A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review

Article *in* AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium · November 2010

Source: PubMed

3 authors, including:

Kyle Ambert
Intel
11 PUBLICATIONS **177** CITATIONS

SEE PROFILE

A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review

Aaron M. Cohen MD MS, Kyle Ambert, and Marian McDonagh PharmD
Department of Medical Informatics and Clinical Epidemiology,
Oregon Health & Science University, Portland, Oregon, USA

Abstract

Systematic reviews (SR) are an important and labor-intensive part of the Evidence-based Medicine process that could benefit from automated literature classification tools. We conducted a *prospective* study of a support vector machine-based classifier for supporting the SR literature triage process. Over 50,000 training data samples were collected for 18 topics prior to March 2008, and used to make predictions on 11,000 test data samples collected during the subsequent two years. Test performance (AUC) was comparable to that estimated by cross-validation on the training set, and ranging from 0.75 - 0.99. Mean AUC macro-averaged across all topics was 0.89, demonstrating that these methods can achieve accurate results in near-real world conditions and are promising tools for deployment to groups conducting SRs.

Introduction

Systematic reviews (SRs) are essential to the practice of Evidence-based Medicine (EBM). The process of conducting a SR involves locating, appraising and synthesizing the best available evidence from clinical studies of diagnosis, treatment, prognosis, or etiology, to provide informative empiric answers to specific medical questions. SRs inform medical recommendations, guiding both practice and policy.[1]

Medicine is continually changing, incorporating new information as it becomes available, so SRs must undergo periodic updates in order to remain useful and accurate. Currently, the process of SR creation or update requires EBM experts to manually review thousands of articles related to a given topic.[2] After 10 years of concerted effort by the EBM community, less than half of the estimated 10,000 needed SRs have been completed. In addition, new clinical trials are currently published at a rate of more than 15,000 per year, making the need for improved efficiency in preparing and updating reviews increasingly urgent.[3]

The process of SR creation and update is highly inefficient, typically requiring 6-12 months of effort with the main expense being personnel time. One possibility for improving the efficiency of this process is using automated text classification and machine learning (ML) techniques to help identify and screen possibly relevant literature citations for a given topic. *Work prioritization* is a promising application of ML for SR. Instead of treating a problem as a binary classification task, where the system predicts which documents are likely to be included or excluded from the final systematic review, work prioritization uses past inclusion decisions to prioritize documents in terms of their likelihood for being judged as necessary to include in an SR. These rankings can be used by reviewers to organize and prioritize their manual review work.

While are several steps in completing a SR, determining article inclusion requires applying predefined criteria to citations of articles identified through comprehensive searches. This takes considerable time and effort, such that prioritizing the set of most likely to be included articles to be managed first could be advantageous for achieving practical time and budget limitations. Identifying the most likely-to-be-included documents first allows human reviewers to obtain and read this set of full-text documents sooner, moving these articles through the review process first, and assigning the review of the less-likely documents a lower priority. In reviews with searches that result in a large number of citations to be screened for retrieval, reviewing the documents according to their likely importance would be particularly useful. The remainder of the citations could be screened over the following months, perhaps by the members of the team with less experience, while the work of reviewing the includable studies is on-going. Lastly, a system incorporating a trained ML algorithm could be set up to monitor newly-published literature, and determine whether it is likely to be included in an update of the review topic. This could serve as an aid to determining when a review topic requires an update.[4]

In this study, we continue our work applying ML techniques to the process of creating and updating systematic reviews. Most published research studying the application of ML techniques to biomedicine (including our own) evaluate performance using either cross-validation on a single data set, or training and test sets collected at the same time (e.g., [5, 6]). To fully understand the implications of using ML in the SR process, it is necessary to go beyond current evaluation methods and study performance in as close to a real world scenario as possible.

To model the use of ML for SR work prioritization in a real world-like scenario, we collected separate data sets for training and evaluation. The evaluation data set represents the samples to which the prediction model built from the training set would

Table 1. Topic sample and included/excluded counts for the systematic review topics on which data was collected.

Topic	Prior Data			New Data			Include in Study?
	Included	Excluded	Total	Included	Excluded	Total	
AceInhibitors	138	5420	5558	8	225	233	YES
ADHD	268	2261	2529	73	512	585	YES
Antiemetics	151	2066	2217	34	609	643	YES
Antihistamines	100	719	819	30	1344	1374	YES
AtypicalAntipsychotics	515	2431	2946	273	1928	2201	YES
BetaAgonistsInhaled	100	4857	4957	1	87	88	YES
BetaBlockers	193	4225	4418	17	730	747	YES
CalciumChannelBlockers	220	2758	2978	7	165	172	YES
Diabetes	37	843	880	3	53	56	YES
DiabetesCombinationDrugs	26	454	480	1	2	3	**NO**
HepatiticC	92	610	702	0	0	0	**NO**
HormoneReplacementTherapy	181	342	523	11	15	26	YES
HyperlipidemiaCombinationDrugs	19	280	299	0	4	4	**NO**
MSDrugs	131	1672	1803	64	639	703	YES
NeuropathicPain	95	400	495	1	5	6	**NO**
NSAIDs	269	102	371	1	7	8	**NO**
Opioids	7	4595	4602	16	862	878	YES
OralHypoglycemics	6	943	949	0	39	39	**NO**
OveractiveBladder	98	774	872	32	277	309	YES
ProtonPumpInhibitors	178	761	939	96	1926	2022	YES
Sedatives	133	1522	1655	30	333	363	YES
Statins	173	6514	6687	8	175	183	YES
Thiazolidinediones	228	2178	2406	7	78	85	YES
Triptans	145	697	842	49	298	347	YES
TOTALS	3411	46814	50225	762	10313	11075	N/A

have been applied, had the system been operational and available to the reviewers in the SR process. The evaluation is *prospective*, in that the data and algorithms the predictive models are based on was collected approximately two years before that used for performance evaluation. The ML models are essentially "blinded" to the evaluation data and any circumstances that affected the collection of that data during the intervening two years, including changes in search strategies or process. The test data used for evaluation is the same that a deployed system would have been applied to during that time period. In this manner, we can be confident that the performance obtained reflects the performance that would have been seen in a deployed system over this time period.

Methods

Data Sets: In order to perform this work, we created two separate data sets based on systematic review inclusion data collected by our automated SYRIAC system, which has been described in a previous publication.[7] The collection contains the titles, abstracts, and MeSH terms for over 60,000 documents that have been judged by experts for inclusion in various systematic drug reviews performed for the Drug Effectiveness Review Project (DERP) by researchers at the Evidence-based Practice Center at Oregon Health & Science University. Each review comprises hundreds to thousands of journal article judgments.

The first data set contains all systematic review inclusion decisions for 24 review topics captured as of March 6, 2008. These data have been used previously for training and cross-validation, and reported on in our prior publications.[5, 8] In this work, we call this collection the *prior data* set.

The second data set was newly collected and constructed for this study. A snapshot of all review inclusion decisions captured by the SYRIAC system as of February 12, 2010 was taken. From this, a set of new samples, never previously used in any of our work, was extracted. These samples were extracted by limiting the new data set to the 24 topics seen in the prior data set, and removing any samples that had identical topic and PubMed identifiers to those in the prior data set. This data collection, which here we call the *new data* set, contains only newly collected samples from topics that appear in the prior data set.

Descriptive statistics on the prior and new data sets are shown in Table 1. As can be seen in the table, over 50,000 samples had been collected over 24 topics in the prior data, with over 11,000 samples of new data collected across these 24 topics. Of the 24 topics, six have been excluded from this study because of a lack of new samples. We used two criteria to make this exclusion. If the new data set contained no samples meeting the inclusion criteria for the systematic review, or if the total number of included and excluded new samples was less than ten, then the topic was excluded from this study because the sample size was too small to achieve accurate evaluation results. We used the remaining eighteen topics to conduct our analysis.

Sample counts shown here are slightly different from our prior published work. Multiple annotation records corresponding to the same publication may occur in the Endnote (www.endnote.com) files used as input to SYRIAC as a result of reviewers searching multiple databases. The inclusion fields for these duplicate annotations sometimes differ. Previously, multiple Endnote records referring to the same publication were resolved to the annotation of the highest numbered record. Discussion with the SR team lead determined that it would more accurately reflect DERP processes to treat a publication with any "Included" annotations as included. This affects at most 5% of the records in a topic.

Classification System: To rank the samples for potential inclusion in each systematic review, we applied the support vector machine (SVM) based classification system that we have reported in our prior work.[5, 8] Briefly, this is an SVM-based machine learning method that ranks samples based on the signed-margin distance. Samples with large positive margin distances are ranked strongly positive for inclusion, and samples with very negative margin distances are ranked as strongly excluded.

Features input to the classifier include uni- and bi-grams from the title and abstract, and MeSH terms associated with the publication. We use the SVMLight implementation of the SVM algorithm (http://svmlight.joachims.org/), with a linear kernel at default settings. Previously, we have done extensive testing with alternate kernels and optimizing parameters for the systematic review classification task, and found minimal performance gains compared to using the linear kernel at default settings.

Evaluation: Since the main goal of this study is to determine how well predictive models created with the prior data set perform on prospectively collected data, we conducted two types of evaluation on the prior and new data sets. Both sets of measurements utilize the area under the receiver operating curve (AUC) for making comparisons.[9] Given a set of documents, where some are positive and some are negative for a given task (here, for inclusion), AUC is a good measure of the quality of a specific document ordering, where 1.0 is a perfect score and 0.50 is equivalent to a random ordering.[10] AUC is independent of class prevalence and is thus a good measure to use when comparing performance across systematic reviews, which typically have different class prevalences.

On the prior data set, for each topic, we performed five repetitions of two-way (5x2) cross-validation with stratification to keep the ratio of positive to negative samples consistent between training and test splits. Each two-way cross-validation was randomly and independently split. The resulting 10 AUC measurements where then averaged together to create a mean AUC measurement for each topic. This procedure results in a baseline set of performance for each topic, and allows comparison with our prior results on this data that used the earlier duplicate annotation resolution method.

To conduct the prospective performance evaluation, we trained our classification system on the prior data set, and evaluated the predictions of this model when applied to the new data set. In other words, the prior data set was used as the training collection and the new data set served as the test collection. AUC per topic was computed for the resulting predictions. We then compared the results obtained from applying 5x2 cross-validation on the prior data set with the results obtained by training on the prior data and testing on the new data. Statistical comparisons of the performance across topics were conducted using the paired *t* and Wilcoxon tests.

Lastly, there has been steady interest in the machine learning and biomedical informatics literature in the idea of topic or conceptual drift[11, 12] where the class definitions or distributions corresponding to a task change over time. If a machine learning process is being used to make predictions for a task undergoing drift, then it is possible that performance will suffer over time as the task diverges from its original definition or purpose. Systematic reviews undergo periodic updates, with the knowledge base of medicine constantly changing, and it is certainly

Table 2. Cross-validation and prospective train/test evaluation on 18 systematic review topics.

Topic	Xval Mean AUC	Test AUC	AUC - Mean AUC
AceInhibitors	0.918	0.893	-0.024
ADHD	0.901	0.898	-0.003
Antiemetics	0.906	0.835	-0.071
Antihistamines	0.848	0.962	0.114
AtypicalAntipsychotics	0.855	0.894	0.039
BetaAgonistsInhaled	0.905	0.851	-0.055
BetaBlockers	0.889	0.890	0.001
CalciumChannelBlockers	0.877	0.825	-0.052
Diabetes	0.983	0.994	0.010
HormoneReplacementTherapy	0.888	0.752	-0.136
MSDrugs	0.906	0.903	-0.003
Opioids	0.829	0.887	0.057
OveractiveBladder	0.871	0.876	0.004
ProtonPumpInhibitors	0.887	0.891	0.004
Sedatives	0.908	0.933	0.025
Statins	0.907	0.932	0.025
Thiazolidinediones	0.894	0.934	0.040
Triptans	0.917	0.830	-0.087

possible that topic drift could affect the ability to apply machine learning to the systematic review process. The two-year period in which the data was collected for this study provided an opportunity to investigate the amount of topic drift in the SR domain, and the effect of this drift on automated classification performance. We compiled data describing the opinion of the reviewers about the amount of change introduced to the systematic review key questions and inclusion criteria during the study period. This data adds insight into the performance changes of the prospective evaluation.

Results

The results of the cross-validation and train/test experiments are shown in Table 2. Cross-validation performance is almost identical to that achieved in our prior work, indicating that the change in the multiple annotation resolution method did not affect the performance of our system. Overall performance is high, and on par with the results we have previously achieved. Cross-validation Mean AUC across all 18 topics varied between 0.829 for *Opioids* and 0.918 for *AceInhibitors*. Performance on the test collection varied between a low of 0.752 for *HormoneReplacementTherapy* and a high of 0.994 for *Diabetes*, with an average of 0.888.

Overall, the cross-validation estimates of performance and the actual prospective levels of performance achieved on each topic are very similar. Decreases in performance seen between the evaluations for each topic are in general very small, (less then 0.05 of a unit of AUC for 13 of the 18 topics, with a mean difference of -0.006). For some topics this difference is larger, with both increases and decreases in performance seen on some topics. The biggest increase in performance was observed in *Antihistamines*, which increased by 0.114 over the prior data set. *HormoneReplacementTherapy* suffered the greatest loss, decreasing 0.136 units. Statistical analysis using the paired *t* and Wilcoxon tests showed no overall significant difference between the cross-validation estimates based on the prior data and the train/test evaluation using the new data (p-values 0.66 and 1.0 respectively).

Table 3 summarizes the changes in topics as provided by SR team members. These reflect changes in scope of the reports, including criteria for patient populations, study designs, and outcomes to be reviewed. Only three topics had major scope changes.

Discussion

Overall the performance of the classification system under nearly real world conditions is very similar to that obtained in our previous cross-validation studies. In general, the AUC measures are high, well over 0.80. Statistically, the cross-validation and prospective evaluations performed identically. This is a very encouraging result, and provides strong confirmation that the systematic review infrastructure could be aided by the addition of an ML-based system to the triage process.

However, there were some notable differences in performance. For one topic, *Antihistamines*, the prospective AUC increased 0.114 to 0.962, taking this from a low scoring topic to the second highest scoring topic. Examining the collection sizes in Table 1 gives some insight into this effect. Originally, the training set size for *Antihistamines* is only 819 with 100 included samples, the smallest set used in this study. Cross-validation reduces the training set by half, as compared to the train/test evaluation using the new data set. This topic exhibits a strong dependency on training set size. Learning is not saturated with the amount of data originally available to the cross-validation. Other topics with similar amounts of data, such as *OveractiveBladder*, saturate earlier, achieving almost the same performance with half the training data in cross-validation as with the whole set. We also saw this effect for these topics in our prior research.[8]

The two topics that had the greatest decrease in performance were *HormoneReplacementTherapy*, and *Triptans*. For *Triptans*, the decrease in performance can be attributed not to a change in scope, but to missing data in the original data set. A

Table 3. Changes in topic scope, key questions, or inclusion criteria during two-year period. Explanations for major changes are: *AtypicalAntipsychotics* - new populations added; *HormoneReplacementTherapy* - focus change from long term to short term treatment; *Statins* - children added to the report.

Topic	None	Minor	Major
AceInhibitors	✓		
ADHD	✓		
Antiemetics	✓		
Antihistamines		✓	
AtypicalAntipsychotics			✓
BetaAgonistsInhaled	✓		
BetaBlockers	✓		
CalciumChannelBlockers	✓		
Diabetes	✓		
HormoneReplacementTherapy			✓
MSDrugs	✓		
Opioids	✓		
OveractiveBladder		✓	
ProtonPumpInhibitors		✓	
Sedatives		✓	
Statins			✓
Thiazolidinediones	✓		
Triptans	✓		

number of retrieved but un-annotated references were discovered by the SR team in the Endnote file during a review update. These were fixed during the period of this study, the result being these corrected annotations were not available for training (as they temporally should have been), and instead were included in the new data evaluation set.

The largest performance decrease occurs with *HormoneReplacementTherapy*. Here, a very significant trial was published during the study period that required redefining the scope of the topic. Originally, the SR was scoped as a comparison of estrogen efficacy. After the Women's Health Initiative report became public, the scope was changed from a focus on long-term use to use in the short term with evaluation of the effect of dose or duration of therapy on outcomes.[13] Clearly, this is a large change or "drift" in scope for the review topic, and the classification performance suffers accordingly, although the topic-specific classification is still better than the non-topic system applied to this topic in our prior work.[8] For the two other topics that underwent major changes, *AtypicalAntipsychotics* and *Statins*, classification performance increased. Therefore, population scoping changes had a smaller effect than a change in treatment orientation. Minor changes in scope had no negative effect.

Overall the cross-validation results accurately predicted the performance of the classification system that we expect to see in when deployed under real world conditions, as shown in our prospective results. Sometimes, because of training set sizes, the performance can actually be better than predicted. For topics with significant changes in focus or breadth, performance may suffer. Often the classification system is robust to changes in scoping, but in some cases it may not be. Fortunately, changes that greatly effect classification performance are uncommon, and the systematic review team is aware when scope changes occur that may lead to this effect. This provides the opportunity for the scope changes to be consciously included in the triage process making use of automated classification. The best way to handle this is an area for future research.

Conclusion

We have demonstrated that a SVM-based ML system can achieve high AUC for work prioritization in SR by studying the performance in a prospective evaluation emulating the samples that a deployed system would actually have available to it for training and prediction. The performance achieved within the setting of the Oregon DERP SR process is consistent across topics, and only rarely diminished by major changes in topic scope. Future work will analyze reviewer expectations on required performance levels, as well as the effect of incremental training for topics with scope changes. Furthermore, since all of the data used in this study comes from a single group working on drug effectiveness reports, it remains for future work to show whether classification systems used by other teams working on different areas of SR for EBM will achieve similar performance.

Acknowledgements

This work was supported by grant 5R01LM009501 from the National Library of Medicine.

References

1. Haynes B. Of studies, syntheses, synopses, summaries, and systems: the" 5S" evolution of information services for evidence-based healthcare decisions. British Medical Journal. 2006;11(6):162.
2. Mulrow C, Cook D. Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions. Philadelphia, PA: The American College of Physicians; 1998.
3. Mallett S, Clarke M. How many Cochrane reviews are needed to cover existing evidence on the effects of health care interventions? ACP J Club. 2003 Jul-Aug;139(1):A11.
4. Moher D, Tsertsvadze A, Tricco AC, Eccles M, Grimshaw J, Sampson M, et al. When and how to update systematic reviews. Cochrane Database Syst Rev. 2008(1):MR000023.
5. Cohen A. Optimizing feature representation for automated systematic review work prioritization. AMIA Annu Symp Proc. 2008:121-5.
6. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc. 2009 Jan-Feb;16(1):25-31.
7. Yang J, Cohen A, McDonagh MS. SYRIAC: The SYstematic Review Information Automated Collection System A Data Warehouse for Facilitating Automated Biomedical Text Classification. AMIA Annu Symp Proc. 2008:825-9.
8. Cohen AM, Ambert K, McDonagh M. Cross-topic Learning for Work Prioritization in Systematic Review Creation and Update. J Am Med Inform Assoc. 2009 Jun 30.
9. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. Palo Alto, CA: HP Labs2004.
10. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 2005 Oct;38(5):404-15.
11. Klinkenberg R, Joachims T. Detecting concept drift with support vector machines. Proceedings of the Seventeenth International Conference on Machine Learning (ICML). 2000:11.
12. Srinivasan P. Adaptive classifiers, topic drifts and GO annotations. AMIA Annu Symp Proc. 2007:681-5.
13. Rossouw J, Anderson G, Prentice R, LaCroix A, Kooperberg C, Stefanick M, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA: the journal of the American Medical Association. 2002;288(3):321.