Methods Research Report

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

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Methods Research Report

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

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-2007-10062-I

Prepared by:
Southern California Evidence-based Practice Center
Santa Monica, CA

Investigators:
Siddhartha R. Dalal, Ph.D.
Paul G. Shekelle, M.D., Ph.D.
Susanne Hempel, Ph.D.
Sydne J. Newberry, Ph.D.
Aneesa Motala, B.A.
Kanaka D. Shetty, M.D., M.S.

AHRQ Publication No. 12-EHC069-EF
September 2012

This report is based on research conducted by the Southern California Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10062-I). The findings and conclusions in this document are those of the author(s), who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

No investigators have any affiliations or financial involvement (e.g., employment, consultancies, honoraria, stock options, expert testimony, grants or patents received or pending, or royalties) that conflict with material presented in this report.

Suggested Citation: Dalal SR, Shekelle PG, Hempel S, Newberry SJ, Motala A, Shetty KD. A Pilot Study Using Machine Learning and Domain Knowledge To Facilitate Comparative Effectiveness Review Updating. Methods Research Report (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I). AHRQ Publication No. 12-EHC069-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2012. http://effectivehealthcare.ahrq.gov.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodological issues in systematic reviews. These methods research projects are intended to contribute to the research base and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers; as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Stephanie Chang, M.D. M.P.H
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Jean Slutsky, P.A, M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Suchitra Iyer, Ph.D.
Task Order Officer
Agency for Healthcare Research and Quality

Acknowledgments

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

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

Structured Abstract

Background. Comparative effectiveness reviews need to be updated frequently to maintain their relevance. Results of earlier screening efforts should be useful in reducing the screening of thousands of newer citations for articles relevant to efficacy/effectiveness and adverse effects (AEs).

Methods. We collected 14,700 PubMed® citation classification decisions from a 2007 comparative effectiveness review of interventions to prevent fractures in persons with low bone density (LBD). We also collected 1,307 PubMed citation classification decisions from a 2006 comparative effectiveness review of off-label uses of atypical anti-psychotic drugs (AAP). We first extracted explanatory variables from the MEDLINE® citation related to key concepts, including the intervention, outcome, and study design. We then used the data to empirically derive statistical models (based on sparse generalized linear models with convex penalties [GLMnet] and gradient boosting machine [GBM]) that predicted inclusion in the AAP and LBD reviews. Finally, we evaluated performance on the 11,003 PubMed citations retrieved for the LBD and AAP updated reviews.

Measurements. Sensitivity (percentage of relevant citations corrected identified), positive predictive value (PPV, percentage of predicted relevant citations that were truly relevant), and workload reduction (percentage of screening avoided).

Results. GLMnet- and GBM-based models performed similarly, with GLMnet (results shown below) performing slightly better. 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 AE-relevant articles for the AAP review (sensitivity=0.981) than for the LBD review (0.685). When attempting to maximize sensitivity, GLMnet achieved high sensitivities (0.99 for AAP and 1.0 for LBD) while reducing projected screening by 55.4 percent (1990/3591 articles for AAP) and 63.2 percent (4,454/7,051 for LBD).

Conclusions. In this pilot study, we evaluated statistical classifiers that used previous classification decisions and key explanatory variables derived from MEDLINE indexing terms to predict inclusion decisions on two simulated comparative effectiveness review updates. The system achieved higher sensitivity in evaluating efficacy/effectiveness articles than in evaluating LBD AE articles. In the simulation, this prototype system reduced workload associated with screening updated search results for all relevant efficacy/effectiveness and AE articles by more than 50 percent with minimal or no loss of relevant articles. After refinement, these document classification algorithms could help researchers maintain up-to-date reviews.

Contents

Tables

Figures

Executive Summary

Background

Comparative effectiveness reviews need to be updated to maintain their relevance, but these updates are often impeded by the need to screen thousands of citations to locate the 1–10 percent that are included in the final report ("relevant studies"). Such effort may match or exceed that involved in the original review. Prior studies have used machine learning methods to reduce the burden of comparative effectiveness review screening but have not formally simulated updating.

Objective

We aimed to create a prototype system for assisting researchers with preparing formal updates of comparative effectiveness reviews. In this report, we describe a pilot study using reviewer decisions from two Agency for Healthcare Research and Quality (AHRQ)-sponsored comparative effectiveness reviews to empirically derive statistical models that predict article relevance to efficacy/effectiveness and adverse effect analyses; we then evaluated these models' performance identifying relevant articles from the literature searches retrieved for the updated reviews. We based these statistical models on two algorithms: gradient boosting machine (GBM) and generalized linear models with convex penalties (GLMnet). Each model predicted an article's relevance based on how its indexing terms described a select number of key concepts (such as publication type, intervention, and outcome). The key challenge was accounting for how search strategies, therapies, outcomes, research personnel, and overall objectives may have changed from the original to the updated study. In accord with an earlier study that noted that a high proportion of reviews underwent minor or major changes, both strategies underwent major revisions. To overcome such challenges (known as "concept drift" in other contexts), we represented specific drugs and outcomes as more abstract concepts such as "intervention" and "outcome," with the hypothesis that this procedure would improve generalizability between time periods.

Methods

We obtained PubMed citations retrieved by the AHRQ Southern California Evidence-based Practice Center (SCEPC) for two review topics (including the early and updated search results): the comparative effectiveness of interventions in preventing fractures in persons with osteoporosis (henceforth referred to as Low Bone Density or LBD) and the efficacy and comparative effectiveness of off-label uses of atypical antipsychotics (AAP). We considered articles to be "relevant" if they passed the second stage screening process and would have been considered for analyses of either efficacy/effectiveness or adverse effects (AEs). We did not exclude duplicates or studies included in prior meta-analyses because these studies were not excluded for intrinsic problems in study design or target population. We did not evaluate PubMed citations that had not yet been assigned MEDLINE indexing information (such as MeSH and Publication Type terms).

The body of articles included in the original LBD report (which we refer to as the training document literature) consisted of 14,700 citations retrieved for screening, of which 382 articles would have passed the second-stage screening—218 for efficacy/effectiveness and 279 for AEs

(some articles included data on both efficacy/effectiveness and AEs). The LBD update corpus consisted of 7,051 retrieved articles (of which 127 would have passed the second-stage filter: 63 for efficacy/effectiveness and 92 for AEs). The AAP training corpus consisted of 1,307 retrieved articles, of which 98 articles would have passed the second stage filter-82 for efficacy and 91 for AEs. The AAP update consisted of 3,591 retrieved articles, of which 116 would have passed the second stage filter--101 for efficacy and 105 for AEs.

Prior study designs used all terms from the index and abstract (i.e., a "bag of words" approach) when modeling relevance. We hypothesized that a limited set of variables that were tightly related to certain key concepts would have substantial predictive power when used to model relevance. To test this hypothesis, we created a limited set of important predictor variables using key MEDLINE Subject Heading (MeSH) indexing terms and associated subheadings. Furthermore, transforming specific concepts such as "alendronate" and "fractures" into abstractions such as "intervention" and "outcome" should allow us to account for changes in outcomes and interventions over time. We selected these key terms by matching terms in the search strategies related to interventions and outcomes to MeSH terms within the MeSH database in a semi-automated fashion; in essence, our approach uses both statistical methods and domain knowledge (extracted from the search strategy) to make the modeling problem tractable. We then created a set of 92 binary explanatory variables representing whether intervention and outcome terms were present in the MEDLINE citation, and how they were described. In addition, we created a set of 29 binary explanatory variables related to article-level characteristics including demographic group (gender and age), treatment target (human, animal, in vitro study, and others), and publication type (review, randomized controlled trial [RCT], clinical trial, meta-analysis, and others), the presence of intervention or outcome terms (or synonyms) in the title, and whether "randomized controlled trial" or "meta-analysis" was mentioned in the title or abstract.

We created a series of eight models based on all combinations of two outcomes (inclusion in the final report for either efficacy or AEs), two statistical learning algorithms (GBM and GLMnet), and training data from two reviews (AAP and LBD). For each model, the inclusion outcome was made a function of extracted explanatory variables for each dataset (described above) using either GBM or GLMnet. GBM is a nonparametric tree-based approach while GLMnet is based on parameterized generalized linear models specifically created to produce sparser models by using convex penalties on the coefficients. We also created a "hybrid" approach that used the maximum prediction probability of relevance from both approaches (GBM- or GLMnet-based). This is equivalent to an approach that rejects only if both GBM- and GLMnet-based approaches reject. In addition, we evaluated how well GBM and GLMnet could predict inclusion for any outcome (efficacy/effectiveness or AEs) in a given review update (AAP or LBD).

To simulate performance in a true update, we generated models using the initial search results while being blind to the true update search results. We generated prediction scores for the updated search (2006–2010 literature for LBD and 2007–2010 for AAP—the test data) using the models described above. We generated a set of predicted relevant and irrelevant articles for the LBD and AAP updates that we compared against decisions that members of the EPC team generated independently. We then calculated performance on the updated results: sensitivity (percentage of relevant articles retrieved, also known as recall), positive predictive value (PPV—percentage of articles predicted to be relevant that were truly relevant, also known as precision), and the percentage of literature search screening that might have been avoided had this predictive

model been used exclusively. We evaluated performance at multiple probability thresholds. There is no perfect threshold, because neither error minimization nor sensitivity maximization can be considered absolute goals; a strategy that rejected all articles might have an error rate of 1 percent (though all would be false negatives) while a strategy accepting all articles would have 100 percent sensitivity (though low PPV). To balance these objectives and conform to researcher preferences, we often judged primary results against a probability threshold of $p \geq 0.02$ because this threshold appeared to substantially reduce the error rate while preserving sensitivity. However, we also derived sensitivity-maximizing thresholds based on performance in the original AAP and LBD studies. We also evaluated the performance of these statistical approaches (GLMnet and GBM) by comparing their receiver operating characteristic (ROC) curves visually and via a nonparametric approach described in DeLong and colleagues.

Results

There were substantial and statistically significant differences in the means of key variables between the original and updated searches, and between categories of each search (excluded, included in efficacy analysis, included in AE analysis, and included in both analyses). These differences suggest that combinations of variables could be used to distinguish between relevant and irrelevant studies; however, the design of the search differed between the update and original searches, which made modeling more difficult.

Model performance differed slightly between the three approaches (GBM, GLMnet, and hybrid), although GLMnet performed slightly better overall. Results below refer to GLMnet. For efficacy analyses, performance in predicting relevant articles was similarly strong for both the AAP and LBD comparative effectiveness reviews. The vast majority of irrelevant citations were assigned relevance probabilities of less than 0.02. The GLMnet approach yielded a sensitivity of 0.921 and PPV of 0.185 when predicting articles relevant to the AAP efficacy update. In considering articles relevant to efficacy for the LBD update, GLMnet model achieved sensitivity of 0.905 and PPV of 0.102 (using the $p \geq 0.02$ threshold).

For the AE analyses, performance in predicting relevant articles was strong for AAP but not for LBD. In the AAP analysis, the GLMnet model achieved sensitivity for AE-relevant articles of 0.981 and PPV of 0.09 at a threshold of $p \geq 0.02$. However, for the LBD study, the GLMnet-based model was able to predict AE-relevant articles with a substantially reduced sensitivity (0.685) for a similar PPV (0.116). When we analyzed articles missed for the AE analysis, we noted that there were relatively few relevant large observational studies (cohort and case-control studies) in the original review. As a result, the model assigned lower probabilities to observational studies in the LBD update as well. However, observational studies were more important in the update because the researchers updating the systematic review focused on several newly identified AEs that were largely studied in cohort and case-control studies.

We also simulated (using GLMnet) a process for estimating potential workload reductions while maximizing sensitivity in identifying all articles relevant to AE and efficacy analyses. Sensitivity and PPV for a particular threshold were determined by selecting articles if the maximum predicted relevance for either outcome (efficacy or AE) exceeded the threshold. For the AAP study, the GLMnet-based model yielded projected sensitivity exceeding 0.99 (115/116 true positives) while the proportion of title/abstract screening saved was 55.4 percent or 1990/3591 articles. The GLMnet-based model produced perfect sensitivity when applied to the LBD update and decreased the projected article screening burden from 7,051 to 2,597 (63.2%). The GLMnet method seemed to perform slightly better than GBM in this context. The AUC for

the GLMnet method (in the AAP study) was 0.943 (95% CI: 0.927 to 0.960) versus 0.925 (95% CI: 0.899 to 0.950) with GBM. The p-value for null hypothesis of equality was 0.007. Similarly, the AUC for the GLMnet method (in the LBD study) was 0.954 (95% CI: 0.943 to 0.965) versus 0.947 (95% CI: 0.933 to 0.961) for GBM. In the LBD study, p-value for null hypothesis of equality was 0.06. Both results suggest that the ROC curves differed between the two studies; in addition, GLMnet seems to perform somewhat better than GBM visually as well. Still, it would be difficult to establish GLMnet's superiority (to GBM, SVM, or other algorithms) in this context (comparative effectiveness reviewing updating) without substantial additional research.

The searchers updating the systematic review independently evaluated articles in the update that were included in the final reports but were assigned relatively low probability scores ($p \leq 0.02$) by the statistical classifiers. There were 29 false negatives at a threshold of $p \leq 0.02$ from both the LBD and AAP updates; nearly all were articles that were not tagged as randomized trials by MEDLINE, and 26 were from the LBD update.

Conclusions

In this report we utilized the large numbers of previously screened documents from two comparative effectiveness reviews to develop models that predicted whether citations retrieved for updated searches would have met final inclusion criteria. We tested several approaches based on the GBM and GLMnet statistical methods. Our approach achieved its best performance predicting relevance for efficacy/effectiveness articles; it performed worse when predicting articles relevant to the AE analysis for the LBD update. However, we estimated that these algorithms reduced (simulated) workload associated with screening updated search results by more than 50 percent with minimal or no loss of relevant articles. Furthermore, the researchers might have been able to retrieve the one excluded article (from the AAP update) 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. Based on the slight differences in model performance between the GBM, GLMnet, and hybrid 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.

Future research is needed on methods for reducing the likelihood of false negatives further (thus reducing the tradeoff between missed articles and workload reductions). Promising methods include incorporating active learning approaches and using text features extracted from the title and abstract to improve capture of study design details, such as RCT design or meta-analysis. We will also need to test this method with other comparative effectiveness review topics. If future work validates these preliminary findings more broadly, we hope to integrate a more refined system into the workflow of comparative effectiveness review researchers. Additional research refining this system, expanding its scope, and comparing it to other methods could allow researchers to select an optimal machine learning method for updating their reviews.

Introduction

Clinicians, clinical guideline developers, regulatory agencies, and research granting agencies all use systematic reviews to determine appropriate clinical practice and research needs. As such, systematic reviews need to be updated to maintain their utility; static reviews potentially ignore new research that could change the results of a systematic review, and thus clinical practice, substantially.[1,2] Given these concerns, several experts suggested updating systematic reviews every 2 years and perhaps more often for rapidly advancing fields.[1,3,4]

In actual practice, a minority of systematic reviews are updated that frequently.[5,6] Several researchers have explored why updating frequency may fall short of the standard. First, updates may entail substantial cost, as the entire process of literature retrieval, filtering, data extraction, and interpretation needs to be repeated. Researchers filter citations in two labor-intensive stages: the first stage excludes articles that are obviously irrelevant based on reading the title and/or abstract; the second stage excludes additional studies after reading or screening the full text.[7-9] In typical comparative effectiveness reviews, researchers retain remaining articles that are useful for analyses of efficacy/effectiveness or AEs (or some other outcome such as utilization). In several AHRQ comparative effectiveness review updates of rapidly advancing fields, researchers needed to screen thousands of citations to locate the 1-10 percent that were relevant;[7-11] such efforts matched or exceeded those involved in the original studies. Second, validated updating protocols and algorithms for determining true signals are still under development.[12,13] Finally, getting updates published in peer-reviewed journals may be more difficult than getting the original review published.[6]

As a first step toward automating part of this process, several studies described information retrieval technologies aimed at improving the efficiency of relevant biomedical literature retrieval.[14,15 16-22] Several focused specifically on reducing the human burden of systematic review creation and updating by limiting the number of retrieved citations that require initial human review.[14,15,17,20] As acknowledged by many of the above studies, systematic review facilitation is complicated by the relative paucity of relevant articles when compared to irrelevant literature; the above studies used a variety of approaches to try to mitigate such issues. These studies typically extracted large numbers of features (explanatory variables) based on variants of a "bag-of-words" approach. In this approach, all terms in the text as well as MeSH indexing terms and publication types were used to create variables based on the presence or frequency of particular terms in the text or MeSH index of each article. Some studies added domain knowledge when classifying features using, for example, United Medical Language System (UMLS).[19] A variety of algorithms have been used to model relevance as a function of these many thousands of potential explanatory variables. These include a support vector machines (SVM), a voting perceptron-based classifier, Naïve Bayes, boosting, a specialized AdaBoost algorithm, and linear and polynomial SVM.[14-17 19,20],

In the specific context of systematic review facilitation, one study used a voting perceptron classifier to identify relevant articles on multiple systematic reviews (and explanatory variables derived from a bag-of-words approach).[14] The study reported work reductions for 11 of the 15 topics while maintaining 95 percent sensitivity for relevant articles; for 3 of those 11, the reduction was more than 50 percent. A later related study used a classifier based on the SVMlight algorithm as well as a bag-of-words feature set to predict updates to multiple systematic reviews prospectively (i.e. using earlier studies to predict studies retrieved in later years).[21] 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 research group used an active-learning strategy to aid the creation of new systematic reviews.[15,20] Similar to the above study, the model predicts relevance based on independent variables derived from multiple sources including MeSH and text. However, they used a process that interactively builds a classifier using expert decisions on the most uncertain cases; the underlying hypothesis is that decisions chosen on the most uncertain instances produce better information for a given cost (reviewer time). They were able to reduce the number of citations that needed to be screened in a simulated *de novo* review by roughly 50 percent while retaining 100 percent of relevant articles.

The above studies all employed thousands of explanatory variables. While predictive accuracy is the main goal of machine learning, model parsimony may contribute toward improving out-of-sample predictions.[23-25] We hypothesized that a parsimonious approach making explicit use of domain knowledge might be useful in comparative effectiveness review updating, because the original reviews frequently generate thousands of training observations and include substantial domain knowledge. We further hypothesized that the effort MEDLINE researchers put into indexing key concepts with subheadings can be leveraged for substantial predictive power, without requiring significant human reviewer input.

MeSH indexing identifies key concepts in articles, which has proven very useful as a tool for retrieving literature. However, MeSH indexing further describes those concepts with descriptive subheadings ("chemically induced", "adverse effects", "epidemiology", etc.). Extracting data solely on a few key variables related to publication type, intervention, and outcome may have sufficient power, counterbalancing the slightly greater upfront time required for identifying key concepts beforehand. For example, if researchers are interested whether alendronate is safe and effective in preventing fractures, the most crucial variables might be those indicating whether the indexing term for alendronate is tagged with “therapeutic use” and whether the corresponding term for fracture is associated with “prevention and control”. These and other key variables might be imperfect individually but their combination could yield a model that robustly predicts relevance. We published earlier work validating this approach for extracting articles that tested whether particular drugs caused any type of adverse effect (AE) regardless of study design.[26] In this study, we adapted our earlier approach to make it useful for locating studies relevant to comparative effectiveness reviews, which require all articles assessing either efficacy/effectiveness or AEs using particular study designs.

We tested the utility of a MeSH-based approach to text classification using two comparative effectiveness reviews. The first concerned the prevention of fractures in patients with osteopenia or osteoporosis (low bone density [LBD]) conducted by the Southern California Evidence-based Practice Center (SCEPC), under contract to the Agency for Healthcare Research and Quality (AHRQ).[7,8] EPC researchers conducted the initial comparative effectiveness review in 2006 using literature indexed in multiple sources including PubMed. As it became apparent that the field was advancing rapidly, the EPC group and AHRQ determined that an update was necessary; an updated literature search was conducted in 2010 that searched for new literature (2006-2010) covering the same interventions and conditions, as well as all literature for newly relevant interventions and conditions (discussed below). The second comparative effectiveness review covered off-label indications for atypical anti-psychotic drugs (AAP).[9] The first AAP review covered literature published until December 2006, and the update covered literature

published subsequent to that date. In each case, we aimed to use the earlier study's reviewer decisions to create a predictive model that could classify articles in the updated search results as meeting the criteria of second-stage filtering. Such studies were relevant for efficacy/effectiveness analyses, relevant for AE analyses, or irrelevant for both.

The key challenge in using machine learning to facilitate systematic and comparative effectiveness review updates lies in accounting for the differences between the training (original search) and test (updated search) data. In both the LBD and AAP reviews, new conditions and interventions were added and others were dropped. Furthermore, both research personnel and study objectives changed between the first and updated review. In other contexts, the general problem of training data becoming inapplicable to test data over time has been described as "concept drift".[27-31] Researchers have devised several strategies to address this problem including giving weight to more recent training observations or to the small number of classified test observations (if present). An earlier study explored how well an SVM-based machine learning framework performed in light of this issue on a series of systematic reviews, some of whose search criteria were revised significantly.[21] They found that their framework performed well in many cases, though they noted that it was uncertain whether reviewers would be satisfied with their prospective performance in all cases and they did not explicitly address changes over time.

In this study, we simulated a true update by creating predictive models while blinded to "true positive" and "true negative" articles from the update, while addressing the fact that both the AAP and LBD reviews substantially revised their inclusion criteria. To address concept drift in this context, we therefore elected to represent specific drugs and outcomes as more abstract concepts such as "intervention" and "outcome", with the hypothesis that this procedure would improve generalizability between time periods. We hypothesized that reviewers wanted the same types of studies, even though the exact interventions and outcomes differed. In the rest of the manuscript, 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 its review of the comparative effectiveness of interventions in preventing fractures in persons with osteoporosis (LBD) and its review of the efficacy and comparative effectiveness of off-label uses of atypical antipsychotics (AAP). MEDLINE citations were retrieved in plain text format and parsed using Python 2.7.2 (Python Software Foundation, http://python.org); specifically, we used the Biopython 1.52 package within Python to retrieve full MEDLINE citations from Entrez PubMed databases.[32] We did not evaluate PubMed citations that had not yet been assigned MEDLINE indexing information (such as MeSH and Publication Type terms). We also excluded 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 search strategies and primary selection criteria for LBD have been discussed extensively in other reports.[7-9] Briefly, in the LBD study, the interventions consisted of multiple drugs (including bisphosphonate drugs, calcitonin, selective estrogen receptor modulators, parathyroid hormone derivatives, and menopausal hormone therapy) and exercise therapy. The primary outcomes of interest were fractures and AEs but the search strategy also attempted to capture articles discussing predisposing conditions by searching for terms such as osteoporosis, osteopenia, and bone mineral density, as well as fractures. The search was limited to English language articles, but no limits were placed on publication type. The initial search (1966-2006) yielded 14,700 articles with full MEDLINE citations, and the updated search (containing a slightly different set of interventions) retrieved 7,051 articles with full MEDLINE citations (spanning 2006-2010). We did not analyze 219 PubMed articles from the LBD updated search that were not indexed in MEDLINE, as our algorithms currently require MEDLINE indexing information.

The search strategies and selection criteria for AAP have also been discussed in other reports.[9] 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 post-traumatic stress disorder, and outcomes could be excluded if re-classified by the U.S. Food and Drug Administration (FDA) as an approved indication. The search conducted in 2006 yielded 1,307 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; 3,591 MEDLINE citations were retrieved. We did not analyze 19 PubMed articles from the AAP original search and 142 PubMed articles from the updated search that were not indexed in MEDLINE. For both studies, articles retrieved for update and original searches were mutually exclusive.

During the initial modeling phase, we had access to researcher decisions on the original search results. The LBD training document literature consisted of 14,700 retrieved articles, of which 382 articles would have passed the second stage filter: 218 for efficacy/effectiveness and 279 articles for AEs. The LBD update body of literature consisted of 7,051 retrieved articles (of

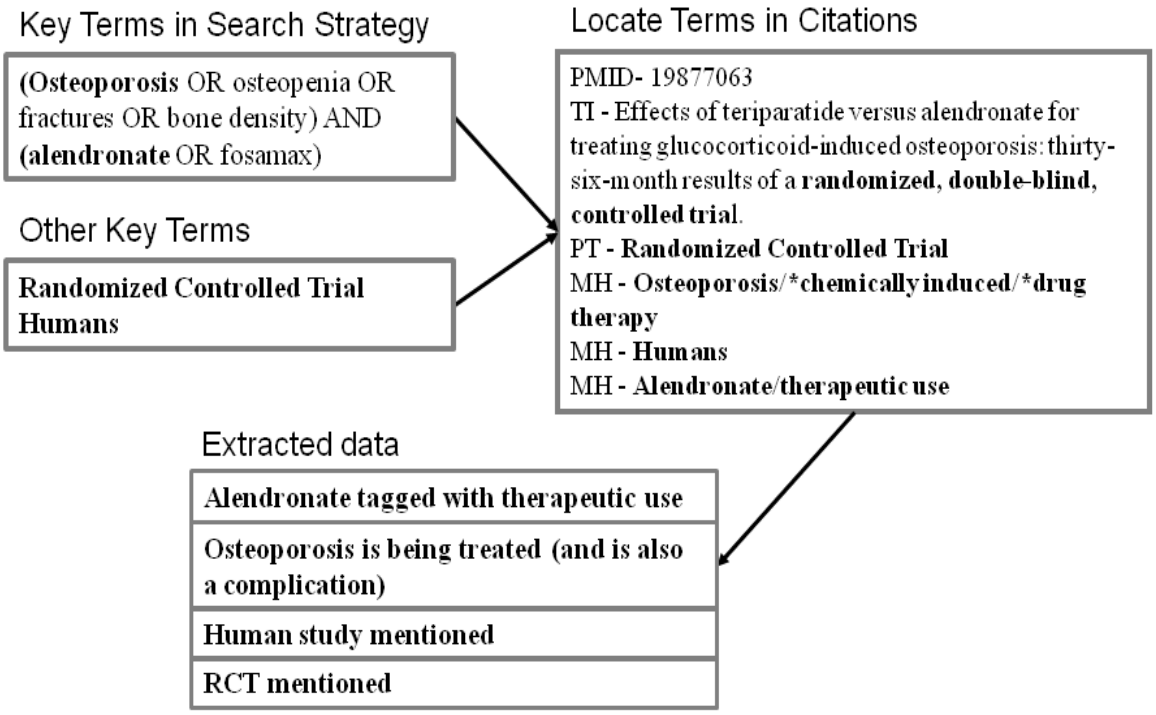
which 127 would have passed the second stage filter: 63 for efficacy/effectiveness and 92 for AEs). The AAP training literature consisted of 1,307 retrieved articles, of which 98 articles would have passed the second stage filter: 82 for efficacy/effectiveness and 91 for AEs. The AAP update consisted of 3,591 retrieved articles, of which 116 would have passed the second stage filter: 101 for efficacy/effectiveness and 105 for AEs. Of note, articles were excluded at the second stage for critical reasons (e.g., inappropriate study design, no mention of fractures, inappropriate intervention, outcome was an on-label indication, 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 they were not excluded for intrinsic problems in study design or target population. We blinded the statistical learning model to researcher decisions involving the updated search results; as such, we could effectively simulate a true update in which the update search results would not have been known.

We did not have access to completely accurate determinations of first-stage filtering outcomes; as a result we did not include this outcome in any model. In addition, while we did not model first-stage outcomes due to data limitations, we also believed that such articles were not important to the final results; indeed, reducing this number was also a researcher goal as all articles passing the first stage required a time-consuming full-text review.

Processing MEDLINE Citations

Each fully indexed MEDLINE citation contains several (usually 10–15) indexing terms; each term is often modified by one or more subheadings. As described above, we aimed to construct a limited set of important variables using key MeSH indexing terms and associated subheadings that are tied to the interventions and outcomes of interest. Figure 1 shows how a data was extracted from one citation using terms adapted from the 2006 LBD search strategy.[7,33] Of note, data related to generic study characteristics (such as RCT or human study) were extracted from all studies. The key MeSH terms were found using the extent search strategies. One author divided each search strategy into terms related to interventions and terms related to 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)".[7] Hence, we could identify interventions (raloxifene, etc.) and outcomes (osteoporosis, osteopenia, etc.) without substantial effort. (Clearly, we will need to expend more effort on poorly defined search strategies.) At that point the MeSH database was programmatically queried to retrieve potential matches (obtained by via direct term and synonym matching) that were reviewed by one of the authors. Several erroneous terms were excluded at this stage. For example, "exercise test"—a diagnostic test—was retrieved when searching for exercise but manually excluded because only therapeutic exercise therapy was relevant to this review. Because this process relies heavily on the original search strategies, it does not use substantially more expertise than what went into the original report.[34]

Figure 1. MEDLINE citation processing example



MH = MEDLINE indexing term; PMID = PubMed identifier; PT = publication type; RCT = randomized controlled trial; TI = title

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 not the outcomes and interventions of interest) were present in the MEDLINE citation and linked to particular subheadings. For example, if the only outcome of interest in a particular article is "fractures," if "fractures" are indexed in the citation in association with "drug therapy", we set the variable "outcome_drug_therapy" equal to one. However, if in the same article, "rheumatic diseases" is also indexed in conjunction with "drug therapy", we would set a variable "other_outcome_drug_therapy" equal to 1 as well. The latter might indicate that other diseases were of primary importance in the article. We used a similar process for interventions.

In addition, we created a set of 29 binary explanatory variables related to broader characteristics from MeSH indexing terms and publication type terms-- including demographic group (gender and age), treatment target (human, animal, in vitro study, and others), and publication type (review, clinical trial, meta-analysis, and others). Finally we created variables indicating whether any intervention or outcome (or synonym of either) was explicitly mentioned in the title or in the article's MeSH index, whether the article was particularly short (1 or 2 pages in length), and whether “randomized controlled trial” or “meta-analysis” was mentioned in the title or abstract. Our approach is parsimonious in that we used only these 121 variables, instead of the full text approach that would have to deal with potentially thousands of explanatory variables and consequently would have the potential for overfitting, resulting in possible loss of out-of-sample predictive power. Although we relied on expert knowledge to some extent while

extracting such data, we relied on statistical algorithms to select the most important features (see below).

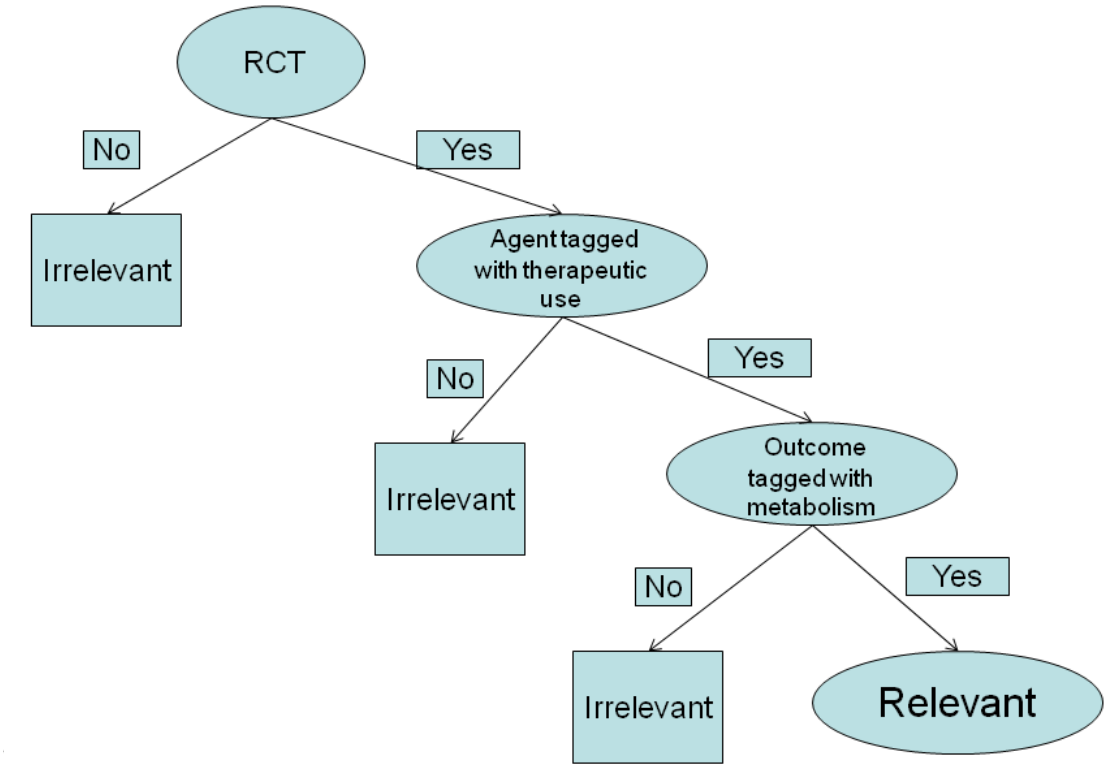
Statistical Classification

We created a series of eight models based on all combinations of: two outcomes (inclusion in the final report for either efficacy or AEs); training data with associated explanatory variables from two reviews (AAP and LBD); and, two statistical learning algorithms (GBM and GLMnet). After deriving each model empirically using training data from original review, we generated predictions for articles in each corresponding update. All statistical modeling was conducted in R 2.10 (R Foundation, www.r-project.org/). We also used these base models to create combination analyses. We tested a "hybrid" approach that used the maximum prediction probability of relevance from both approaches (GBM- or GLMnet-based). 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). We modeled relevance as a function of the explanatory variables discussed above while solely using those articles retrieved in the original search (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. Each step is explained in detail below.

We constructed separate models for predicting inclusion in efficacy/effectiveness or AE analyses because article characteristics predictive of relevance were likely to be quite different between the two analyses. For both, we aimed to retain the maximum number of relevant citations (true positives), while minimizing the number of irrelevant citations detected (false positives). We also evaluated our models' performance in predicting inclusion in either analysis. The latter analysis is most relevant to current AHRQ practice, as both efficacy/effectiveness and AE analyses are required for comparative effectiveness reviews. However, we showed disaggregated results as well because other researchers may be interested in one type of study.

We determined each model specification (efficacy/effectiveness and AE analyses for both LBD and AAP) using several statistical methods. The first method we considered was gradient-boosting machine (GBM), a non-parametric tree based prediction approach based on boosting.[35,36] In the general boosting framework, models are built in a stage-wise fashion, with weak (i.e. moderately inaccurate) classifiers combined to create 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).

Figure 2. Example GBM tree



GBM = gradient boosting machine

A single, simple classifier as above is inadequate for generating accurate predictions. In the example given in Figure 2, it is obvious that automatically discarding articles not tagged as RCTs would exclude relevant articles (such as systematic reviews). However, the GBM algorithm generates a model based on a series of simple classifiers, including, for example, decision trees that discard articles that are not systematic reviews. The algorithm sequentially evaluates each simple model and assigns it 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 models, but each simple tree is assigned a weight proportional to its accuracy. By taking a weighted average across simple, weak classification trees, it is possible to generate more accurate predictions.

We validated the results on these training data using five-fold cross validation (which reduces overfitting). Each fold of cross validation randomly selects 20 percent of the data to serve as test data; then the process fits a model on the remaining 80 percent of the data; finally, model performance is measured on the reserved test data. The process is repeated on all 5 folds and one ultimately finds the model which would minimize the prediction error averaged across all 5 folds and models. This approach reduces both overfitting (using cross validation) and improves overall performance (using boosting). The output results were probabilities that the articles were relevant; we examined a receiver operator characteristic (ROC) curve to determine

the optimal probability threshold for minimizing both false negatives and false positives using only the original search results.

We also used the GLMnet method, which is a parametric approach in that one fits a linear logistic model with convex penalty on the magnitude of coefficients. As above, we model the outcome variable (inclusion in the report for efficacy/effectiveness or AEs) as a function of the explanatory variables described above. In a standard linear model, the outcome would be made a function of all explanatory variables, but this may lead to over-fitting. The Lasso shrinkage and selection method for linear regression (and generalizations such as Elastic-Net) minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients.[23] The GLMnet method shrinks coefficients of less important variables to zero with a more general convex penalty, resulting in fewer independent variables that have better predictive power. GLMnet also employs cyclical coordinate descent (computed along a regularization path) to efficiently solve these problems.[37] Both algorithms (GLMnet and GBM) outputted prediction probabilities that could be judged against the gold standard results obtained by the SCEPC team. Finally, to bridge some of the large differences between the parametric GLMnet and non-parametric GBM procedures, we created a hybrid approach that would reject articles only if both procedures rejected them. Equivalently, we accepted articles if either procedure assigned sufficiently high probability of relevance to them. Of note, we considered using SVM and other established methods such as latent semantic indexing, 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 and GLMnet.[16,38,39]

We generated prediction scores for the updated searches (2006-2010 literature for LBD and 2007-2010 for AAP–the test data) using the models and thresholds generated above. We generated a set of predicted relevant and irrelevant articles that we compared against decisions that members of the EPC team generated independently. We then calculated performance: sensitivity (% relevant articles retrieved, also known as recall) and positive predictive value (PPV: % predicted relevant articles that were truly relevant, also known as precision). We also computed the proportion of workload (literature search screening of both relevant and irrelevant articles) that might have been avoided had this predictive model been used exclusively.

We evaluated performance at multiple probability thresholds. There is no perfect threshold, because neither error minimization nor sensitivity maximization can be considered absolute goals; a strategy that rejected all articles might have an error rate of 1 percent (though all would be false negatives) while a strategy accepting all articles would have 100 percent sensitivity (though low PPV). To balance these objectives and conform to researcher preferences, we often judged primary results against a probability threshold of $p \geq 0.02$ because this threshold appeared to substantially reduce the error rate while preserving sensitivity. However, we also derived sensitivity-maximizing thresholds based on performance in the original AAP and LBD studies. As described below, these empirically derived thresholds differed between the two studies. We also evaluated the performance of these approaches (GLMnet and GBM) by comparing their Receiver Operating Characteristic (ROC) curves visually and via a non-parametric approach. [40] Of note, pure statistical comparisons may not produce the best results (from the perspective of comparative effectiveness review researchers) because maximizing the area under the curve (AUC) does not automatically select the proper balance of sensitivity and PPV/specificity.

To estimate model variability, we calculated bootstrapped standard errors for the sensitivity and PPV results.[41] We generated 100 models by sampling with replacement from the original literature review articles. We then generated 100 sets of predictions by applying each of the

models to the actual data (from the original and updated reports); we calculated standard errors from the resulting simulated sensitivity and PPV estimates. However, at the plausible thresholds discussed in the report ($0<p<=.1$), the standard errors were extremely small (due to the large sample sizes of the training data used to fit the original models) and are not shown for each case. For example, for sensitivity at a threshold of 0.1 for the original LBD study (efficacy), the estimated sensitivity was 0.995 and the standard error was 0.0008.

Results

Literature Characteristics

Table 1 shows the characteristics of the original and updated AAP literature searches; each column (original and update) represents both excluded and relevant studies. We compared the proportions of each variable within the original and update search results using Fisher's exact test. Substantial and statistically significant differences were observed between the means of variables in the AAP original and updated searches. This finding suggests that the composition of the search results (if not necessarily the included studies) differed substantially between the update and original searches.

Table 1. AAP characteristics: original versus update

Variable	Original (Count, Proportion)	Update (Count, Proportion)	Comparison of Means (p-value)*
Number of Studies	1307	3591	
Year§	2000.9	2005.7	
(range)	1972-2006	1988-2011	
Any Outcome In Title	432 (0.331)	1394 (0.388)	<0.001
Any Agent In Title	893 (0.683)	1979 (0.551)	<0.001
Agent & Administration	254 (0.194)	586 (0.163)	0.011
Agent & Therapeutic Use	937 (0.717)	2334 (0.650)	<0.001
Agent & Toxicity	581 (0.445)	1368 (0.381)	<0.001
Demographic Tags Include Child	233 (0.178)	822 (0.229)	<0.001
Outcome & Complications	104 (0.080)	300 (0.084)	0.681
Outcome & Drug Therapy	542 (0.415)	1284 (0.358)	<0.001
Outcome & Prevention	5 (0.004)	39 (0.011)	0.024
Outcome & Psychology	290 (0.222)	648 (0.180)	0.001
Other Outcome & Psychology	305 (0.233)	657 (0.183)	<0.001
Clinical Trial	375 (0.287)	451 (0.126)	<0.001
Comparative Study	259 (0.198)	608 (0.169)	0.02
Meta-Analysis	24 (0.018)	83 (0.023)	0.377
RCT	214 (0.164)	501 (0.140)	0.035
Text Contains RCT	133 (0.102)	414 (0.115)	0.2

*P-value derived from Fisher's Exact Test; RCT, Randomized Controlled Trial
§ Year is reported as mean year of publication.

Table 2 shows the characteristics for the AAP original search by category (excluded, included for AE analysis, included only for the efficacy/effectiveness analyses, included for both analyses). There are obviously substantial differences, as revealed by the one-way Anova test comparing means in all four groups; these differences were highly significant for most key variables including "RCT." The importance of each variable is unknown, but the differences suggest that combinations of variables could be useful in distinguishing between included and excluded studies.

Table 2. Characteristics of the original AAP review (by category of article)

Variable	Excluded	Efficacy only*	AE Only	Both Types of Outcomes	Comparison of Means (p-value)**
Number of Studies	1209	7	16	75	
Year§	2000.8	2002	2003.8	2002.6	NA
Any Outcome In Title	0.307	0.714	0.312	0.68	<0.001
Any Agent In Title	0.667	0.714	0.75	0.933	<0.001
Agent & Administration	0.187	0.286	0.188	0.307	0.077
Agent & Therapeutic Use	0.706	0.857	0.688	0.88	0.011
Agent & Toxicity	0.432	0.143	0.938	0.573	<0.001
Demographic Tags Include Child	0.17	0.286	0.062	0.333	0.002
Outcome & Complications	0.079	0.143	0	0.107	0.469
Outcome & Drug Therapy	0.405	0.429	0.375	0.573	0.04
Outcome & Prevention	0.004	0	0	0	0.939
Outcome & Psychology	0.207	0.429	0.062	0.48	<0.001
Other Outcome & Psychology	0.227	0.286	0	0.387	0.002
Clinical Trial	0.246	1	0.062	0.933	<0.001
Comparative Study	0.174	0.286	0.625	0.493	<0.001
Meta-Analysis	0.02	0	0	0	0.576
RCT	0.108	1	0.062	1	<0.001
Text Contains RCT	0.086	0.286	0.062	0.347	<0.001

*Efficacy includes effectiveness analyses
**P-value derived from Pearson's Chi-squared Test;
§ Year is reported as mean year of publication
RCT = randomized controlled trial; NA = not applicable

Table 3 shows select characteristics of the LBD literature; we show the same characteristics as in the AAP update (Tables 1 and 2) to demonstrate how characteristics may vary between different review topics. The original search results were published from 1966 to 2009 (articles published after 2006 were electronically published in 2006). The updated search results were predominantly published from 2007 to 2010, with some articles published from 1997 to 2006 and in 2011. Roughly 10 percent of the retrieved studies were classified as RCTs in MEDLINE in both the original and updated literature searches. As noted in the third column of Table 3, the presence of several key variables differed substantially between the original and updated searches in univariate comparisons. In particular, the update included non-human studies and proportionally fewer articles in which the outcome was associated with drug therapy. This finding suggests that the original and updated data were somewhat different, which made creation of a generalizable model more difficult.

Table 3. Characteristics of LBD search results (original vs. updated)

Variable	Original (Count, Proportion)	Update (Count, Proportion)	Comparison of Means (p-value)*
Number of Studies	14,700	7,051	
§ Year	1997.6	2007.5	
(range)	1966-2009	1997-2011	
Any Outcome In Title	6478 (0.441)	2431 (0.345)	<0.001
Any Agent In Title	5770 (0.393)	3572 (0.507)	<0.001
Agent & Administration	1364 (0.093)	1218 (0.173)	<0.001
Agent & Therapeutic Use	3900 (0.265)	1916 (0.272)	0.318
Agent & Toxicity	1149 (0.078)	1046 (0.148)	<0.001
Demographic Tags Include Child	2545 (0.173)	986 (0.140)	<0.001
Outcome & Complications	1187 (0.081)	544 (0.077)	0.363
Outcome & Drug Therapy	2929 (0.199)	1246 (0.177)	<0.001
Outcome & Prevention	2606 (0.177)	1266 (0.180)	0.691
Outcome & Psychology	67 (0.005)	29 (0.004)	0.743
Other Outcome & Psychology	142 (0.010)	76 (0.011)	0.467
Clinical Trial	1992 (0.136)	277 (0.039)	<0.001
Comparative Study	1711 (0.116)	544 (0.077)	<0.001
Meta-Analysis	88 (0.006)	121 (0.017)	<0.001
RCT	1542 (0.105)	711 (0.101)	0.366
Text Contains RCT	0.061	0.087	0.000

*P-value derived from Fisher's Exact Test; § Year is reported as mean year of publication
RCT = randomized controlled trial

Table 4 shows the original literature search results for LBD in greater detail, and compares characteristics among four categories (excluded studies, considered only for efficacy/effectiveness analyses, considered only for AE analysis, and considered for both AE and efficacy/effectiveness analyses). As is clear from the table, none of the predictors function perfectly. However, substantial differences exist for multiple variables, which make modeling based on some combination of these variables feasible via a regression approach. As expected, the vast majority of relevant studies were either meta-analyses or RCTs; in contrast, the results in irrelevant studies were occasionally tagged as *in vitro* or animal studies (not shown). Furthermore, large majorities of studies in every included category (efficacy, AE, or both analyses) contained indexing information that described the therapeutic use of a preferred intervention or the treatment of a preferred outcome. By contrast, relatively few excluded studies contained indexing information that linked the therapeutic use of a preferred intervention (0.257) or the treatment of a preferred outcome (0.192).

Table 4. Characteristics of the original LBD review (by category of article)

Variable	Excluded	Efficacy Only*	AE Only	Both Types of Outcomes	Comparison of Means (p-value)**
Number of Studies	14318	103	164	115	
§ Year	1997.5	2001.1	2000.7	2001.4	
Any Outcome In Title	0.433	0.806	0.604	0.8	<0.001
Any Agent In Title	0.378	0.806	0.963	0.983	<0.001
Agent & Administration	0.084	0.33	0.5	0.417	<0.001
Agent & Therapeutic Use	0.253	0.777	0.634	0.809	<0.001
Agent & Toxicity	0.071	0.097	0.445	0.426	<0.001
Demographic Tags Include Child	0.177	0.029	0.03	0.043	<0.001
Outcome & Complications	0.078	0.34	0.043	0.209	<0.001
Outcome & Drug Therapy	0.188	0.689	0.537	0.661	<0.001
Outcome & Prevention	0.169	0.67	0.317	0.6	<0.001
Outcome & Psychology	0.004	0.019	0.006	0	0.13
Other Outcome & Psychology	0.01	0	0.006	0	0.494
Clinical Trial	0.116	0.816	0.878	0.887	<0.001
Comparative Study	0.114	0.126	0.22	0.217	<0.001
Meta-Analysis	0.005	0.136	0	0	<0.001
RCT	0.083	0.835	0.902	1	<0.001
Text Contains RCT	0.052	0.495	0.354	0.461	<0.001

*Efficacy includes effectiveness analyses
**p-value derived from Pearson's Chi-squared test
§ Year is reported as mean year of publication

Performance Predicting Efficacy/Effectiveness Results

Predicting Articles Relevant to Efficacy/Effectiveness for AAP Review

We developed a model for predicting the inclusion of efficacy/effectiveness articles using the original search results. Figure 3 shows the relative weights of different variables for GBM; variables with larger relative weights account for large fractions of the total explanatory power. In keeping with some of the differences in frequency distributions between included and excluded studies, "RCT" contains a substantial portion of the model's explanatory power. Weights for GLMnet were similar, with "RCT" providing the greatest explanatory power.

Figure 3. Relative weights for variables in AAP efficacy analysis

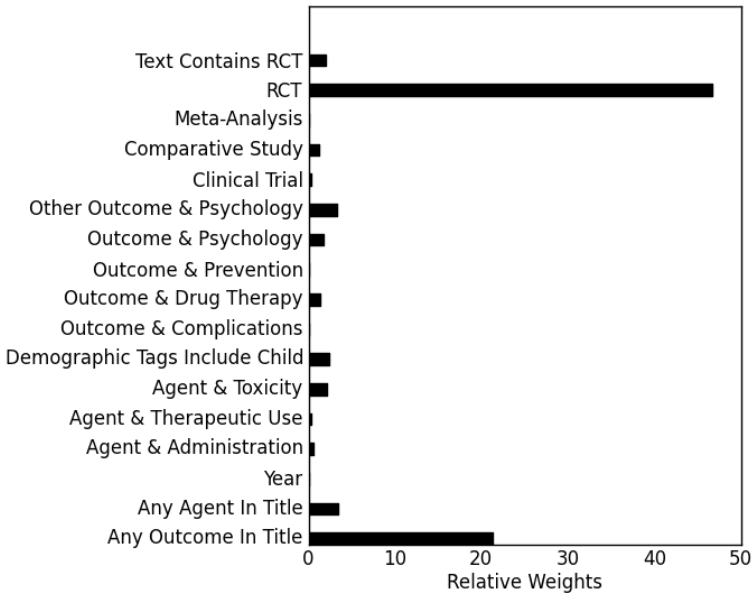


Table 5 shows efficacy/effectiveness results for all models (GLMnet, GBM, and hybrid) at multiple thresholds. For AAP, all models achieved high sensitivity when predicting on the original sample at relatively high thresholds ($p \leq 0.02$). For example, the GLMnet-based predictive model achieved a sensitivity of 1 and PPV of 0.38 using a threshold of 0.02 for predicting relevant articles in the original sample. Achieving good results on the original sample was expected because the underlying model was derived from the same outcomes and explanatory variables. Applying the GLMnet model to the updated AAP literature search results yielded a sensitivity of 0.921 and PPV of 0.185; GBM and hybrid models performed similarly.

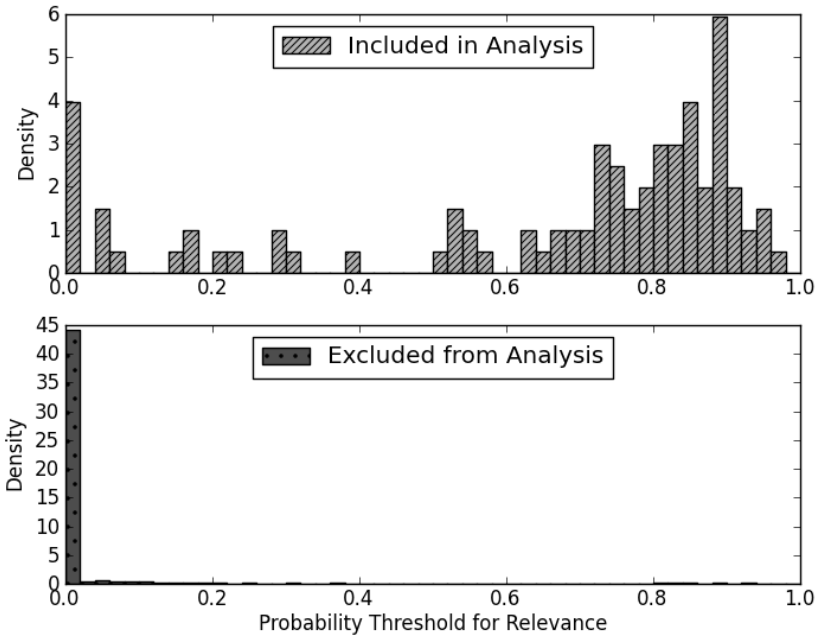
Table 5. Model performance for efficacy/effectiveness

			GLMnet		GBM		Hybrid	
Study	**Phase**	**Threshold**	**Sensitivity**	**PPV**	**Sensitivity**	**PPV**	**Sensitivity**	**PPV**
AAP	**Original**	0.001	1	0.144	1	0.383	1	0.144
		0.01	1	0.366	1	0.383	1	0.366
		0.02	1	0.383	1	0.383	1	0.383
		0.1	1	0.421	0.976	0.476	1	0.418
	Update	0.001	1	0.066	0.921	0.186	1	0.066
		0.01	0.921	0.162	0.921	0.186	0.921	0.162
		0.02	0.921	0.185	0.921	0.187	0.921	0.185
		0.1	0.901	0.206	0.881	0.232	0.901	0.205
LBD	**Original**	0.001	1	0.07	1	0.108	1	0.068
		0.01	0.991	0.143	0.991	0.142	0.991	0.133
		0.02	0.982	0.174	0.982	0.179	0.986	0.168
		0.1	0.862	0.322	0.872	0.378	0.894	0.321
	Update	0.001	1	0.038	0.968	0.06	1	0.037
		0.01	0.937	0.08	0.889	0.08	0.937	0.075
		0.02	0.905	0.102	0.889	0.106	0.905	0.098
		0.1	0.778	0.203	0.635	0.181	0.794	0.192

GLMnet = Generalized Linear Models with Convex Penalties; GBM = gradient boosting machine; Hybrid = Maximum prediction from either GLMnet or GBM; PPV = Positive predictive value
Note: We calculated bootstrapped standard errors for the GLMnet estimates. In all cases, the standard errors were substantially smaller (<0.005) than the estimates for sensitivity or PPV.

Figure 4 shows these results graphically using a histogram of the prediction probabilities for the update, divided according to whether the article met final inclusion criteria. Excluded articles were predominantly given probabilities very close to zero, while articles considered for efficacy/effectiveness had probabilities that spanned the entire spectrum. Of note, this histogram displays densities; even small densities of false positive articles (from the much larger group of negative articles) entail a relatively high proportion of false positives among model predictions, which limits the PPV to 0.185.

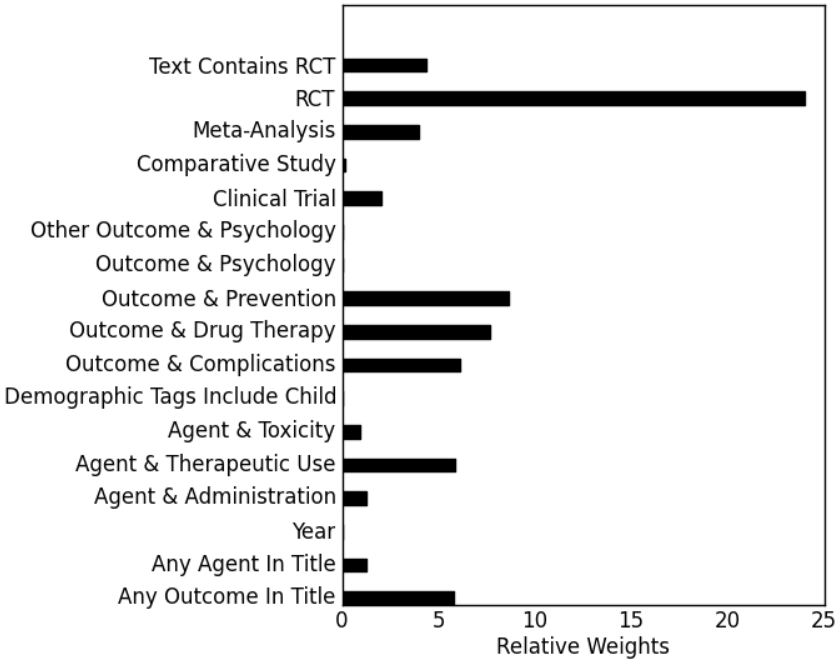
Figure 4. Histogram AAP efficacy analysis: distribution of predictions



Predicting Articles Relevant to Efficacy/Effectiveness for LBD Review

Figure 5 shows the relative weights of variables included in the GBM model of efficacy for LBD (weights for GLMnet were similar, in that RCT contained the greatest explanatory power). As in the AAP analysis, terms such as RCT and meta-analysis are important. Clearly, other variables carried different weights in the AAP analysis, suggesting that predictive models may need to be topic-specific.

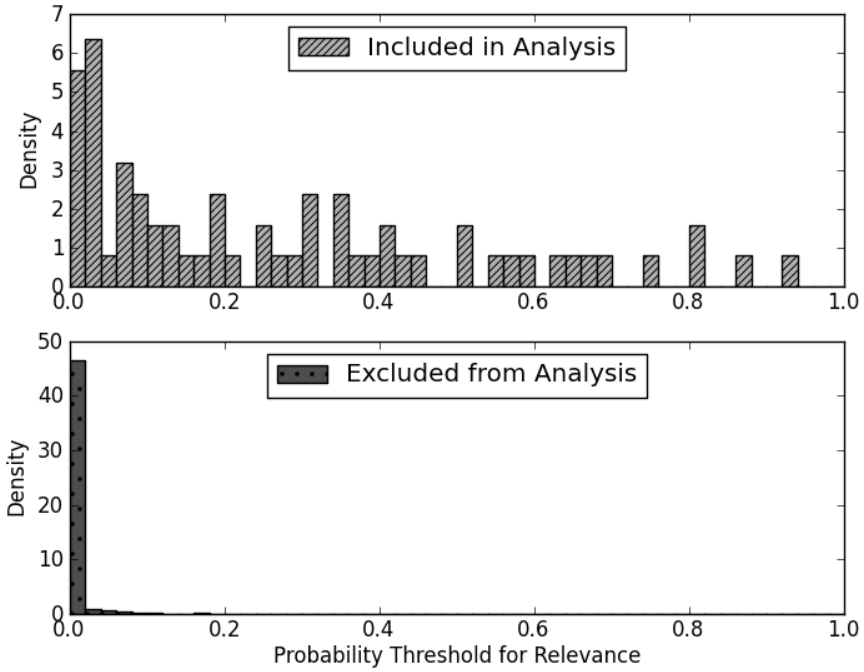
Figure 5. Relative weights for variables in LBD efficacy analysis



The efficacy/effectiveness results were similar for the LBD review (Table 5.) The GLMnet-based predictive model achieved 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 sensitivity of 0.905 and PPV of 0.102.

Figure 6 shows model prediction performance on the LBD updated search graphically using a histogram of the prediction probabilities. Excluded articles were generally assigned very low probabilities. As in Figure 4 (for AAP), the small percentage of false positive articles reduced the PPV to 0.102 due to the much greater number of negative articles overall.

Figure 6. Histogram LBD efficacy analysis: distribution of predictions

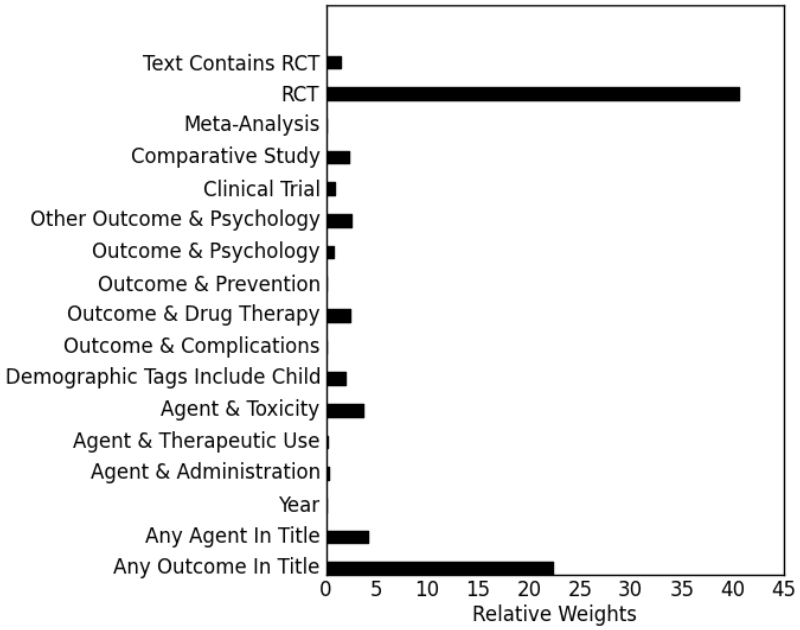


Performance Retrieving Articles Considered for AE Analysis

Predicting AE-Relevant Articles for AAP Update

We empirically developed a model for predicting AE articles using the original search results. We show the relative importance of the same select variables in Figure 7 for GBM (though GLMnet produced similar weights). Again, the "RCT" variable remains extremely important, even as the importance of the remaining explanatory variables differs from the efficacy/effectiveness models.

Figure 7. Relative weights for variables in AAP AE analysis



We show results from all models in Table 6. 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 hybrid model yielded similar sensitivity to the GLMnet model, but worse PPV.

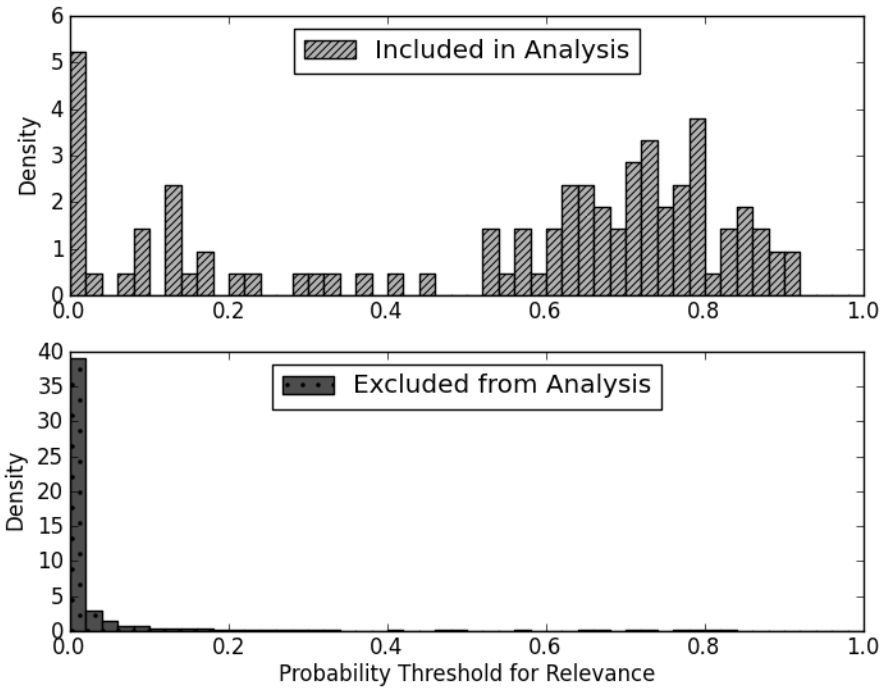
Table 6. Model performance for AEs

			GLMnet		GBM		Hybrid	
Study	**Phase**	**Threshold**	**Sensitivity**	**PPV**	**Sensitivity**	**PPV**	**Sensitivity**	**PPV**
AAP	**Original**	0.001	1	0.078	1	0.07	1	0.07
		0.01	1	0.168	1	0.138	1	0.118
		0.02	0.978	0.215	1	0.274	1	0.194
		0.1	0.901	0.392	0.934	0.436	0.956	0.385
	Update	0.001	1	0.033	1	0.029	1	0.029
		0.01	0.99	0.065	0.971	0.056	0.99	0.047
		0.02	0.981	0.09	0.895	0.11	0.981	0.078
		0.1	0.867	0.172	0.848	0.2	0.886	0.162
LBD	**Original**	0.001	1	0.065	1	0.073	1	0.057
		0.01	0.993	0.175	0.975	0.192	0.996	0.166
		0.02	0.964	0.21	0.971	0.229	0.978	0.203
		0.1	0.885	0.338	0.903	0.365	0.918	0.328
	Update	0.001	0.946	0.04	0.957	0.039	0.967	0.033
		0.01	0.739	0.097	0.674	0.098	0.739	0.09
		0.02	0.685	0.116	0.663	0.119	0.707	0.112
		0.1	0.511	0.179	0.478	0.191	0.522	0.167

GLMnet = Generalized Linear Models with Convex Penalties; GBM = gradient boosting machine; Hybrid = Maximum prediction from either GLMnet or GBM; PPV = Positive predictive value
Note: We calculated bootstrapped standard errors for the GLMnet estimates. In all cases, the standard errors were substantially smaller (<0.005) than the estimates for sensitivity or PPV.

Figure 8 shows these results graphically using a histogram of the prediction probabilities, divided according to whether the article met final inclusion criteria. Articles not considered for AE analyses were predominantly assigned probabilities very close to zero; included articles had probabilities that spanned the entire spectrum including the 2 percent that were assigned a probability of inclusion <0.02.

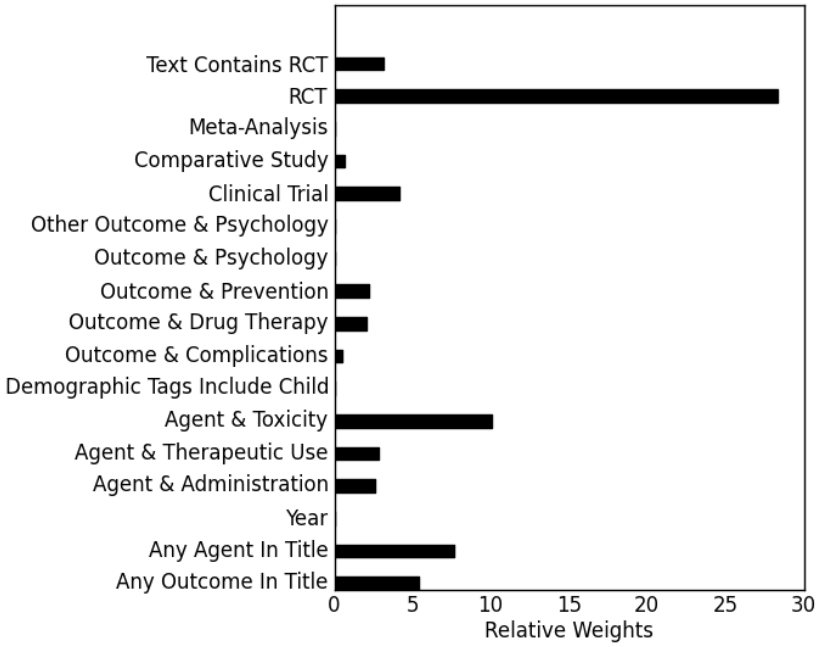
Figure 8. Histogram AAP AE analysis: distribution of predictions



Predicting AE-Relevant Articles for LBD Update

Figure 9 shows key variables for this analysis (GBM only, though weights for GLMnet were similar, in that RCT contained the greatest explanatory power). By inspection, these importance weights do not appear extremely dissimilar to those from the AAP analysis.

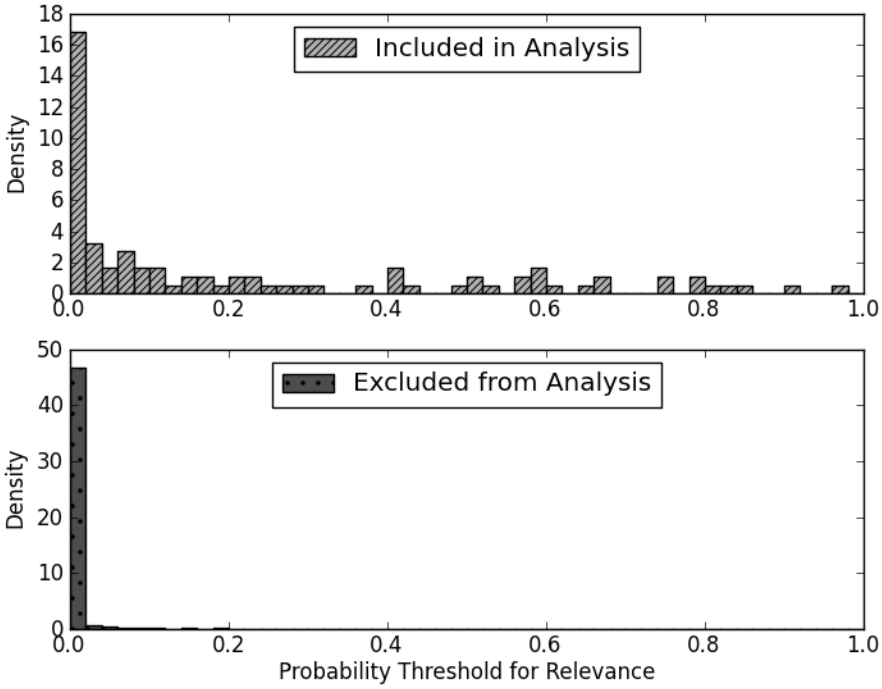
Figure 9. Relative weights for variables in LBD AE analysis



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 (Table 6.) However, we were able to predict AE-relevant articles 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. Our results for GBM-based and hybrid models were not substantially better at threshold $p \geq 0.02$, with the hybrid model achieving sensitivity of 0.707 and PPV of 0.112.

Figure 10 shows these results graphically as many AE articles relevant to the LBD update were assigned relatively low prediction probabilities. In fact, 11.6 percent of AE-relevant articles were assigned probabilities <0.005. When we examined missed AE articles, we noted that there were relatively few relevant large observational studies (cohort and case-control studies) in the original review. As a result, the both the GLMnet- and GBM-based models assigned lower probabilities to observational studies in the LBD update as well. However, observational studies were more important in the update because the SCEPC researchers focused on several newly identified AEs that were largely studied in cohort and case-control studies.

Figure 10. Histogram LBD AE analysis: distribution of predictions



Performance Predicting Any Relevant Result and Potential Workload Reductions

The workflow in many AHRQ comparative effectiveness reviews includes a first step in which reviewers select all articles that might be relevant to AEs or efficacy, and as the second step, a process that reviews the full text of articles to determine their relevance to efficacy/effectiveness or AE analyses. To simulate how our approach might improve the workflow for updates, we determined the GLMnet-based model's sensitivity and PPV at various thresholds for retrieving *all* AE and efficacy/effectiveness analyses. Sensitivity and PPV for a particular threshold were determined by selecting articles if the maximum predicted relevance from either model (efficacy/effectiveness or AE) exceeded the threshold. We show how sensitivity and the number needed to screen change as the threshold changes in Table 7. (We do not show sensitivities < 0.75 as these results are unlikely to be useful to comparative effectiveness review researchers.) 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 percent of screening saved. When we applied this threshold to the update predictions, the projected sensitivity model exceeded 0.99, whereas the proportion of title/abstract screening saved was 55.4 percent. In other words, the total number of articles to be screened would have been reduced from 3,591 to 1,601. By contrast, the hybrid model had identical sensitivity, but more limited workload reductions at the same threshold ($p \geq 0.01$).

Table 7. GLMnet model performance in retrieving any relevant article (AAP Update)

	Prediction Threshold	True Positives	False Negatives	Sensitivity	Total Screening Burden	Screening Saved (%)
Original	**0**	98	0	1	1307	0
	0.001	98	0	1	1169	10.6
	0.005	98	0	1	765	41.5
	0.01	98	0	1	547	58.1
	0.015	97	1	0.99	463	64.6
	0.02	96	2	0.98	415	68.2
	0.025	96	2	0.98	373	71.5
	0.05	92	6	0.939	284	78.3
	0.1	90	8	0.918	225	82.8
	0.2	87	11	0.888	176	86.5
	0.3	76	22	0.776	122	90.7
Update	**0**	116	0	1	3591	0
	0.001	116	0	1	3237	9.9
	0.005	115	1	0.991	2191	39
	0.01	115	1	0.991	1601	55.4
	0.015	114	2	0.983	1312	63.5
	0.02	113	3	0.974	1144	68.1
	0.025	112	4	0.966	1026	71.4
	0.05	106	10	0.914	737	79.5
	0.1	102	14	0.879	549	84.7
	0.2	95	21	0.819	452	87.4
	0.3	89	27	0.767	366	89.8
	0.4	88	28	0.759	308	91.4

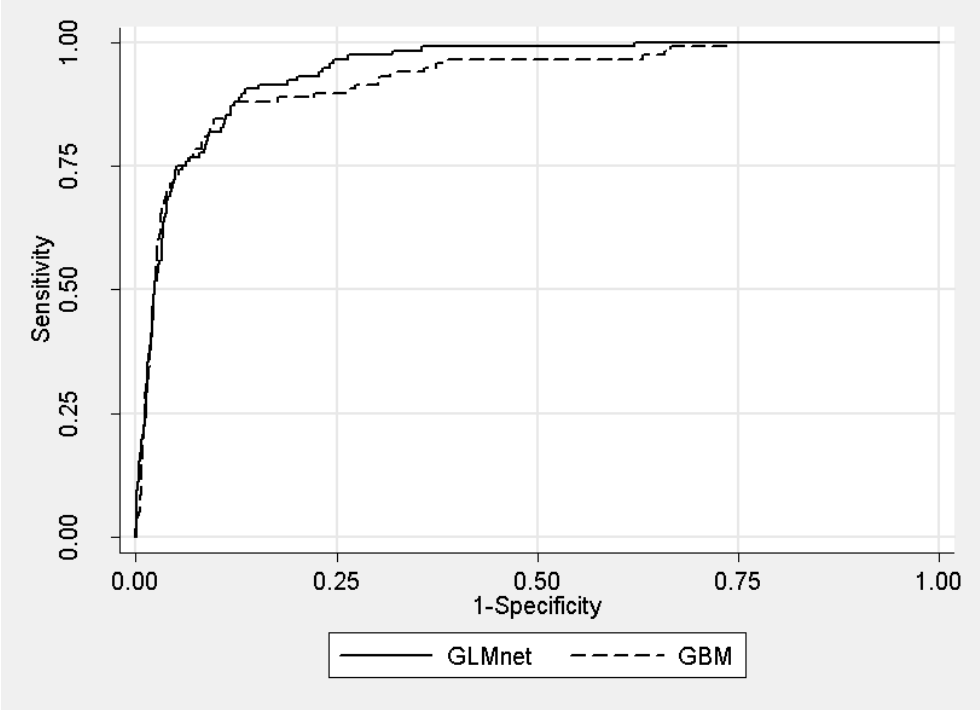
The GLMnet-based model for LBD performed worse, in that the model selected articles for the update with a sensitivity of 0.795 at a threshold of $p \geq 0.02$ (compared to 0.974 for AAP). (Tables 7 and 8.) However, this approach still provided potential benefits once we selected a suitable threshold. We chose a threshold of $p \geq 0.001$ based on the performance of the model in the original search results, in which a threshold of $p \geq 0.001$ yielded perfect sensitivity with 66.8 percent of screening saved. Using the same threshold when evaluating results in the update yielded perfect sensitivity accompanying the drop in the projected article screening burden from 7,051 to 2,597 (63.2%). While the probability thresholds differed between the AAP and LBD updates (0.001 in LBD and 0.01 in AAP), both thresholds could be derived from the original modeling process.

Table 8. GLMnet model performance in retrieving any relevant article (LBD update)

	Prediction Threshold	True Positives	False Negatives	Sensitivity	Total Screening Burden	Screening Saved (%)
Original	**0**	382	0	1	14700	0
	0.001	382	0	1	4880	66.8
	0.005	380	2	0.995	2346	84
	0.01	379	3	0.992	1836	87.5
	0.015	378	4	0.99	1615	89
	0.02	372	10	0.974	1477	90
	0.025	370	12	0.969	1369	90.7
	0.05	362	20	0.948	1098	92.5
	0.1	338	44	0.885	837	94.3
Update	**0**	127	0	1	7051	0
	0.001	127	0	1	2597	63.2
	0.005	117	10	0.921	1180	83.3
	0.01	107	20	0.843	882	87.5
	0.015	102	25	0.803	749	89.4

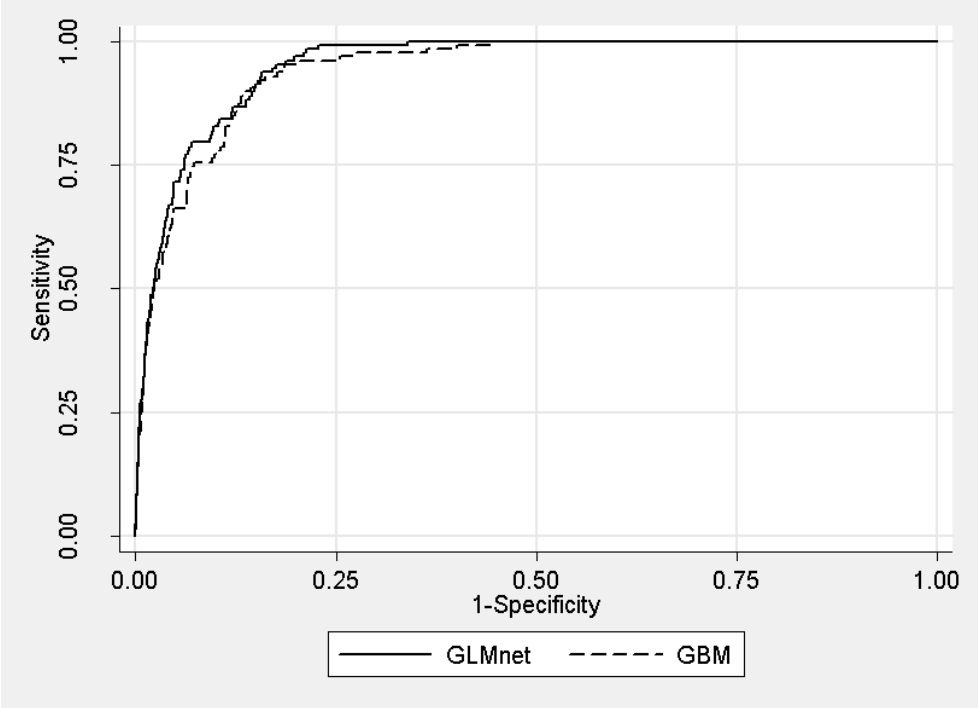
We show these results graphically using ROC curves (Figures 11 and 12). The AUC for the GLMnet method (in the AAP study) was 0.943 (95% CI: 0.927 to 0.960) versus 0.925 (95% CI: 0.899 to 0.950) with GBM. The p-value for null hypothesis of equality was 0.007. Similarly, the AUC for the GLMnet method (in the LBD study) was 0.954 (95% CI: 0.943 to 0.965) versus 0.947 (95% CI: 0.933 to 0.961) for GBM. In the LBD study, p-value for null hypothesis of equality was 0.06. Both results suggest that the ROC curves differed between the two studies; in addition, GLMnet seems to perform somewhat better than GBM visually as well. Still, it would be difficult to establish GLMnet's superiority in this context (comparative effectiveness reviewing updating) without further studies.

Figure 11. ROC curve for classifying AAP articles



AAP = antipsychotic systematic review; GBM = gradient boosting machine; GLMnet = generalized linear models with convex penalties; ROC = receiver operating characteristic

Figure 12. ROC curve for classifying LBD articles



GBM = gradient boosting machine; GLMnet = generalized linear models with convex penalties; LBD = low bone density systematic review; ROC = receiver operating characteristic

Evaluation of Model Prediction Errors

SCEPC researchers independently evaluated articles in the update that were included in the final reports but were assigned low probability scores by the statistical classifiers. We initially chose a probability threshold ($p \geq 0.02$) that reduced workload substantially; however, this threshold entailed 29 false negatives. Nearly all false negatives were non-RCT studies (along with an RCT that was not tagged as such by MEDLINE). Of the 29 false negatives (at threshold $p \geq 0.02$ from both updates), 26 were from the LBD update. The GLMnet model for LBD missed one RCT because the drug of interest ("raloxifene") was tagged with "pharmacology" and not a more revealing subheading. The remaining LBD false-negatives were non-RCT studies (including meta-analyses, case-control studies, retrospective analyses of claims databases, case-control studies, and analyses of government registries). It is difficult to determine whether similar studies were present in the original data without actually re-reading all earlier studies, but we did note that words such as "cohort" and "database" were poorly represented among both included and excluded studies in the original LBD report.

In considering the models used to predict inclusion of any relevant articles (Tables 7 and 8), just one article (from the AAP update) would have been excluded.[42] This article was likely assigned a low probability because it was tagged as a letter although it reported on a clinical trial. Of note, despite missing this trial using machine learning, EPC researchers might have been able to retrieve this trial 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.[43]

EPC researchers also evaluated several citations that were assigned high relevance probabilities but were deemed irrelevant by the original comparative effectiveness review researchers; none of these decisions changed on re-evaluation. These studies included one small RCT on calcitriol (that did not report fracture outcomes) and another RCT in a modest sized specialized population (Parkinson's patients).[44,45]

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 two comparative effectiveness reviews (LBD and AAP), and then modeled article relevance (i.e., inclusion in the final updated review) using several approaches based on the GBM and GLMnet statistical methods. In two simulated comparative effectiveness review updates, our approach achieved its best performance predicting relevance for efficacy/effectiveness articles; it performed worse when predicting articles relevant to the AE analysis for the LBD update. However, we estimated that these algorithms could reduce workload associated with screening updated search results for relevant efficacy/effectiveness and AE articles by more than 50 percent with minimal or no loss of relevant articles. Based on the slight differences in model performance between the GBM, GLMnet, and hybrid 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.

Evaluating Model Performance

Performance was similar when screening AAP citations for those relevant to efficacy/effectiveness and AE analyses. However, in the LBD analysis, we achieved substantially higher PPV for the same levels of sensitivity when predicting whether citations were relevant for the efficacy/effectiveness analyses as opposed to the AE analysis. Prior work has not focused heavily on AEs, so the benchmark is unclear here. However, we speculate that many of these false negatives (and the consequent poorer performance on the LBD study) can be attributed to the changed criteria for relevant AE citations. In the original LBD review, most of the articles relevant to AEs were RCTs because epidemiologic studies and retrospective database analyses are difficult to conduct prior to widespread use. Therefore, relevant citations in the original data set consisted (almost entirely) of RCTs; this would not have presented a problem if researchers only wanted RCTs in the update. However, the paucity of relevant non-RCT studies in the original data probably limited the ability of the model to efficiently retrieve relevant non-RCT studies. In addition, key included outcomes may be present in the full text, and yet not mentioned in the abstract or MeSH indexing terms. For example, in the LBD study, the key outcome was fracture prevention. However, often the articles mentioned just bone density in the abstract, while fractures were a secondary outcome described 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.

Our results concur with prior attempts at using machine learning to facilitate systematic review data collection; those studies used manually classified citations to predict inclusion in unclassified studies.[14,15,20,21] These efforts were met with substantial success, particularly an active learning model, which achieved 50 percent workload reductions and 100 percent sensitivity. [20] Previous studies used all indexing and text terms when employing statistical algorithms to classify documents.[14,20] The advantage of the prior approach is that little or no upfront investment is required outside of collecting an original data set. For a *de novo* search, removing upfront workload offers some advantages.

In contrast to prior studies, we adopted a more parsimonious approach that focused on a few key terms related to study design characteristics (publication type, demographic groups, and statistical design), intervention-specific characteristics, and outcome-specific characteristics. For many comparative effectiveness review updates, research librarians have already invested substantial time in creating optimal search strategies; we leveraged this effort using a prototype that parses previously created search strategies in a semi-automated fashion to locate key indexing terms. Furthermore, the vast majority of work was involved in creating the training data, which had already been completed. Therefore, the additional cost of making explanatory variables specific to each review was small when performing this simulated comparative effectiveness review update.

Furthermore, our algorithms explicitly dealt with updating, which afforded us far more initial training data than active learning models. However, our approach needed to surmount several new challenges because we needed to predict updated citations even though the literature was different, the reviewers changed, the search strategies changed, and (possibly) some of the underlying goals changed. Our approach achieved some success in combating data changes over time (known as concept drift in other applications).[21,27,28,30,31] Achieving similar levels of success suggests benefits to an approach incorporating domain-specific knowledge about key interventions and outcomes. However, 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.[15,31] In addition, this approach allows us to separate efficacy/effectiveness and AE analyses; although most comparative effectiveness reviews do not separate these analyses, independent filtering mechanisms may be of interest to other researchers.

Workload Reductions

For researchers seeking both AE and efficacy-relevant citations, we were able to remove approximately 50 percent of articles with loss of 1/116 articles for AAP and 0/127 articles of 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 systematic reviews and comparative effectiveness reviews aim for 100 percent sensitivity; despite the high sensitivity rates achieved, the loss of one article suggests that researchers will have to make some tradeoffs between sensitivity and efficiency as it will be difficult to guarantee 100 percent 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 missed reference might have been found by searching among references for included articles.

Our results also suggest possible improvements as well. The classifier's false negatives were more related to indexing variability than to model development. This observation suggests that capturing additional key variables might be more helpful than further statistical development. One method of doing so would be to use text features to improve capture of study design details, such as RCT design or meta-analysis. We used limited text features in generating predictions, but we anticipate that adding features from the entire text would be helpful, much as other machine learning document classification systems have done. In addition, the GLMnet model performed well with a limited number of training examples (1,307) for AAP suggesting that this method

could be implemented within an active learning framework,[15,20] and thus might be used to facilitate de novo reviews as well.

Implication for EPC Processes

The results we present (along with previous work in document classification) show that workload associated with updating could be substantially reduced if earlier classification decisions were used to reduce the workload involved in screening articles. We estimated that roughly 50 percent of title/abstract screening might be rendered unnecessary using a predictive model to reduce the screening burden. However, several outstanding issues need to be resolved prior to making these tools widely available.

First, the classifier relied on having complete data (database identifier, decision regarding relevance to efficacy/effectiveness analyses, decision regarding relevance to AE analysis). If such data were not fully compiled in the initial report, creating a machine learning model would be unlikely to be cost-effective as excessive effort would be required to format the data properly.

Second, although our statistical model relied on dozens of citation characteristics, it was very sensitive to MEDLINE's publication type field and MEDLINE indexing generally. NLM validates MEDLINE indexing against its own internal criteria and is responsive to re-indexing requests.[46] However, NLM's criteria did not match our criteria perfectly, which made model predictions less accurate. Several authors associated with the EPC group independently assessed false negatives (relevant citations that the model assigned a low probability of inclusion); typically, the low prediction probabilities for these included articles were due to problematic MEDLINE indexing of the publication type field. If such discrepancies could be accounted for, our other encouraging results suggest that this document classification prototype could be used to improve the efficiency of comparative effectiveness review updating. To that end, we are developing techniques for extracting information from the text to allow for greater consistency in determining the publication type (from our perspective) and other variables independent of MEDLINE indexing.

Third, predicting performance on update data using original data is imperfect thus far. While there was minimal performance declines in some cases (inclusion in any analysis), using identical thresholds in an update would have reduced sensitivity in other cases (inclusion in LBD effectiveness/efficacy analyses). Further testing on additional topics should allow us to provide researchers with better information regarding projected performance.

Finally, these systems currently work only with fully indexed PubMed citations. One mitigating factor is that the vast majority of relevant articles are located in PubMed. As described in greater detail below, we plan to generalize this model to articles lacking MEDLINE indexing by developing additional text analysis tools.

Future Research

1. Comparative effectiveness review methodologists will need to agree on a common data format and save all literature review decisions at the time of collection, as these data are much more easily accumulated over time than reconstructed later. At a minimum, the following elements are needed: data source (MEDLINE, EMBASE, PsycInfo, etc); source-specific identifier (e.g., PMID); study-specific identifier (e.g., LBD #1034); inclusion in final report for efficacy/effectiveness or AE analyses (or both); and title/abstract (if not in MEDLINE). Other information, e.g., inclusion after first stage screening and reason for exclusion from final study (if excluded) would be helpful as well.
2. The current model was built entirely upon MEDLINE classifications and is heavily reliant on their accuracy. Clearly, this characteristic would result in delaying the classification of newer articles. If reviews are being conducted every 2-3 years, this limitation would exclude only a small percentage of articles from the analysis (and leave them entirely for human review). However, if researchers wished to update reviews continuously (or monthly) and use citations from non-MEDLINE databases (such as EMBASE), absent or delayed MEDLINE indexing would render MEDLINE-only modeling inadequate. Further research on adding structured text characteristics to the statistical model would be helpful. Adding more (and presumably useful) features would improve accuracy as well. Other researchers have made extensive use of approaches based on "bag-of-words" when classifying documents for systematic reviews.[14,20] The underlying hypothesis in these studies is that term frequencies will differ between relevant and irrelevant documents. For example, a relevant document might be more likely to contain the phrases "randomized trial" or "RCT", whereas an irrelevant article might contain words such as "mouse" or "case-control." One can then use statistical algorithms (such as GBM or SVM) to model relevance as a function of these many text features.

   A modified approach using both text- and MeSH-derived features could be helpful. For example, one could classify citations that lack MEDLINE indexing by determining whether their text features are most similar to articles that are predicted to be highly irrelevant or to those that are predicted to be highly relevant (among MEDLINE-indexed articles). Using these shared text features, the MeSH indexing could be leveraged to provide additional information to articles lacking indexing.
3. We will need to test our models on additional systematic reviews, surgical interventions, and on non-therapeutic applications; in addition, we will also test other commonly-used algorithms such as SVM. Additional research is needed before a particular updating approach can be recommended for practical use.
4. We will examine whether training data can be used across systematic review topics, if the underlying inclusion criteria are similar enough. This experiment has been attempted before but has not been applied to true updating.[17] If this attempt is successful, we could vastly increase the volume of useful training data at our disposal.
5. We will need to streamline the process to make it production-ready and efficient. However, if classification decisions are readily available (see #1 above), the remainder of the process will not be labor-intensive; the key step will involve a clinical reviewer or research librarian spending 1–2 hours transforming the review's search strategy into

groups of terms (interventions, diagnostic tests, outcomes, etc.). We plan on developing a platform that would allow further data processing and modeling without human input. While predictions will still need to be evaluated by human reviewers, we plan on making this step time-neutral by providing samples of articles to be evaluated by comparative effectiveness review researchers as part of their normal workflow.

6. An active learning model could be adapted to perform in the updating context as well.[15,20] For example, one could generate predictions for updated data and sample predicted relevant articles in a stratified fashion – that is, all articles in the updated search predicted to be highly relevant and a sample of indeterminate and lower-ranked citations. The model could then be re-run using these new training data to generate a new model. This effort would offer two advantages: (a) Newer models could account for changes in the literature; and (b) less reviewer time would be wasted because many of the reviewed articles would likely be relevant and require review.
7. Finally, we identified a small false negative rate associated with our approach. Using the references of included reports could reduce the false negative rate by identifying missed reports.

Conclusions

In this pilot study, we created a prototype system that classified PubMed literature search results from two 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. Additional research refining this system, expanding its scope, and comparing it to other methods could allow researchers to select optimal machine learning methods for updating their reviews frequently and efficiently.

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, et al. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. 2007 Aug 21;147(4):224-33. PMID 17638714.

3. Garritty C, Tsertsvadze A, Tricco AC, et al. Updating systematic reviews: an international survey. PloS one. 2010;5(4):e9914. PMID 20376338.

4. Moher D, Tsertsvadze A, Tricco AC, et al. When and how to update systematic reviews. Cochrane database of systematic reviews. 2008(1):MR000023. PMID 18254126.

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 Jul 15;280(3):278-80. PMID 9676681.

6. Moher D, Tetzlaff J, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews. PLoS Med. 2007 Mar 27;4(3):e78. PMID 17388659.

7. 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. . Comparative Effectiveness Review. Rockville, MD: 2007. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

8. 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 Feb 5;148(3):197-213. PMID 18087050.

9. Shekelle P, Maglione M, Bagley S, et al. Efficacy and Comparative Effectiveness of Off-Label Use of Atypical Antipsychotics. Efficacy and Comparative Effectiveness of Off-Label Use of Atypical Antipsychotics. Rockville (MD); 2007.

10. 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 Sep 28;306(12):1359-69. PMID 21954480.

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. Sampson M, Shojania KG, Garritty C, et al. Systematic reviews can be produced and published faster. J Clin Epidemiol. 2008 Jun;61(6):531-6. PMID 18471656.

13. Sampson M, Shojania KG, McGowan J, et al. Surveillance search techniques identified the need to update systematic reviews. J Clin Epidemiol. 2008 Aug;61(8):755-62. PMID 18586179.

14. Cohen AM, Hersh WR, Peterson K, et al. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc. 2006 Mar-Apr;13(2):206-19. PMID 16357352.

15. Wallace BC, Small K, Brodley CE, et al. Active learning for biomedical citation screening. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining; 2011 Washington, DC, USA. ACM.

16. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, et al. Text categorization models for high-quality article retrieval in internal medicine. Journal of the American Medical Informatics Association : JAMIA. 2005 Mar-Apr;12(2):207-16. PMID 15561789.

17. Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):690-704. PMID 19567792.

18. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform. 2005 January 1, 2005;6(1):57-71.

19. Kilicoglu H, Demner-Fushman D, Rindflesch TC, et al. Towards automatic recognition of scientifically rigorous clinical research evidence. Journal of the American Medical Informatics Association : JAMIA. 2009 Jan-Feb;16(1):25-31. PMID 18952929.

20. Wallace BC, Trikalinos TA, Lau J, et al. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 2010;11:55. PMID 20102628.

21. Cohen AM, Ambert K, McDonagh M. A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. 2010;2010:121-5. PMID 21346953.

22. Cohen AM. Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. Journal of the American Medical Informatics Association : JAMIA. 2011 Jan-Feb;18(1):104; author reply -5. PMID 21169622.

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

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

25. Tibshirani R. Regression Shrinkage and Selection by Lasso. Journal of the Royal Statistical Society, Series B. 1996;58:267-88.

26. Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. Journal of the American Medical Informatics Association : JAMIA. 2011 Jun 2PMID 21546507.

27. Widmer G, Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts. Machine Learning. 1996;23(1):69-101.

28. Tsymbal A, Pechenizkiy M, Cunningham P, et al. Dynamic integration of classifiers for handling concept drift. Information Fusion. 2008;9(1):56-68.

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

30. Pan SJ, Yang Q. A Survey on Transfer Learning. 2010;22(10):15.

31. . Frustratingly Easy Domain Adaptation. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics; 2007; Prague, Czech Republic. Association for Computational Linguistics.

32. Cock PJ, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422-3. PMID 19304878.

33. 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 and rheumatism. 2009 Nov;60(11):3346-55. PMID 19877063.

34. U.S. National Library of Medicine. MeSH Browser. U.S. National Institutes of Health; 2010.

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

36. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001;29(5):1189-232.

37. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software. 2010;33(1):1-22. PMID 20808728.

38. Deerwester S, Dumais S, Furnas G, et al. Indexing by Latent Samentic Analysis. Journal of the American Society for Information Science. 1990;41:391-407.

39. Deerwester S, Dumais S, Landauer T, et al. Improving information-retrieval with latent semantic indexing. Proceedings of the AS1S annual meeting. 1988;25:36-40.

40. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837-45. PMID 3203132.

41. Efron B, Tibshirani R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science. 1986;1(1):54-75.

42. Tsuang J, Marder SR, Han A, et al. Olanzapine treatment for patients with schizophrenia and cocaine abuse. The Journal of clinical psychiatry. 2002 Dec;63(12):1180 -1. PMID 12530415.

43. Hamilton JD, Nguyen QX, Gerber RM, et al. Olanzapine in cocaine dependence: a double-blind, placebo-controlled trial. The American journal on addictions / American Academy of Psychiatrists in Alcoholism and Addictions. 2009 Jan-Feb;18(1):48-52. PMID 19219665.

44. Sato Y, Manabe S, Kuno H, et al. Amelioration of osteopenia and hypovitaminosis D by 1alpha-hydroxyvitamin D3 in elderly patients with Parkinson's disease. Journal of neurology, neurosurgery, and psychiatry. 1999 Jan;66(1):64-8. PMID 9886454.

45. Ebeling PR, Wark JD, Yeung S, et al. Effects of calcitriol or calcium on bone mineral density, bone turnover, and fractures in men with primary osteoporosis: a two-year randomized, double blind, double placebo study. The Journal of clinical endocrinology and metabolism. 2001 Sep;86(9):4098-103. PMID 11549632.

46. Glanville JM, Lefebvre C, Miles JN, et al. How to identify randomized controlled trials in MEDLINE: ten years on. J Med Libr Assoc. 2006 Apr;94(2):130-6. PMID 16636704.

Abbreviations

AAP	Atypical antipsychotic drug
AE	Adverse effect
AHRQ	Agency for Healthcare Research and Quality
AUC	Area under the receiver operating curve
EPC	Evidence-based Practice Center
FDA	U.S. Food and Drug Administration
GBM	Gradient boosting machine
GLMnet	Generalized linear models with convex penalties
LBD	Low bone density
MeSH	Medical subject heading
PPV	Positive predictive value
ROC	Receiver operating characteristic
SCEPC	Southern California Evidence-based Practice Center
SVM	Support vector machines
UMLS	United Medical Language System