annotated articles from 9 topics that had received an update during the study period. The algorithm was then evaluated in terms of the overall correct and incorrect alert rate for publications meeting the topic inclusion criteria, as well as in terms of its ability to identify important, update-motivating publications in a topic area.”

**On page 5, in the background section, authors talk about “publications that may be more important than others”. I would suggest mentioning that these documents are subset of documents that are eventually included in the update.**

We have clarified this point as suggested: “Furthermore, some publications may be more important than others, in that, in addition to being included in the final SR, they may include specific novel, or higher quality evidence that could motivate the scheduling, priority, or initiation of a review update. We specifically annotate and study these important publications in the work described below. For the work described here, alerts are triggered when any potentially included publication is detected, whether this is a motivating publication or not.”

**It is also not clearly defined whether NewAlert is triggered when a positive document is encountered for update or a “motivating” document. While both of these issues become clear further down the road, they are not clearly defined when introduced.**

We have clarified this by adding the sentence “For the work described here, alerts are triggered when any potentially included publication is detected, whether this is a motivating publication or not.”

**It would be helpful to define your training and test sets in the description of Table 1.**

We have expanded the caption for Table 1 to make this more clear: “**Table 1 - Data sets used in this study.** Eleven systematic review topics had both a prior report and an update completed within our data collection window. Included articles are those included in the final systematic review report, excluded articles are those not included in the report. Two topics, *Antiplatelets* and *NSAIDs* were not included in this study due to insufficient numbers of included and/or excluded articles. DERP review inclusion judgments for articles with MEDLINE entry dates prior to the *End of the Report Cycle* (for the prior report) we used as training data for that topic. DERP review inclusion judgments for articles indexed in MEDLINE during the *Pre-Update*
Period (after the End of the Report Cycle date and prior to the Report Search Begins for the next update) as the testing data for that topic.”

The following sentences need revision.
P9, 3rd sentence describing Table 1; Two topics, … were not included in this study due to insufficient …
Thank you for catching this. We have fixed in this error in the caption for Table 1.

P10, 1st sentence describing Table 2;
We have revised the caption for Table 2 as follows: “Table 2 - Annotation guide for articles that were deemed to potentially motivate a review update on their topic. Each article included in the actual systematic review update was analyzed and assigned either the one most descriptive annotation, or no annotations if the article was not deemed to be potentially motivating for a review update.”

P10, next to the last sentence on the page; Four specific (A, P, B, and L) …
Thank you for pointing this out. We have revised the passage as follows:
“The annotation scheme includes four specific (A, P, B, and L) annotation codes, and one general (M) code intended to cover the different kinds of new evidence that an article might provide. This evidence could motivate the SR expert to consider (e.g., schedule, or try to pursue funding to support) an update of the SR topic.”

P13, 2nd paragraph, last sentence; to be used here to study to
Thank you for catching this. Passage revised to “We therefore targeted a recall of 0.55 to study the performance of the classification system on the important publications in each topic.”

P20, The most important publications are the ones that are most likely to inform
Split and revised sentence: “The most important publications add new information to the evidence base for a given topic. These publications are the ones that are most likely to inform the medical community about new indications or potential harms, and influence the conclusions or recommendations of an evidence report or meta-analysis.”

P20, If a choice needed to be made … have a higher priority than …
Thanks for catching this error. Revised sentence: “If a choice about assigning SR update resources needed to be made, it would be reasonable to assume that ProtonPumpInhibitors would have a higher priority than OveractiveBladder.”

P22, 1st paragraph last sentence; … unstudied adverse events within the literature
Fixed, thank you for catching this. Revised and split sentence: “Furthermore, since novel evidence is an important part of why a particular article may motivate an SR update, it may be useful to specifically recognize new forms of evidence. For example, this would include data such as previously unreported or unstudied adverse events within the literature for an SR topic.”

P22, last sentence on the page; In terms of their potential to motivate
Thanks for identifying this error. Revised and split sentence: “In terms of their potential to motivate a review update, some publications are more important than others, and contribute specific types of new knowledge to the topic evidence base. Therefore we have designed and applied an annotation schema to identify and characterize the publications particularly important in motivating the need for an SR review.”

Comments from reviewer: Andrea Tricco

Major Compulsory Revisions

1. Title: I wonder if the title should be revised to “Studying the impact of automated document classification on when to update a systematic review” – this might help explain what the article is focusing on
We have simplified the title to “Studying the Potential Impact of Automated Document Classification on Scheduling a Systematic Review Update” to help clarity this point.

2. Abstract - Background: The last line of the background states that “relatively little work has been performed studying when and how to update ”. I don’t agree. I think that there are quite a few studies in this area. Instead, may wish to rephrase to something along the lines of “Although work has been performed in this area, further research is warranted because when to update is unclear”. Since this article does not discuss how to update I would not put this in the abstract.
Thank you for this suggestion. Changed to “Although the question of how and when to update SRs has been studied, the best method for determining when to update is still unclear, necessitating further research.

3. Abstract – Methods: The methods seemed scant. May wish to discuss that the investigators annotated articles included in 9 reviews from a database with over 70,000 articles. May wish to discuss the annotation scheme and how the system was trained to annotate the articles.
Thank you for this suggestion. Expanded the abstract methods section to: “In this work we study the potential impact of a machine learning-based automated system for providing alerts when new publications become available within an SR topic. Some of these new publications are especially important, as they are more likely to initiate a review update. To this end, we have designed a classification algorithm to identify articles likely to be included in an SR update, along with an annotation scheme designed to identify the most important publications in a topic area. Using an SR database containing over 70,000 articles, we annotated articles from 9 topics that had received an update during the study period. The algorithm was then evaluated in terms of the overall correct and incorrect alert rate for publications meeting the topic inclusion criteria, as well as in terms of its ability to identify important, update-motivating publications in a topic area.”

4. Abstract – Conclusions – Might want to point out that this is a method that can be used to supplement when to update systematic reviews – the process is not perfect (yet) so you are not suggesting to replace other methods. May also wish to state that this is an “initial test” (as you do in the conclusions from the main
We agree, and have expanded the abstract conclusions as follows: “We performed an initial analysis of the opportunities and challenges in aiding the SR update planning process with an informatics-based machine learning approach. Alerts could be a useful tool in the planning, scheduling, and allocating resources for SR updates, providing an improvement in timeliness and coverage for the large number of medical topics needing SRs. While the performance of this initial method is not perfect, it could be a useful supplement to current approaches to scheduling an SR update. Approaches specifically targeting the types of important publications identified by this work are likely to further improve results.”

5. Methods, fifth paragraph, second sentence: It was unclear as to why February 12, 2010 was chosen. Shouldn’t the investigators have used all updated reviews in the database to test the automated system? Wouldn’t this be more reliable? If this is intended as a pilot test, perhaps this needs to be specified up front in the abstract and maybe even in the title.

This was the date that we took a snapshot of the raw data in the database in order to build and evaluate the predictive models. It was all the data that we had at the time that we began this phase of the work, and included all updated reviews in the database up to that time. Perhaps one point of confusion is that the SYRIAC database is continually updated. We have to take snapshots in order to build models and perform comparisons on a consistent and fixed data set. Also, the SYRIAC database contains “raw” citations and inclusion/exclusion judgments. The predictive models require extensive text and data preprocessing to construct feature sets for machine learning.

6. Methods, eight paragraph: the process of annotating the articles was a bit vague. For example, it was unclear who annotated the publications. All of the authors? Two of the authors? Was this done independently? Did the authors classify all of the 3654 publications or only the 80? How were the 80 originally identified? Further detail here is warranted.

Thank you for identifying this ambiguity. We have tried to clarify this in the text: “The two senior authors (AMC and MM) discussed and modified the article annotation assignments until consensus was reached.” “Only the 332 included publications out of 3654 publications in the test set were considered for annotation, and out of these, only 80 were assigned “important publication” annotations.”

7. Discussion: Might want to point out that this is a method that can be used to supplement when to update systematic reviews – the process is not perfect (yet) so you are not suggesting to replace other methods.

Agreed. Added a similar sentence to the conclusions as was added to the abstract: “While the performance of this initial method is not perfect, it could be a useful supplement to current approaches to scheduling an SR update.”

8. Discussion: May wish to discuss the utility of the machine. For example, who can use this program and what types of skills are necessary to use it? Do you think it will be used for all systematic reviewers performing an update or is this
unrealistic? Is there a chance that investigators won’t trust the results of the machine?
Agreed. Thank you for pointing this out. Added some text to the discussion to help clarify our thinking: “At the current level of performance we expect that our approach will be most useful to the senior reviewer or leader of an SR team. The senior review team lead will be in the best position to effectively combine their domain expertise and other SR topic knowledge with results of our system to best determine when to schedule a review update.”

- Minor Essential Revisions

General:
9. Please use the SR acronym after you spell it out the first time – this was inconsistent throughout the paper
This has been fixed.

10. All acronyms need to be spelled out in the tables
This has been fixed.

Specific:
11. Abstract – Background: Evidence-based Medicine should be “evidence-based medicine”
Corrected, thank you.

12. Background, first paragraph, first sentence: Evidence-based Medicine should be “evidence-based medicine”
Corrected, thank you.

13. Background, second paragraph, sixth sentence: suggest rewording to “Although research guidance on when to update SRs exists [references], this area is still not well understood”
Changed to: “Although there exists research guidance on when and how to update SRs [references], the process is not well understood.”

14. Background, fifth paragraph, sentence beginning “Furthermore, the reviewers can get a better idea of the”: I don’t understand how New Update Alerts can be used to appraise the quality of new information. Perhaps you can clarify?
Clarified the sentence: “The overall impact on the process would be that, by examining the articles that result in alerts, the reviewers would get a better initial idea of the quantity and quality of new information pertinent to an SR before actually scheduling or conducting the review update.”

15. Methods, seventh paragraph, last sentence: “(such as other Ms)” was unclear. Perhaps you can elaborate with an example of what is meant by this?
Expanded to: …information (for example, from additional articles such as other articles that meet the M annotation criteria)
16. Table 4 – unclear why sensitivity and specificity were not calculated
Recall and precision are more standard in text classification work. Recall is equivalent to sensitivity. Precision is equivalent to positive predictive value = TP/(TP+FP). F1 is the harmonic mean of precision and recall. These three measures are usually shown together. For new update alert it is more useful to know the proportion of alerts that are correct (precision), as compared to the proportion of non-relevant publications that did not cause an alert (specificity). However, specificity can be readily computed from the table as TN/(TN + FP).

17. Table 4 – please specify what “F1” stands for
Added definition to table 4.

18. Discussion, last paragraph, last sentence: may wish to explain what “topic drift” means, perhaps by giving an example
Added “(the change in the language or essential concepts within topic discourse over time)”.

- Discretionary Revisions
19. Background, first paragraph, second sentence: should revise to “While this is certainly a desirable goal, given a typical physician’s heavy workload, it can be difficult to realize.” And delete the rest of the sentence
Revised text: “While this is certainly a desirable goal, a typical physician’s heavy workload can make it difficult to realize. Practicing physicians may not have time to consult the primary literature to identify the best-available evidence for each and every patient.”

20. Background, first paragraph, fourth sentence: suggest using the Cochrane definition of a systematic review from the Cochrane Handbook
Added quote of definition from Cochrane Handbook.

21. Background, second paragraph, third sentence: suggest deleting “moving target”, as “continually changing” is sufficient
Removed “moving target” as suggested.

22. Background, third paragraph, first sentence: add an ‘s’ after SR
Added as suggested.

23. Background, third paragraph, sentence beginning with “We seek to study”: add an ‘s’ after the word update
Isn’t “the process of SR update” singular and not plural?

24. Background, third paragraph, sentence beginning with “Here, we define”: suggest making this a new paragraph
Changed as suggested.

25. Background, third paragraph, sentence beginning with “When an article likely to be included”: spell out RSS
Changed as suggested.
26. Methods, first paragraph, first sentence: spell out SYRIAC
   Changed as suggested.

27. Methods, third paragraph, first sentence: suggest deleting “For this study” (it’s already implied)
   Changed as suggested.

28. Methods, eighth paragraph, first sentence: put the word “the” in front of “Pre-Update Period”
   Changed as suggested. Thank you for catching this error.

29. Discussion, fourth paragraph, sentence beginning “Note that this is true for this example even”: put “and colleagues” after “Shojania”
   Changed as suggested.

Comments from reviewer: Byron Wallace

**Major Compulsory Revisions**

The biggest issue with the work as presented is that the task is not clearly defined. The authors first define the 'new update alert', which occurs when the system designates a newly published article as potentially eligible for a review. The project lead will periodically receive 'new update alerts'; but what the authors envision lead doing with this information isn't made explicit. A clearer discussion on how exactly the tool will fit into the existing systematic would make this paper much stronger, especially in light of it describing an applied work. In particular, the primary result is shown in Table 4, and contains precision and recall values for the 'new update alert' classification task. In terms of classification performance from a general perspective, the results aren't that great.

We have expanded the section in the background detailing how new update alerts could be incorporated into the review process:

"New Update Alerts could be useful to the process of SR in several ways. For example, they can be used by the SR team to determine whether an SR needs an update, the urgency of the update, or when it should be scheduled. Seeing potentially includable publications accumulate as they are published may be helpful in scheduling a review update. With a system providing New Update Alerts, reviewers could be made aware of studies potentially impacting the SR scope, conclusions or recommendations at an earlier time. The overall impact on the process would be that, by examining the articles that result in alerts, the reviewers would get a better initial idea of the quantity and quality of new information pertinent to an SR before actually scheduling or conducting the review update. This would provide support for determining when to schedule an SR update, such as whether a review update would be needed as soon as possible, or could a review update be postponed for a time. Given that the resources to conduct SRs are specialized and limited, the ability to coordinate review update scheduling across the full set of a team’s review topics would be a great advantage in best applying those resources and supporting the current needs of the practice of EBM. Furthermore, this could play an important role in obtaining funding to support the review. Since many SRs are dependent upon outside funding, new update
alerts could provide the SR team lead with timely and important information to share with a funding organization.”

While the performance numbers are not as high as we would like, the actual impact on the SR process can best be understood by looking at the alert rates in Table 5. Furthermore, this paper admittedly reports initial work, intended to motivate additional research in this area.

A more enlightening evaluation, in my opinion, would come from a discussion that attempts to quantify the utility of the proposed 'article update alert' system during in terms of time (or money). In other words, the authors should quantify the costs associated with false alarms, discuss the implications of missing relevant studies (false negatives) and contrast these with the existing protocols for updating systematic reviews. At current, the article does not reference existing strategies (e.g., RAND) for updating systematic reviews, but this discussion needs to be contextualized with respect to these current processes for updating systematic reviews. Notably the authors should consult the following articles.

This is a very good idea, but somewhat beyond the scope intended for this initial article. The user-intensive study suggested will require both additional funding as well as IRB approval to use the systematic reviewers as subjects in order to study their time usage and costs. The current article has more modest goals: to propose a task that could be helpful to the SR, and show that it is feasible. Currently the DERP SR team reviews the new literature for past reviews only once or twice a year, leading to a lag of at least 6-12 months for recognizing new evidence and beginning the planning of a SR update.

We have expanded the background to reflect these points:

“Best practice in medicine is continually changing, requiring incorporation of new information as it becomes available, so SRs must undergo periodic updates in order to remain useful and accurate. Updates are costly in terms of both time and money, and can take as much time and effort as the original SR. Typically SR programs, such as the Drug Effectiveness Review Project (DERP) can only review a past SR topic for new literature once or twice a year, leading to a 6-12 month lag in for recognizing new evidence and beginning the planning of an SR update.”

Notably the authors should consult the following articles.


* "When and how to update systematic reviews" by Moher et al. (2008)

The Moher 2008 article is already referenced in the manuscript, as is another Moher article. We have previously referenced other articles by Shekelle, thank you for pointing out this article. We have added discussion of this article and tied it to the context for the current work:

“Although there exists research guidance on when and how to update SRs, the process is not well understood. A comparison by Shekelle of the RAND and Ottawa methods for determining the need for an SR to be updated found that both begin with an initial literature search. Neither..."
method provides guidance on when to conduct the required literature search. The machine learning method proposed here provides exactly this guidance, and fits into the SR update process ahead of review commitment decision methods such as that of RAND and Ottawa.”

The case must be made that the proposed system improves some tangible aspect of the systematic review process compared to such methods. It is not enough to say that 'alerts could be useful'; the authors should argue that they indeed *would* be. Specifically, how does this method contrast to, say, just updating every 2 years, irrespective of any 'signals'? Respectfully, we think that the article provides a sufficient argument for the face and practical validity of introducing a method to inform the SR process that is not dependent upon periodic scheduling of manual steps, such as the literature and expert review processes required by the RAND and Ottawa methods. As stated above, the research described in the current article fits into the process before these other others, and is not in conflict with them. We have added some text to the conclusion to further define how the system fits into the update process:

“The system proposed here could be used continually after a topic is completed, with very little additional manpower required. This would provide a clear indication on which topics need updating before the typical two-year cycle, and which are unlikely to need it. This 1) saves time on the part of reviewers, 2) reduces time delays in updating topics that develop faster, and 3) prevents time and effort spent on reviewing topics not yet in need of an update. The system fits in well with the current RAND and Ottawa approaches, serving as a continuous prior step before the decision is made to allocate the substantial resources required by these approaches.”

With respect to the results, the recall is rather low, even keeping in mind the author's aim to mitigate false alarms. Perhaps the authors could improve this using methods to mitigate the problem of class imbalance (this is not addressed); having worked on this problem myself, I know how much of an issue imbalance can cause. And while I typically dislike when people plug their own work in reviews, but I feel compelled to here, given that our results may be of use. Here at the Tufts EPC, we have been working on semi-automating citation screening for systematic reviews, and I feel that some of our methods may be of use to you. In particular, it seems that you don't explicitly handle the class imbalance problem inherent to the abstract screening scenario; i.e., there are far fewer 'irrelevant' articles than there are 'relevant'. Have you experimented with undersampling the majority class (or other methods to mitigate imbalance)? We have found that 'bagging' classifiers induced over randomly undersampled datasets (i.e., trained on an equal number of 'relevant' and 'irrelevant' citations) performs very well, substantially beating baseline SVM. We have achieved very high specificity and recall using these strategies, given that we are provided a large amount of training data (as you are here).

We have had positive results in prior work with undersampling the majority class as you state. In the present work we have dealt with the class imbalance using a different method. We have found that the undersampling method primarily works by adjusting the margin threshold (‘b’ in the SVM formula) and that, if the appropriate margin cutoff is already known, perhaps because
of prior modeling, then similar results can be achieved more efficiently by applying that cut off directly. That is the case here. Since we know from prior work on the training data the threshold for the chosen recall/precision tradeoff, as stated we use the predetermined margin distance as the cutoff between alerts and non-alerts.

We think that in future work, improving the performance should be focused on smarter features. As noted in the manuscript, the most important papers showed new harms, or results in conflict with prior work. Designing methods to extract features that reflect these situations should improve performance substantially.

An additional question I have is: why was a recall of .55 targeted? The authors mention that the PI preferred this, but it seems unlikely that he or she would 'prefer' such a low recall (why not prefer 100% recall?). More likely, the PI preferred some level of precision that happened to coincide with this level of recall, but based on my experience working on this task myself I suspect you could achieve better results (higher recall with the same precision) if you better handled imbalance. In any case, this explanation should be re-written to emphasize that the PI selected a tolerable *precision* rather than a *recall*.

A recall of 0.55 was targeted because in a user study we found that was the preferred recall point for this task on the actual recall/precision curve for the classification method. Going towards 100% recall leads to too many alerts, towards 100% precision to many publications of interest being missed. Certainly a classification system with higher performance would provide the reviewers different choices. However, note that since SR is a recall driven process (they want all articles, hence the exhaustive literature review), the choice of 0.55 recall can more be properly framed as choosing the lowest acceptable recall for the task, and taking advantage of the relatively higher precision that goes along with that. The context of DERP is important in the choice of 0.55. The team lead would be reviewing multiple topics every month for years – not just a one-off SR now and then.

We have updated the paper accordingly:

“In particular, the principle investigator of the DERP (one of the senior authors of this paper) consistently preferred a recall of 0.55 and the achievable precision corresponding to that level of recall over all other available levels of recall between 0.99 and 0.55. The context of DERP is important in the choice of 0.55. The team lead would be reviewing multiple topics every month for years – not just a one-off SR every now and then. The level of the continual workload is an important factor. This means that the reviewer found 0.55 as the lowest acceptable recall for this task, leading to the highest precision that the current system can deliver at an acceptable recall. We therefore targeted a recall of 0.55 to study the performance of the classification system on the important publications in each topic.”

Finally, I think the annotation scheme the authors introduced is interesting, but more details are necessary. For example, it’s not clear that the inclusion of a 'new patient population for a drug' would necessitate the updating of a review, assuming the new population is outside of the target population defined in the PICO criteria for the original review. Other details should also be included, e.g.
what constitutes a 'significantly' larger sample size? What makes a study 'better designed'? Perhaps these are subjective decisions, but this should be made explicit.

Admittedly, the definition of thresholds such as “significantly larger sample size” are dependent upon the expertise of the reviewer and it would be difficult to create simple objective and complete decisions for these things. We have expanded the description of the annotation scheme to clarify:

“These categories were determined in an iterative manner after discussing the types of new evidence that can contribute to a review update and examining publications from the update period of each topic. Certain aspects of the annotation definitions rely on the SR experience and expertise of the annotator. For example, “significantly larger sample size” must be interpreted by the annotator in the context of all prior studies performed in the given domain.”

Minor Essential Revisions
I think the authors should speak more the discrepancy of results in the training (cross-validation) sets versus the test sets. Is this a case of concept drift?

We mention this in the revised manuscript: “Finally, the recall and precision of the classifier varied much more widely in the test set than in the cross-validation estimates obtained on the training set. We attribute at least some of this variation to the small sample sizes in the test sets. For the largest test set topic, Sedatives, the achieved recall of 0.50 is reasonably close to the target recall of 0.55, and the achieved precision of 0.591 on the test set is reasonably close to that predicted on the training set, 0.516. Further study will be required to determine whether additional effects, such as topic drift\textsuperscript{18,19} (the change in the language or essential concepts within topic discourse over time), are also coming into play here.”

A figure in the Methods section that illustrates the different time periods enumerated (search begins, pre-update period, etc.) would be appreciated.

Figure 1 provides most of this information with dates specific to each topic. To aid the reader, we have added a new table (Table 1 now), defining these time periods.

The authors did not perform any tuning for the c parameter in their linear SVM; they should make a note as to why they didn't (did grid search not boost performance?).

We have added text describing our experience and rationale for this decision. Also, a complete description of the algorithm is included as appendix 1.

“We have done extensive past work on systematic review text classification with SVMLight, using grid search and cross-validation to investigate the impact of adjusting tuning parameters. We have found that there is little impact in adjusting the C tuning parameter with text classification problems involving many more features than training samples, such as is the case here. With the large numbers of features typically used in text classification, training error is minimal, and the size of the C parameter has little impact. The default value of one over the average of the sample norm usually works well. Un-validated grid search on the training data may result in misleading, over-trained results. Additionally, using the default value of C reduces the need for tuning and therefore eliminates the need for additional tuning data or reusing the
training data for tuning, which may bias other aspects of an evaluation, such as feature selection.”

Discretionary revisions and other comments

I would have liked to see the specificity listed in Table 4, though the authors helpfully include the 2x2 tables, so this isn't strictly necessary. Personally, I don't find either the F_1 measure nor the threshold scalar used helpful in this case. With the large class skew, in this problem specificity mostly a reflection of the number of negative samples and is not highly informative. The specificity for the topics are all about 0.90 or above, except for the NasalCorticosteriods (specificity = 0.45), which has many fewer negative samples, and much less class skew than the other topics. As noted, the 2x2 table information allows interested readers to compute this measure, as well as others of potential interest, such as MCC.

Comments from reviewer: Debra Revere

…in this manuscript, incorrect grammar, run-on sentences, and unprofessional writing style distract from the content.

We have extensively reviewed and revised the text to resolve this issue with the language and writing style. See the specific changes noted above, as well as many small revisions throughout the manuscript. In particular we have split run-on sentences to multiple sentences.