A Pilot Study Using Machine Learning and Domain Knowledge to Facilitate Comparative Effectiveness Review Updating

Siddhartha R. Dalal, Paul G. Shekelle, Susanne Hempel, Sydne J. Newberry, Aneesa Motala and Kanaka D. Shetty

Med Decis Making 2013 33: 343 originally published online 7 September 2012
DOI: 10.1177/0272989X12457243

A Pilot Study Using Machine Learning and Domain Knowledge to Facilitate Comparative Effectiveness Review Updating

Siddhartha R. Dalal, PhD, Paul G. Shekelle, MD, PhD, Susanne Hempel, PhD, Sydne J. Newberry, PhD, Aneesa Motala, BA, Kanaka D. Shetty, MD, MS

Background. *Comparative effectiveness and systematic reviews require frequent and time-consuming updating. Results of earlier screening should be useful in reducing the effort needed to screen relevant articles.* ***Methods.*** *We collected 16,707 PubMed citation classification decisions from 2 comparative effectiveness reviews: interventions to prevent fractures in low bone density (LBD) and off-label uses of atypical antipsychotic drugs (AAP). We used previously written search strategies to guide extraction of a limited number of explanatory variables pertaining to the intervention, outcome, and study design. We empirically derived statistical models (based on a sparse generalized linear model with convex penalties [GLMnet] and a gradient boosting machine [GBM]) that predicted article relevance. We evaluated model sensitivity, positive predictive value (PPV), and screening workload reductions using 11,003 PubMed citations retrieved for the LBD and AAP updates.* ***Results.*** *GLMnet-based models performed slightly better than GBM-based models. When attempting to maximize sensitivity for all relevant articles, GLMnet-based models achieved high sensitivities (0.99 and 1.0 for AAP and LBD, respectively) while reducing projected screening by 55.4% and 63.2%. The GLMnet-based model yielded sensitivities of 0.921 and 0.905 and PPVs of 0.185 and 0.102 when predicting articles relevant to the AAP and LBD efficacy/effectiveness analyses, respectively (using a threshold of* $P \geq 0.02$*). GLMnet performed better when identifying adverse effect relevant articles for the AAP review (sensitivity = 0.981) than for the LBD review (0.685). The system currently requires MEDLINE-indexed articles.* ***Conclusions.*** *We evaluated statistical classifiers that used previous classification decisions and explanatory variables derived from MEDLINE indexing terms to predict inclusion decisions. This pilot system reduced workload associated with screening 2 simulated comparative effectiveness review updates by more than 50% with minimal loss of relevant articles.* ***Key words:*** *machine learning; comparative effectiveness reviews; text classification* ***(Med Decis Making 2013;33:343–355)***

Received 21 March 2012 from Southern California Evidence-based Practice Center, RAND Corporation, Santa Monica, CA (SRD, PGS, SH, SJN, AM, KDS); and the Greater Los Angeles Veterans Affairs Healthcare System, Los Angeles, CA (PGS). This work was largely produced under Agency for Healthcare Research and Quality Contract No. 290-2007-10062-I. The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report. This work was also funded by an internal grant from the RAND Corporation. The RAND Corporation played no role in the design or conduct of the study. Revision accepted for publication 29 May 2012.

Address correspondence to Siddhartha R. Dalal, PhD, RAND Corporation, 1776 Main Street, Santa Monica, CA 90401; e-mail: sdalal@rand.org.

DOI: 10.1177/0272989X12457243

Clinical providers, researchers, and government agencies use comparative effectiveness and other systematic reviews to determine appropriate clinical practice and research needs. Several experts have suggested that systematic reviews should be updated at least every 2 years to maintain their utility, although this is not common practice.[1–6] One barrier to more timely reviews remains the substantial cost involved in repeating the entire process of literature retrieval and data extraction (which may exceed that needed for the first review).[7–11] Researchers often filter citations in 2 labor-intensive stages: first, researchers scan titles and/or abstracts to exclude obviously irrelevant articles; second, researchers use the full text to exclude additional

studies after reading or screening the full text.[8–10] In many comparative effectiveness reviews, researchers retain articles that are useful for analyses of effectiveness.

To reduce such effort in screening systematic reviews, several studies described technologies aimed at limiting the number of retrieved citations that require initial human review.[12–18] These studies typically extracted large numbers of features (explanatory variables) based on variants of a ''bag-of-words'' approach. In this approach, each document is given a set of explanatory variables representing the presence or frequency of text and Medical Subject Heading (MeSH) indexing terms. Some studies added domain knowledge when classifying features using, for example, United Medical Language System (UMLS).[17] Researchers have also used numerous algorithms to model relevance as a function of these many thousands of potential explanatory variables, including variants of a support vector machine (SVM),[13,14,17,18] a voting perceptron-based classifier,[12] naïve Bayes,[14,17] and algorithms related to boosting.[14,17] In one study on predicting articles relevant to systematic reviews, the authors used a voting perceptron-based classifier to model relevance as a function of these explanatory variables.[12] The study reported work reductions for 11 of the 15 topics while maintaining 95% sensitivity for relevant articles. A later related study used a classifier based on the SVM-light algorithm as well as a bag-of-words feature set to generate predictions prospectively (i.e., the approach used articles included in earlier systematic reviews to classify articles retrieved for updated reviews).[16] The investigators found that predictive performance as measured by area under the receiver operating curve (AUC) was stable in the update when compared to predictions generated for training data.

Another group used an active learning strategy to aid the creation of new systematic reviews.[13,18] Similar to the above study, the model predicts relevance based on explanatory variables derived from a bag-of-words approach. However, they used a process that interactively built a classifier using expert decisions on the most uncertain cases; their underlying hypothesis was that decisions chosen on the most uncertain instances would produce better information for a given cost (reviewer time). They were able to reduce the article screening burden in a simulated de novo review by roughly 50% in 2 of 3 reviews while retaining all relevant articles.

While these earlier machine learning studies using thousands of explanatory variables have met with some success, in other fields, model parsimony has contributed toward improving out-of-sample predictions.[19–21] We hypothesized that the effort MEDLINE researchers put into indexing key concepts with subheadings and the domain knowledge within earlier comparative effectiveness reviews can be leveraged for generating parsimonious models with substantial predictive power. MeSH indexing identifies key concepts in articles and further describes those concepts with descriptive subheadings (''chemically induced,'' ''adverse effects,'' ''epidemiology,'' etc.). We hypothesized that extracting data solely on a few key variables (including publication type, intervention, and outcome) could have sufficient power, counterbalancing the slightly greater upfront time required for identifying key concepts beforehand. Earlier, we described using this approach to extract articles that tested whether particular drugs caused any type of adverse effect (AE).[22] In this study, we modified our earlier approach to make it useful for locating studies relevant to comparative effectiveness reviews, which analyze all articles assessing either effectiveness or AEs that use particular study designs.

We tested this hypothesis in a pilot study that used 2 comparative effectiveness reviews conducted by the Southern California Evidence-based Practice Center (SCEPC), under contract to the Agency for Healthcare Research and Quality (AHRQ). The first concerned the prevention of fractures in patients with osteopenia or osteoporosis (low bone density [LBD]).[9,10] The second review covered off-label indications for atypical antipsychotic drugs (AAP).[8] For both reviews, we aimed to use the earlier study's reviewer decisions to predict whether updated search results would have been judged irrelevant by the second-stage filter or would have been classified as relevant for either the effectiveness or AE analyses.

The key challenge in this study is that the training (original search) and test (updated search) data are independent samples. In both reviews, the conditions and interventions of interest, research personnel, and study objectives changed between the initial and update reports. In other contexts, the general problem of training data becoming inapplicable to test data over time has been described as concept drift.[23–27] Researchers have devised several strategies to address this problem including giving weight to more recent training observations or to the small number of classified update observations (if present). An earlier study explored how well an SVM-based machine learning framework would perform on a series of simulated review updates (some of whose

search criteria changed significantly) without additional accommodations to the update.[16] They found that their framework performed well in many cases, although they noted that it was uncertain whether reviewers would be satisfied with their prospective performance in all cases. In this case, the LBD and AAP researchers changed their objectives substantially when conducting the updates; also, they had not classified additional citations from either update. To address this substantial concept drift without additional data, we represented specific drugs and outcomes as more abstract concepts such as ''intervention'' and ''outcome.'' Our underlying hypothesis was that more general attributes would be stable over time, although specific features might change. In the remainder of the article, we describe this method in detail and simulate how such a system might perform in predicting articles that would have passed the second-stage screening process and been included in the update report for either AEs or efficacy/effectiveness.

METHODS

Data Sources

We obtained PubMed citations retrieved by the SCEPC (until January 2011) for the LBD and AAP reviews. We excluded PubMed citations that had not yet been assigned MeSH and publication type terms as well as articles obtained exclusively from non-PubMed databases (such as PsycInfo and EMBASE). Excluding non-PubMed databases is a limitation whose importance varied by study. For the LBD update, all relevant studies were found in PubMed. For the AAP update, 31 articles included in the final report were not located in PubMed. Of those, 14 were scientific information packets (which will always require human review), 9 were identified by mining references of included reports, and 8 were found in poster presentations.

The methodologies for both LBD and AAP reviews have been discussed extensively in other reports.[8–10] Briefly, in the LBD study, the interventions consisted of exercise therapy and multiple drugs (including bisphosphonate drugs, calcitonin, selective estrogen receptor modulators, parathyroid hormone derivatives, and menopausal hormone therapy). The primary outcomes of interest were fractures and AEs, but the search strategy also retrieved articles discussing predisposing conditions such as osteoporosis and osteopenia to improve yield. The initial search (1966-2006) yielded 14,700 articles with full MEDLINE citations, and the updated search (containing a modified set of interventions and new AEs) retrieved 7051 articles with full MEDLINE citations (spanning 2006-2010). In the original AAP review, the interventions consisted of atypical antipsychotic drugs, including olanzapine, risperidone, quetiapine, and clozapine. Outcomes of interest included dementia, obsessive-compulsive disorder, and posttraumatic stress disorder, and outcomes could be excluded if reclassified by the US Food and Drug Administration as an approved indication. The search conducted in 2006 yielded 1307 MEDLINE citations requiring human classification. The updated search added outcomes such as anorexia nervosa, bulimia, and substance abuse to the list of off-label uses under consideration; 3591 MEDLINE citations were retrieved. For both studies, articles retrieved for update and original searches were mutually exclusive.

During the initial modeling phase, we had access to second-stage researcher decisions on the original search results (relevant for AE analysis, effectiveness analysis, or both). The LBD training document literature included 382 relevant articles: 218 for effectiveness and 279 articles for AEs. The LBD update body of literature included 127 relevant articles: 63 for effectiveness and 92 for AEs. The AAP training literature contained 98 relevant articles: 82 for effectiveness and 91 for AEs. The AAP update contained 116 relevant articles: 101 for effectiveness and 105 for AEs. Of note, at the second stage, articles were excluded for more critical reasons (e.g., inappropriate study design, inappropriate intervention, etc.) and reasons of timing (duplicate data, inclusion in prior meta-analysis). We considered the latter articles to have passed a second stage of review because their study design was fundamentally sound.

Processing MEDLINE Citations

We aimed to construct a limited set of important variables from MEDLINE citations, which usually contain 10 to 15 indexing terms that may be modified by 1 or more subheadings. Figure 1 shows a stylized example adapted from the LBD search strategy and 1 citation.[9,28] Data related to generic study characteristics (such as randomized controlled trial [RCT] or human study) were extracted from all studies. We first identified key MeSH terms by programmatically matching all terms in previously constructed search strategies to terms within the MeSH database.[29] One author divided each search strategy into terms related to interventions and terms related to

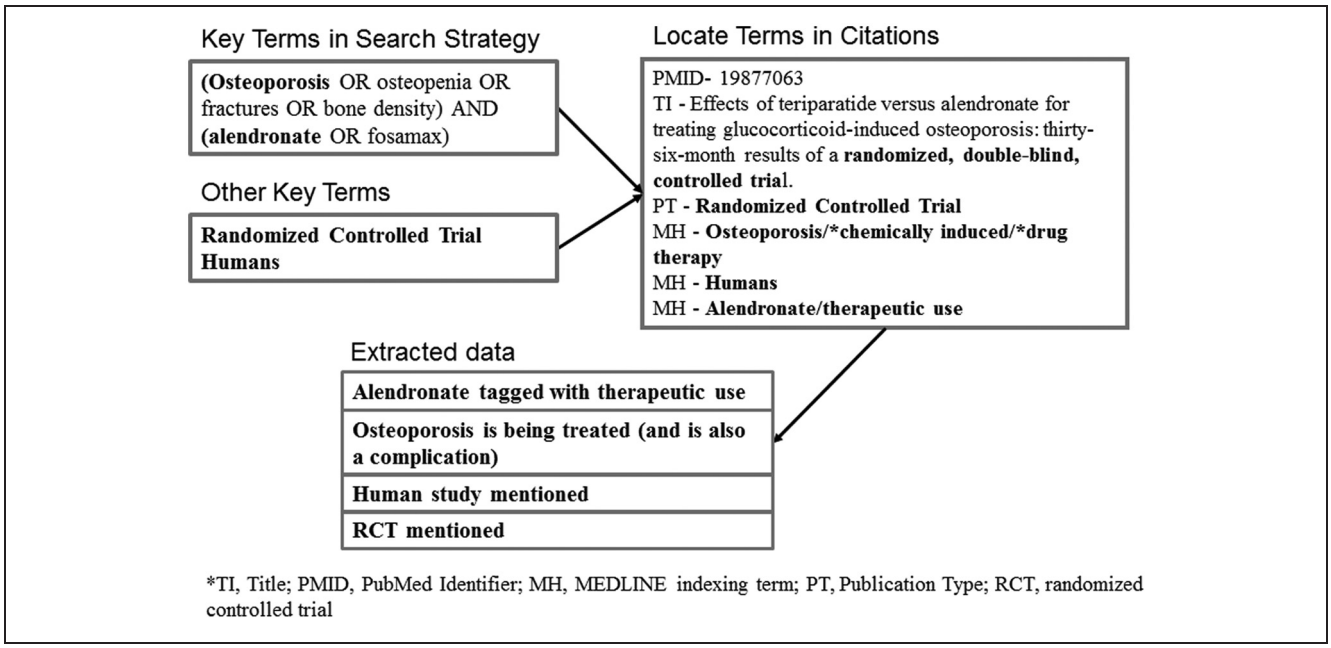

Figure 1 MEDLINE citation processing example.

outcomes. This task was made easier by the fact that outcome and intervention terms were usually grouped within each search. For example, one search in the original LBD review was ''(osteoporosis OR osteopenia OR osteopaenia OR fracture* OR bone mineral OR fractures[mh] OR bone density) AND (raloxifene* OR evista OR tamoxifen* OR nolvadex OR emblon OR fentamox OR soltamox OR tamofen).''[9] Hence, we could identify interventions (raloxifene, etc.) and outcomes (osteoporosis, osteopenia, etc.) without substantial effort. We then created a set of 46 binary explanatory variables based on whether the exact intervention or outcome terms were present in the MEDLINE citation and linked to particular subheadings. We further created a set of 46 matching explanatory variables based on whether other interventions or outcomes (that are unrelated to the outcomes and interventions of interest) were present in the MEDLINE citation and linked to particular subheadings. This might indicate that other diseases or interventions were of primary importance in the article.

In addition, we created a set of 29 binary explanatory variables related to general article characteristics including demographic group (gender and age), treatment target (human, animal, and others), and publication type (review, RCT, meta-analysis, and others). Finally, we created variables indicating whether any intervention or outcome was explicitly mentioned in the article's title or indexing terms, whether the article was particularly short (1 or 2 pages long), and whether ''randomized controlled trial'' or ''meta-analysis'' was mentioned in the title or abstract. In total, we used only these 121 variables instead of a text-based approach that might create thousands of explanatory variables along with a greater risk for overfitting and possible loss of out-of-sample predictive power.

Statistical Classification

We created a series of models based on all combinations of the following: 2 outcomes (inclusion in the final report for efficacy or AEs), training data with associated explanatory variables from 2 reviews (AAP and LBD), and 2 statistical learning algorithms (gradient boosting machine [GBM] and generalized linear model with convex penalties [GLMnet]). After deriving each model empirically using training data from the original review, we generated predictions for articles in each corresponding update. We also evaluated how well GBM and GLMnet could predict inclusion for any outcome (efficacy/effectiveness or AEs) in a given review update (AAP or LBD). Each step is explained in detail below.

We modeled relevance as a function of the explanatory variables discussed above while solely using those articles retrieved in the original search

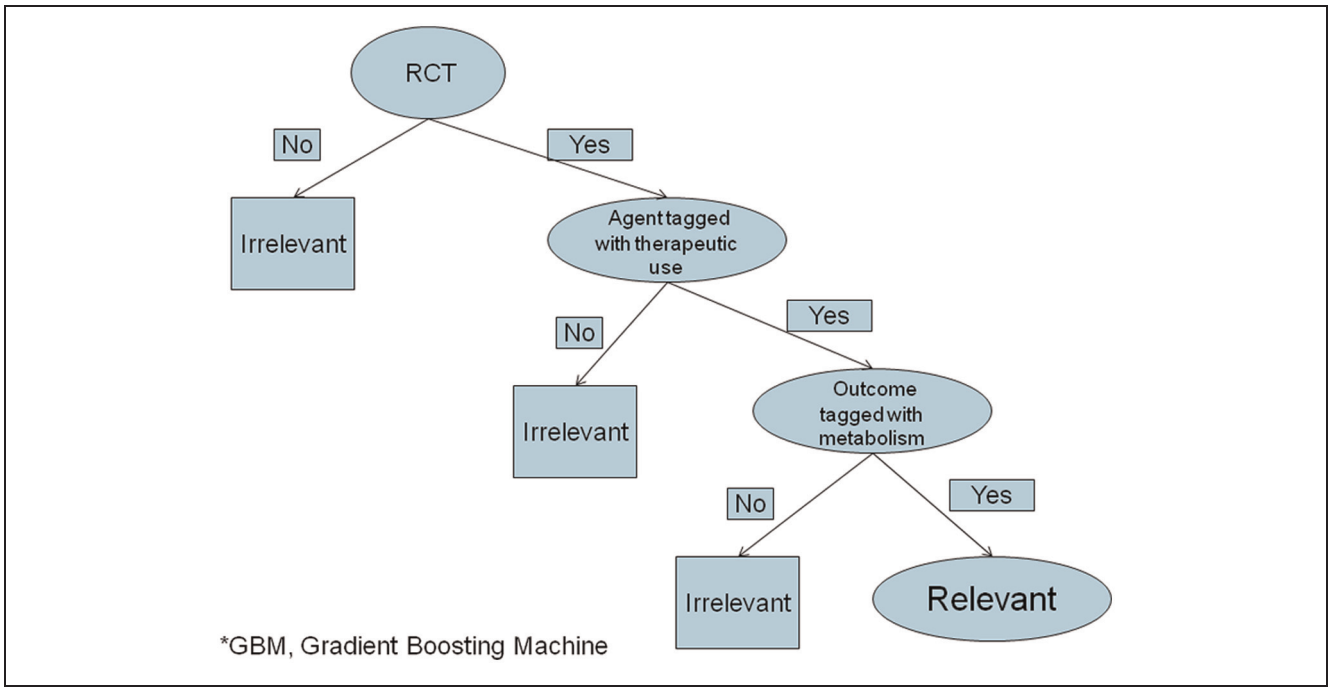

Figure 2 Example of gradient boosting machine tree.

(1966-2005/6 literature: the training data). To simulate a true update in which the update search results would not have been known, we blinded the statistical learning model to researcher decisions from study updates. We constructed separate models for predicting inclusion in effectiveness or AE analyses because article characteristics predictive of relevance likely differed substantially between the 2 analyses. Then, we generated a final prediction using the maximum prediction from the effectiveness and AE analyses. This analysis is most relevant to current AHRQ practice for comparative effectiveness reviews because automated processes will need to capture both effectiveness and AE analyses. However, we also evaluated our models' performance in predicting inclusion in either analysis because other researchers may be interested in 1 type of study.

We first evaluated the GBM statistical method, a nonparametric tree-based prediction approach based on boosting.[30,31] In the general boosting framework, models combine weak (i.e., moderately inaccurate) classifiers iteratively into a strong final classifier. GBM is a specific implementation of boosting and consists of a general, automated, data-adaptive modeling algorithm that can estimate the nonlinear relationship between a variable of interest and a large number of covariates using a sequence of simple classifiers combined in an optimal way. The algorithm generates a large sequence of simple classification trees. Each tree is fit to the prediction residuals for the preceding tree (i.e., the deviations between the observed and predicted values) (see Figure 2 for an example tree). The GBM algorithm also assigns each tree a weight computed to minimize the entire model's overall loss function (in this case, based on the logistic function). The final model therefore includes all simple (weak) models, but each is weighted proportionally to accuracy. We validated the results on training data using 5-fold cross-validation (which reduces overfitting). Each fold of cross-validation randomly selects 20% of the data to serve as test data; then, the process fits a model on the remaining 80% of the data; finally, model performance is measured on the reserved test data. One ultimately finds the model, which would minimize the prediction error averaged across all 5 folds and models.[19]

We also tested GLMnet, which fits a linear logistic model with convex penalty on the magnitude of coefficients.[19,32] A standard linear model would model relevance as a function of all explanatory variables, but this may lead to overfitting. The LASSO shrinkage and selection method for linear regression minimizes the usual sum of squared errors, with

a bound on the sum of the absolute values of the coefficients.[19] The GLMnet method shrinks coefficients of less important variables to zero with a more general convex penalty, resulting in fewer independent variables that often have better out-of-sample predictive power. GLMnet also employs cyclical coordinate descent (computed along a regularization path) to efficiently solve these problems.[32] Of note, we considered using SVM (and other methods), but we chose GLMnet and GBM based on our prior experience and because several reports suggested that it would be unlikely that SVM would be markedly superior to GBM or GLMnet.[14,33,34]

We used both GLMnet and GBM to generate prediction scores (probabilities) for articles in updated searches (the test data). We compared these predictions against ''gold-standard'' independent decisions from the SCEPC at various thresholds; that is, if a relevant article was assigned a probability of 0.015, it would be counted as a true positive at threshold $P \geq 0.01$ and a false negative at a threshold of $P \geq 0.05$. We then calculated the following performance metrics at various thresholds: recall/sensitivity (percentage of relevant articles retrieved), precision/positive predictive value (PPV: percentage of predicted relevant articles that were truly relevant), and simulated workload saved (i.e., percentage of simulated literature search screening reduced by using the predictive model exclusively).

Of note, there is no perfect threshold because neither error minimization nor sensitivity maximization are absolute goals; a strategy that rejected all articles might have an error rate of 1% (and an excessive false-negative rate), while a strategy accepting all articles would have 100% sensitivity (and an excessive false-positive rate).[12,18] To balance these objectives and conform to researcher preferences, we evaluated performance at multiple probability thresholds. We derived sensitivity-maximizing thresholds based on actual performance in the original AAP and LBD studies; we also judged results against a threshold of $P \geq 0.02$, which appeared to have preserved good sensitivity while substantially reducing the error rate. We also compared the performance of GLMnet and GBM using their receiver operating characteristic (ROC) curves and via a nonparametric approach.[35]

To estimate model variability, we calculated bootstrapped standard errors for the sensitivity and PPV results.[36] We derived models from 100 samples drawn with replacement from the original literature review articles; we used these bootstrapped models to generate independent predictions on the actual original and updated reports. We calculated standard errors from the resulting simulated sensitivity and PPV estimates. At plausible thresholds discussed in the report ($P \leq 0.1$), the standard errors were extremely small (probably due to the large sample sizes of the training data used to fit the original models) and are not shown. For example, at a threshold of 0.1, the estimated sensitivity for LBD efficacy articles was 0.995, and the standard error was 0.0008. MEDLINE citations were retrieved in plain text format and parsed using Python 2.7.2 (Python Software Foundation, http://python.org); all statistical modeling was conducted in R 2.10 (R Foundation, http://www.r-project.org/).

This work was produced under contract for AHRQ, which evaluated the work using its own peer review system but did not participate in data acquisition, statistical analysis, or article preparation. This work was also funded in part by the RAND Corporation. The RAND Corporation played no role in the design or conduct of the study.

RESULTS

Literature Characteristics

Table 1 shows the characteristics of the AAP and LBD literature searches, divided into original and updated searches. Substantial and statistically significant differences were observed between the means of variables in the original and updated searches for both LBD and AAP. This finding suggests that the composition of the search results differed substantially between the update and original searches in both studies, making modeling more difficult. Table 1 demonstrates how characteristics may vary between different review topics (e.g., ''RCT'' and ''Agent and toxicity'').

Performance Predicting Any Relevant Result and Potential Workload Reductions

Sensitivity decreases and the number needed to screen increases as the probability threshold is raised for AAP (Table 2). We selected a threshold of $P \geq 0.01$ based on the performance of the model in the original search results, in which a threshold of $P \geq 0.01$ yielded perfect sensitivity with 58.1% of screening saved. When we applied this threshold to the GLMnet predictions for the update, sensitivity exceeded 0.99, and the proportion of title/abstract screening saved was 55.4% or 1990 of 3591 articles.

Table 1 AAP Characteristics: Original versus Update

Variable	AAP			LBD		
	Original	Update	*P* Value for Comparison[a]	Original	Update	*P* Value for Comparison[a]
No. of studies	1307	3591		14,700	7051	
Year, mean (range)	2000.9 (1972-2006)	2005.7 (1988-2011)	<0.001	1997.6 (1966-2009)	2007.5 (1997-2011)	<0.001
Any outcome in title	0.331	0.388	<0.001	0.441	0.345	<0.001
Any agent in title	0.683	0.551	<0.001	0.393	0.507	<0.001
Agent and administration	0.194	0.163	0.011	0.093	0.173	<0.001
Agent and therapeutic use	0.717	0.65	<0.001	0.265	0.272	0.318
Agent and toxicity	0.445	0.381	<0.001	0.078	0.148	<0.001
Demographic tags include child	0.178	0.229	<0.001	0.173	0.14	<0.001
Outcome and complications	0.08	0.084	0.681	0.081	0.077	0.363
Outcome and drug therapy	0.415	0.358	<0.001	0.199	0.177	<0.001
Outcome and prevention	0.004	0.011	0.024	0.177	0.18	0.691
Outcome and psychology	0.222	0.18	0.001	0.005	0.004	0.743
Other outcome and psychology	0.233	0.183	<0.001	0.01	0.011	0.467
Clinical trial	0.287	0.126	<0.001	0.136	0.039	<0.001
Comparative study	0.198	0.169	0.02	0.116	0.077	<0.001
Meta-analysis	0.018	0.023	0.377	0.006	0.017	<0.001
RCT	0.164	0.14	0.035	0.105	0.101	0.366
Text contains RCT	0.102	0.115	0.2	0.061	0.087	<0.001

Note: Each column (original and update) represents both excluded and relevant studies. AAP = atypical antipsychotic drug systematic review; LBD = low bone density systematic review; RCT = randomized controlled trial.
a. *P* value derived from the Fisher exact test comparing update versus original for each study.

Table 2 Performance in Retrieving Any Relevant Article (AAP Update)

Prediction Threshold	GLMnet				GBM			
	False Negatives	Sensitivity	Total Screening Burden	Screening Saved (%)	False Negatives	Sensitivity	Total Screening Burden	Screening Saved (%)
0	0	1	3591	0	0	1	3591	0
0.001	0	1	3237	9.9	0	1	3591	0
0.005	1	0.991	2191	39	0	1	3591	0
0.01	1	0.991	1601	55.4	4	0.966	1807	49.7
0.015	2	0.983	1312	63.5	11	0.905	1021	71.6
0.02	3	0.974	1144	68.1	13	0.888	857	76.1
0.025	4	0.966	1026	71.4	13	0.888	780	78.3
0.05	10	0.914	737	79.5	14	0.879	604	83.2
0.1	14	0.879	549	84.7	18	0.845	471	86.9
0.2	21	0.819	452	87.4	25	0.784	341	90.5

Note: Sensitivity for a particular threshold was determined by selecting articles if the maximum predicted relevance from either model (effectiveness or adverse effect) exceeded the threshold. We do not show sensitivities <0.75 as these results are unlikely to be useful to systematic review researchers. AAP = atypical antipsychotic drug systematic review; GLMnet = generalized linear model with convex penalties; GBM = gradient boosting machine.

The GLMnet-based model for LBD performed worse than in AAP: the model selected articles for the update with a sensitivity of 0.795 at a threshold of $P \geq 0.02$ (compared to 0.974 for AAP) (Tables 2 and 3). However, this approach still provided potential benefits at less stringent thresholds. We chose a threshold of $P \geq 0.001$ because it was the largest threshold that yielded perfect sensitivity in the original search results; it also yielded a 66.8% screening reduction. After applying the same $P \geq 0.001$ threshold to the update, we found perfect sensitivity and a 63.2% reduction in projected article screening

Table 3 Performance in Retrieving Any Relevant Article (LBD Update)

	GLMnet				GBM			
Prediction Threshold	False Negatives	Sensitivity	Total Screening Burden	Screening Saved (%)	False Negatives	Sensitivity	Total Screening Burden	Screening Saved (%)
0	0	1	7051	0	0	1	7051	0
0.001	0	1	2597	63.2	3	0.976	2354	66.6
0.005	10	0.921	1180	83.3	24	0.811	878	87.5
0.01	20	0.843	882	87.5	30	0.764	765	89.2
0.015	25	0.803	749	89.4	31	0.756	687	90.3
0.02	26	0.795	678	90.4	31	0.756	630	91.1

Note: Sensitivity for a particular threshold was determined by selecting articles if the maximum predicted relevance from either model (effectiveness or adverse effect) exceeded the threshold. We do not show sensitivities <0.75 as these results are unlikely to be useful to systematic review researchers. LBD = low bone density systematic review; GLMnet = generalized linear model with convex penalties; GBM = gradient boosting machine.

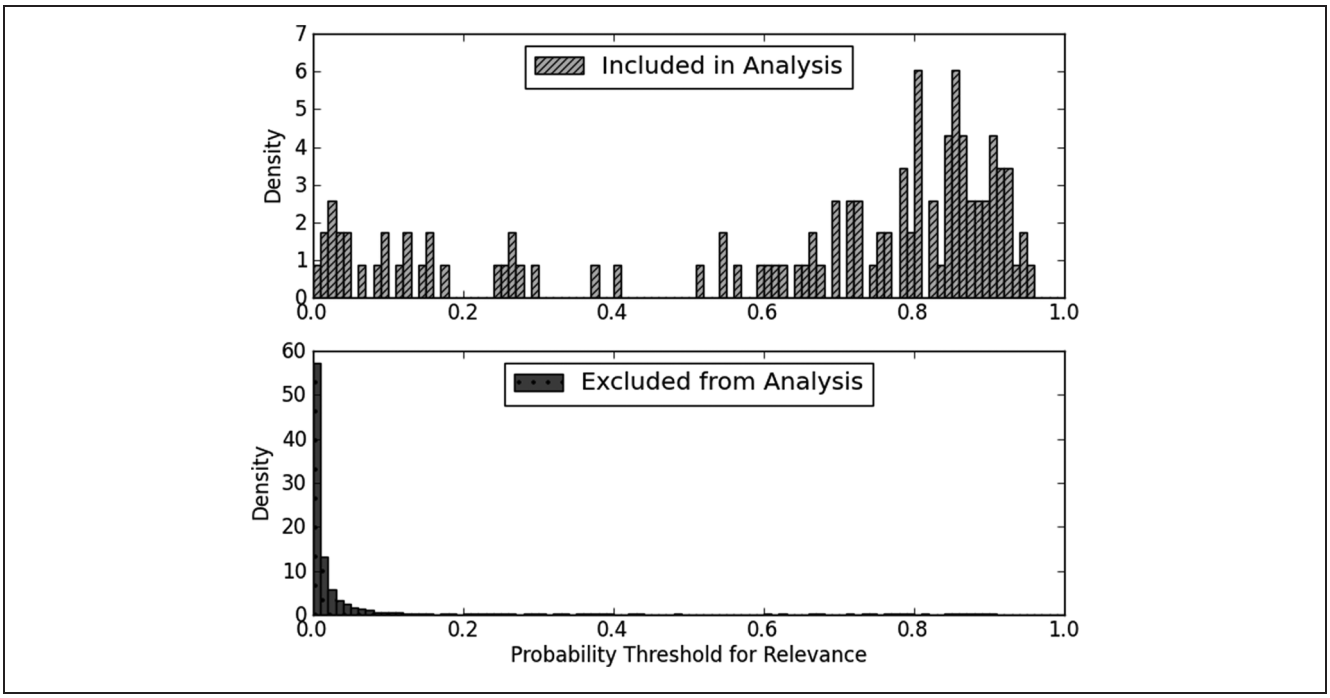

Figure 3 Distribution of predictions for inclusion in any analysis (atypical antipsychotic drug study).

burden. GBM performed slightly worse in both the AAP and LBD reviews, as supported by the ROC analysis below.

We show model prediction performance for the AAP and LBD updated searches graphically using histograms of prediction probabilities (Figures 3 and 4, respectively). Excluded articles were generally assigned very low probabilities in both. However, in the LBD study, a substantial percentage of relevant studies were assigned relatively low probabilities. We also show these results using ROC curves (Figures 5 and 6). The AUC for the GLMnet method (in the AAP study) was 0.943 versus 0.925 with GBM (P = 0.007 under null hypothesis of equality). The AUC for the GLMnet method (in the LBD study) was 0.954 versus 0.947 for GBM (P = 0.06 under null hypothesis of equality). GLMnet seemed to slightly outperform GBM using visual inspection and using AUC methods. Still, it would be difficult to establish GLMnet's superiority in this context (comparative effectiveness review updating) without further studies.

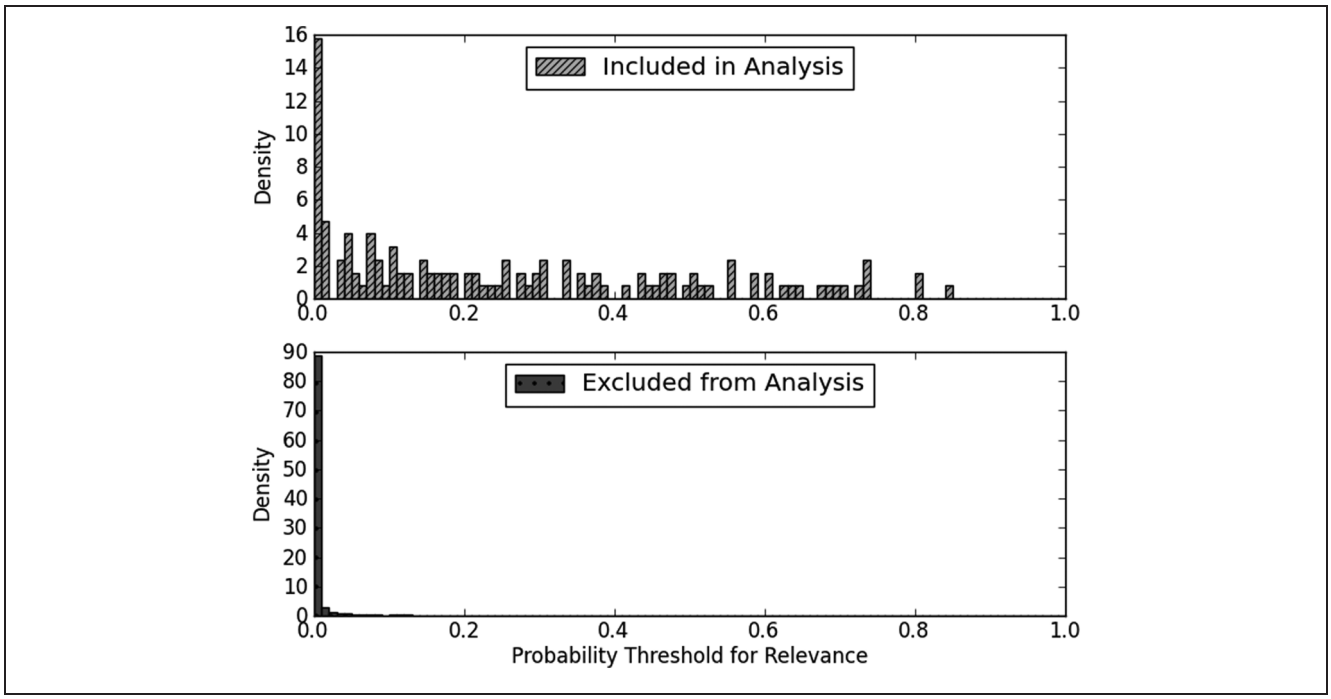

Figure 4 Distribution of predictions for inclusion in any analysis (low bone density study).

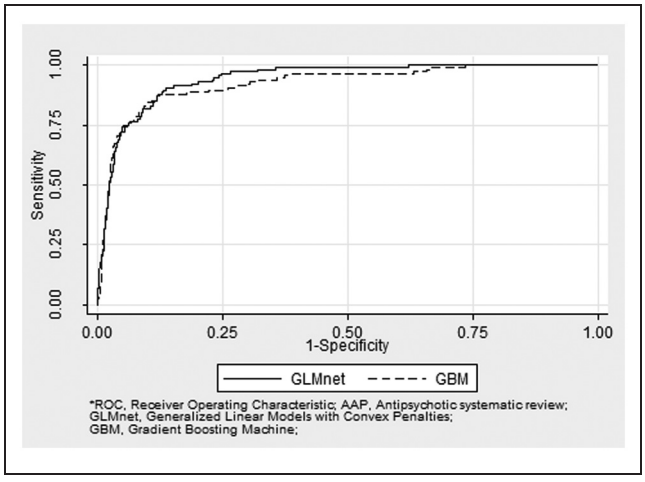

Figure 5 Receiver operating characteristic curve for classifying atypical antipsychotic drug articles.

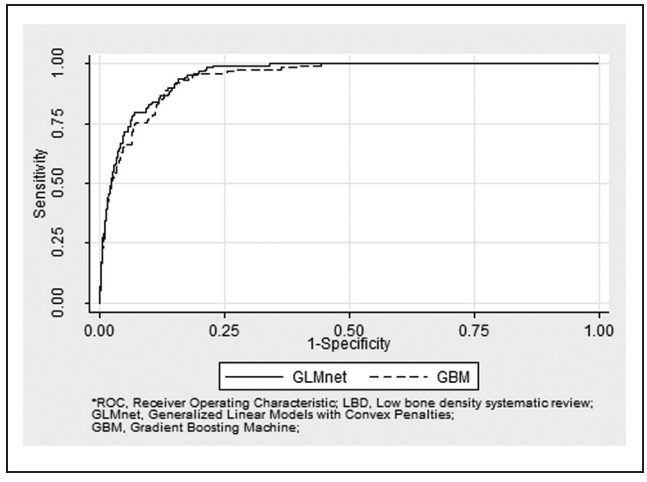

Figure 6 Receiver operating characteristic curve for classifying low bone density articles.

Performance Predicting Articles Relevant to Specific Analyses

We show disaggregated effectiveness (Table 4) and AE (Table 5) results for both studies (AAP/LBD) and methods (GLMnet/GBM). For AAP, all analyses achieved high sensitivity when predicting effectiveness articles on the original sample at relatively stringent thresholds ($P \geq 0.02$). For example, the GLMnet-based predictive model achieved a sensitivity of 1 and PPV of 0.38 using a threshold of 0.02 for predicting relevant articles in the original sample. Applying the GLMnet model to the updated AAP literature search results yielded a sensitivity of 0.921 and PPV of 0.185 using the same threshold; GBM performed similarly.

Table 4 Model Performance for Efficacy/Effectiveness

			GLMnet		GBM	
Study	**Phase**	**Threshold**	**Sensitivity**	**PPV**	**Sensitivity**	**PPV**
AAP	Original	0.001	1	0.144	1	0.383
		0.01	1	0.366	1	0.383
		0.02	1	0.383	1	0.383
		0.1	1	0.421	0.976	0.476
	Update	0.001	1	0.066	0.921	0.186
		0.01	0.921	0.162	0.921	0.186
		0.02	0.921	0.185	0.921	0.187
		0.1	0.901	0.206	0.881	0.232
LBD	Original	0.001	1	0.07	1	0.108
		0.01	0.991	0.143	0.991	0.142
		0.02	0.982	0.174	0.982	0.179
		0.1	0.862	0.322	0.872	0.378
	Update	0.001	1	0.038	0.968	0.06
		0.01	0.937	0.08	0.889	0.08
		0.02	0.905	0.102	0.889	0.106
		0.1	0.778	0.203	0.635	0.181

Note: GLMnet = generalized linear model with convex penalties; GBM = gradient boosting machine; PPV = positive predictive value; AAP = atypical antipsychotic drug systematic review; LBD = low bone density systematic review.

The effectiveness results were similar for the LBD review. The GLMnet-based predictive model achieved a sensitivity of 0.982 and PPV of 0.174 using a threshold of 0.02 for predicting relevant articles in the original sample. We then tested these results on the updated literature search results; GLMnet yielded a sensitivity of 0.905 and PPV of 0.102.

For the AAP review, the GLMnet-based predictive model achieved a sensitivity of 0.978 and PPV of 0.215 using a threshold of 0.02 for predicting articles relevant to AEs in the original sample. Applying the GLMnet-based model to the updated literature search results yielded a sensitivity of 0.981 and PPV of 0.09. The GBM-based model performed better in the original (sensitivity = 1; PPV = 0.274) but worse in the update (sensitivity = 0.895; PPV = 0.11).

The GLMnet-based predictive model achieved a sensitivity of 0.964 and PPV of 0.21 using a threshold of 0.02 for predicting articles relevant for the AE analysis in the original LBD review. However, we were able to predict AE-relevant articles in the update with a substantially reduced sensitivity (0.685) when compared to the AAP results. Reducing the threshold substantially (i.e., retaining all articles with $P \geq 0.001$) would increase sensitivity to 0.946 but decrease PPV to 0.04. The GBM-based model did not produce substantially better results for this analysis.

Table 5 Model Performance for Adverse Effects

			GLMnet		GBM	
Study	**Phase**	**Threshold**	**Sensitivity**	**PPV**	**Sensitivity**	**PPV**
AAP	Original	0.001	1	0.078	1	0.07
		0.01	1	0.168	1	0.138
		0.02	0.978	0.215	1	0.274
		0.1	0.901	0.392	0.934	0.436
	Update	0.001	1	0.033	1	0.029
		0.01	0.99	0.065	0.971	0.056
		0.02	0.981	0.09	0.895	0.11
		0.1	0.867	0.172	0.848	0.2
LBD	Original	0.001	1	0.065	1	0.073
		0.01	0.993	0.175	0.975	0.192
		0.02	0.964	0.21	0.971	0.229
		0.1	0.885	0.338	0.903	0.365
	Update	0.001	0.946	0.04	0.957	0.039
		0.01	0.739	0.097	0.674	0.098
		0.02	0.685	0.116	0.663	0.119
		0.1	0.511	0.179	0.478	0.191

Note: GLMnet = generalized linear model with convex penalties; GBM = gradient boosting machine; PPV = positive predictive value; AAP = atypical antipsychotic drug systematic review; LBD = low bone density systematic review.

Model Prediction Errors

SCEPC researchers independently evaluated articles that were included in the final updated reviews but were assigned relatively low probability scores by the statistical classifiers. At a probability threshold of 0.02 (which reduced workload substantially), the GLMnet algorithm produced 29 false negatives. Of the 29 false negatives, 26 were from the LBD update, and 28 were non-RCT studies (which included meta-analyses, case-control studies, retrospective analyses of claims databases, and analyses of government registries); nearly all such studies were irrelevant in the original LBD review.

The GLMnet-based model rejected 1 study relevant to the AAP update at $P \geq 0.01$.[37] This clinical trial was likely assigned a low probability because it was also tagged as a letter. Despite missing this trial using machine learning, SCEPC researchers might have been able to retrieve it because it was referenced in a relevant article and would plausibly have been caught using the researchers' analyses of references accepted in the final reports.[38]

DISCUSSION

We created a prototype machine learning system that is designed to reduce the workload associated

with comparative effectiveness review updating. Our system first extracted domain knowledge and thousands of previously classified documents from 2 comparative effectiveness reviews and then modeled article relevance using the GBM and GLMnet statistical methods. In 2 simulated comparative effectiveness review updates, our approach reduced the workload associated with screening updated search results for relevant effectiveness and AE articles by more than 50% with minimal or no loss of relevant articles. GLMnet performed slightly better than GBM in this context, but overall model performance was similar despite their substantial theoretical differences. Based on the slight differences in model performance between the various approaches, improving identification of RCTs and refining methods for correcting differences between the original and updated reviews may be more important than algorithm selection in future research.

When simulating a process for selecting all citations relevant to either effectiveness or AEs, we excluded over 50% of irrelevant articles with a loss of 1 of 116 relevant articles for AAP and 0 of 127 relevant articles for LBD. Clearly, the false-positive rate is high (~50%), but this process still could provide substantial value to researchers. One potential problem is that researchers conducting comparative effectiveness reviews aim for 100% sensitivity; despite the high sensitivity rates achieved, the loss of 1 article suggests that researchers will have to make some tradeoffs between sensitivity and efficiency as it will be difficult to guarantee 100% sensitivity without excessively high false-positive rates. On the other hand, it is unclear whether human reviewers can guarantee perfect sensitivity using current processes. In addition, other methods (such as reference mining) can be used to raise sensitivity further. In this case, the single missed reference might have been found by searching within the references of included articles.[38] Furthermore, some errors were the result of key included outcomes being present in the main article text and yet not mentioned in the abstract or MeSH indexing terms. For example, relevant LBD articles often mentioned bone density in the abstract while describing fracture prevention only in the full text. As a result, we assigned a number of articles to the intermediate range because both relevant and irrelevant articles were frequently indexed under bone density. Such data extraction errors were unrelated to improper feature encoding and might only be resolved by analyzing the full text of these articles.

Both GLMnet-based and GBM-based statistical models performed well when selecting AAP citations relevant to effectiveness and AE analyses. By contrast, in the LBD study, both methods achieved substantially lower sensitivity (for a given level of PPV) when predicting AE-relevant articles than when predicting articles relevant to effectiveness analyses. When investigating these false-negative AE articles, we noted that nearly all articles relevant to AEs in the original LBD review were RCTs, possibly because epidemiological studies and retrospective database analyses are difficult to conduct prior to widespread use. This would not have presented a problem if researchers want only RCTs in the update. However, in the update, SCEPC researchers focused on new AEs that were studied in observational studies. As a result, both methods mistakenly assigned lower probabilities to relevant non-RCT studies in the update.

Our results are in accord with prior attempts using machine learning to facilitate comparative effectiveness review data collection.[12,13,18] For example, an active learning model achieved 50% workload reductions and 100% sensitivity in several cases.[18] In contrast to prior studies, we adopted a more parsimonious approach that focused on relatively few study design characteristics (publication type, demographic groups, and statistical design), intervention-specific characteristics, and outcome-specific characteristics. We were able to efficiently identify key terms by mapping terms in search strategies to key terms in MeSH, thus leveraging the substantial time that research librarians invested in creating search strategies. (While this required some upfront time, experienced clinical researchers and research librarians should not require more than 1 to 2 hours to perform this task and possibly considerably less after we integrate automated concept mapping tools such as the National Library of Medicine's MetaMap.[39]) Furthermore, our algorithms explicitly dealt with updating, which required predicting updated citations even though the literature, research personnel, search strategies, and (possibly) some of the underlying goals all changed. Our approach achieved some success in combating data changes over time (known as concept drift in other applications).[16,23,24,26,27] Achieving similar levels of success suggests benefits to a parsimonious approach incorporating domain-specific knowledge about key interventions and outcomes. In addition, this approach allows us to separate effectiveness and AE analyses; although most comparative effectiveness reviews do not separate these analyses, independent filtering mechanisms may be of interest to other researchers.

Our methods carry several limitations. First, our statistical model was very sensitive to MEDLINE's publication type field and MEDLINE indexing generally[40]; developing algorithms that account for discrepancies in MEDLINE classification could dramatically improve model performance. Second, these systems currently work only when PubMed citations are fully indexed, which occurs several weeks or months after publication; excluded publications would require standard human review. This limitation would only affect a small percentage of articles if reviews are being conducted every 2 to 3 years; however, the current system would be inadequate for reviews that require continuous updating and large numbers of citations from non-MEDLINE databases (such as EMBASE). To address the first 2 limitations, we are investigating methods for using additional text features to improve capture of study design details and for classifying unindexed articles, which would make these methods more timely, comprehensive, and accurate. Third, these algorithms also assigned moderately low relevance probabilities to numerous non-RCT articles relevant to AEs, suggesting that this approach cannot mitigate all issues related to concept drift. This suggests some role for an active learning approach that classifies a small number of update articles to maximize accuracy on the update.[13,27] Fourth, this was tested on just 2 comparative effectiveness reviews and will need validation on additional comparative effectiveness reviews and on nontherapeutic applications. Furthermore, we will need to test this system on actual comparative effectiveness review updates to estimate cost savings. To that end, agreement among comparative effectiveness review methodologists on a common data format would facilitate such trials. If such data (article identifier, decision regarding relevance, etc.) were not tabulated in the original report, creating a machine learning model might not be cost-effective given the effort required to format the data properly.

CONCLUSIONS

In this pilot study, we created a prototype system that classified PubMed literature search results from 2 simulated comparative effectiveness review updates. We achieved good performance on both updates using statistical models that were empirically derived from earlier review inclusion judgments as well as explanatory variables selected using domain knowledge. Future research analyzing the article text and incorporating active learning approaches could expand the scope and the accuracy of this method. A more refined system could allow researchers to update their reviews more frequently and efficiently.

ACKNOWLEDGMENTS

We thank Roberta Shanman for the literature searches that were provided for the comparative effectiveness reviews. We also thank Margaret Maglione, Martha Timmer, Elizabeth Roth, and Tanja Perry for their assistance with the data collection.

REFERENCES

1. Higgins J, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.2. The Cochrane Collaboration; 2009.

2. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. 2007;147(4):224–33.

3. Garritty C, Tsertsvadze A, Tricco AC, Sampson M, Moher D. Updating systematic reviews: an international survey. PLoS One. 2010;5(4):e9914.

4. Moher D, Tsertsvadze A, Tricco AC, et al. When and how to update systematic reviews. Cochrane Database Syst Rev. 2008(1): MR000023.

5. Jadad AR, Cook DJ, Jones A, et al. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. JAMA. 1998;280(3):278–80.

6. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med. 2007;4(3):e78.

7. Maher AR, Maglione M, Bagley S, et al. Efficacy and comparative effectiveness of atypical antipsychotic medications for off-label uses in adults: a systematic review and meta-analysis. JAMA. 2011;306(12):1359–69.

8. Shekelle P, Maglione M, Bagley S, et al. Efficacy and Comparative Effectiveness of Off-Label Use of Atypical Antipsychotics. Rockville, MD: Agency for Healthcare Research and Quality; 2007.

9. MacLean C, Alexander A, Carter J, et al. Comparative Effectiveness of Treatments to Prevent Fractures in Men and Women with Low Bone Density or Osteoporosis. Rockville, MD: Agency for Healthcare Research and Quality; 2007.

10. MacLean C, Newberry S, Maglione M, et al. Systematic review: comparative effectiveness of treatments to prevent fractures in men and women with low bone density or osteoporosis. Ann Intern Med. 2008;148(3):197–213.

11. Shekelle P, Takata G, Newberry S, et al. Management of Acute Otitis Media: Update. Rockville, MD: Agency for Healthcare Research and Quality; 2010.

12. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc. 2006;13(2):206–19.

13. Wallace BC, Small K, Brodley CE, Trikalinos TA. Active learning for biomedical citation screening. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC: Association for Computing Machinery; 2011.

14. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc. 2005; 12(2):207–16.

15. Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. J Am Med Inform Assoc. 2009;16(5):690–704.

16. Cohen AM, Ambert K, McDonagh M. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. AMIA Annu Symp Proc. 2010; 2010:121–5.

17. 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;16(1):25–31.

18. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 2010;11:55.

19. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer Verlag; 2009.

20. Genkin A, Lewis D, Madigan D. Large-scale bayesian logistic regression for text categorization. Technometrics. 2007;49: 291–304.

21. Tibshirani R. Regression shrinkage and selection by lasso. J R Stat Soc Series B. 1996;58:267–88.

22. Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. J Am Med Inform Assoc. 2011; 18(5):668–74.

23. Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. Mach Learn. 1996;23(1):69–101.

24. Tsymbal A, Pechenizkiy M, Cunningham P, Puuronen S. Dynamic integration of classifiers for handling concept drift. Inf Fusion. 2008;9(1):56–68.

25. Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia: Association for Computational Linguistics; 2006. p 120–8.

26. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):15.

27. Daume H, ed. Frustratingly easy domain adaptation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague: Association for Computational Linguistics; 2007.

28. Saag KG, Zanchetta JR, Devogelaer JP, et al. Effects of teriparatide versus alendronate for treating glucocorticoid-induced osteoporosis: thirty-six-month results of a randomized, double-blind, controlled trial. Arthritis Rheum. 2009;60(11):3346–55.

29. US National Institutes of Health. MeSH Browser. 2010. Available from: URL: http://www.nlm.nih.gov/mesh/MBrowser.html

30. Freund Y, Schapire R. Experiments with a new boosting algorithm. In: Machine Learning: Proceedings of the Thirteenth International Conference. San Francisco: Morgan Kaufman; 1996. p 148–56.

31. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232.

32. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.

33. Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990;41: 391–407.

34. Deerwester S, Dumais S, Landauer T, Furnas G, Beck L. Improving information-retrieval with latent semantic indexing. In: Proceedings of the ASIS Annual Meeting. Atlanta, GA: American Society for Information Science; 1988. p 36–40.

35. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3): 837–45.

36. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci. 1986;1(1):54–75.

37. Tsuang J, Marder SR, Han A, Hsieh W. Olanzapine treatment for patients with schizophrenia and cocaine abuse. J Clin Psychiatry. 2002;63(12):1180–1.

38. Hamilton JD, Nguyen QX, Gerber RM, Rubio NB. Olanzapine in cocaine dependence: a double-blind, placebo-controlled trial. Am J Addict. 2009;18(1):48–52.

39. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010; 17(3):229–36.

40. Glanville JM, Lefebvre C, Miles JN, Camosso-Stefinovic J. How to identify randomized controlled trials in MEDLINE: ten years on. J Med Libr Assoc. 2006;94(2):130–6.