# Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine

RECEIVED 22 April 2014
REVISED 5 November 2014
ACCEPTED 15 November 2014
PUBLISHED ONLINE FIRST 5 February 2015

Aaron M Cohen[1], Neil R Smalheiser[2], Marian S McDonagh[1], Clement Yu[3], Clive E Adams[4], John M Davis[2], Philip S Yu[3]

## ABSTRACT

**Objective:** For many literature review tasks, including systematic review (SR) and other aspects of evidence-based medicine, it is important to know whether an article describes a randomized controlled trial (RCT). Current manual annotation is not complete or flexible enough for the SR process. In this work, highly accurate machine learning predictive models were built that include confidence predictions of whether an article is an RCT.

**Materials and Methods:** The LibSVM classifier was used with forward selection of potential feature sets on a large human-related subset of MEDLINE to create a classification model requiring only the citation, abstract, and MeSH terms for each article.

**Results:** The model achieved an area under the receiver operating characteristic curve of 0.973 and mean squared error of 0.013 on the held out year 2011 data. Accurate confidence estimates were confirmed on a manually reviewed set of test articles. A second model not requiring MeSH terms was also created, and performs almost as well.

**Discussion:** Both models accurately rank and predict article RCT confidence. Using the model and the manually reviewed samples, it is estimated that about 8000 (3%) additional RCTs can be identified in MEDLINE, and that 5% of articles tagged as RCTs in Medline may not be identified.

**Conclusion:** Retagging human-related studies with a continuously valued RCT confidence is potentially more useful for article ranking and review than a simple yes/no prediction. The automated RCT tagging tool should offer significant savings of time and effort during the process of writing SRs, and is a key component of a multistep text mining pipeline that we are building to streamline SR workflow. In addition, the model may be useful for identifying errors in MEDLINE publication types. The RCT confidence predictions described here have been made available to users as a web service with a user query form front end at: http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/RCT_Tagger.cgi.

**Key words:** Support Vector Machines, Natural Language Processing, Randomized Controlled Trials as Topic, Evidence-Based Medicine, Systematic Reviews, Information Retrieval

## BACKGROUND AND SIGNIFICANCE

For many biomedical literature search tasks, it is important to know whether articles provide primary data about randomized controlled trials (RCTs). RCTs are considered the highest form of primary evidence for Evidence-based Medicine (EBM), and are the current foundation on which much of medical knowledge is based.[1,2] When RCT evidence exists in the form of well-conducted studies, the articles describing the RCT studies have a strong influence on medical knowledge summarization, guidelines, and practice. In particular, the process of creating systematic reviews (SRs), which seeks to objectively evaluate and summarize the current state of knowledge about a specific medical question, relies heavily on the identification and content of published RCTs.

Unfortunately, identifying all the published RCTs in a given area is far from simple. While bibliographic databases such as MEDLINE include an article publication type *Randomized Controlled Trial (RCT_PT)*, the annotation is not applied with 100% accuracy or coverage. Studies have found that only about 85% of articles in MEDLINE considered RCTs for the purpose of SR are actually annotated with the *RCT_PT*.[3,4] As a result, we have found in our research that many SR groups do not use the publication type filter in their search criteria

Correspondence to Aaron M. Cohen, MD MS, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, Phone: 1-503-494-0046, Fax: 1-503-346-6815, cohenaa@ohsu.edu



For numbered affiliations see end of article.

because they aim to identify all RCTs relevant to the subject of interest, and missing ~15% of articles based on publication type filtering is not acceptable. Instead SR groups review a much larger set of articles manually, resulting in a huge amount of additional work. In a prior study we found that the single largest reason for excluding an article from an SR performed by the Cochrane Collaboration was because the review required RCTs and the articles retrieved by the literature search were screened manually and determined to not be RCTs.[4]

We are currently conducting research on building a text mining-based pipeline to aid in the creation of SRs.[5] A major goal of the pipeline is to reduce the workload of systematic reviewers by reducing the number of articles they assess that are eventually excluded from the final SR. This represents essentially wasted effort, and the SR could use the time saved for much better purposes, such as performing deeper meta-analysis, writing and publishing more reports, and conducting additional SRs. There are also efforts underway to produce quicker, smaller reviews, called "Rapid Reviews" that will require faster ways to get at the most important publications.[6,7]

Since excluding an article from an SR because it is not an RCT is the most common exclusion reason in many reviews, it is clear that a highly accurate and complete confidence assessment of whether all available topical articles are or are not RCTs is an essential component of our text mining pipeline. This knowledge would be best applied after the initial literature search and before the articles are manually reviewed. A probabilistic measure instead of a binary annotation would be most appropriate. Articles could then be ranked for "RCT-ness" based on this measure. The ranking could then be used for several purposes, such as article filtering, work prioritization, and more accurate estimation of the size of the relevant literature.[8–10]

## OBJECTIVE

The objective is to build highly accurate machine learning predictive models that can be used to access whether or not an article is an RCT. The models will provide estimates of the probability that a given article is an RCT. These models must be applicable to both MEDLINE annotated articles as well as unannotated article citation records. These confidence predictions can then be used to *retag* MEDLINE articles, providing more information about whether an article is an RCT compared to binary publication type annotations.

## MATERIALS AND METHODS

There is no single token feature or simple obvious model that can be used to identify RCTs with high performance. Preliminary research found that the best single feature is the term "randomized," which has an F1 of ~0.72, with the performance of other top-ranked single features dropping precipitously (see the Supplementary material). Furthermore, single binary tokens do not provide a way to rank articles.

Instead, we chose to use a data-driven machine learning approach. We applied machine learning techniques to build RCT confidence prediction models using a large human-related subset of PubMed. Model feature selection was performed using forward selection of large sets of related features. These models were then tested using held-out data, and compared to other baseline approaches to evaluate the results of the specific techniques chosen. We also performed a manual review of a random selection of article predictions.

### Data set creation

A dataset intended to include all human-related article citations published between 1987 and 2012 was created by querying PubMed and downloading the publication record in XML format. This was split into a training set, and year 2011 and 2012 evaluation test sets. Full details are given in the Supplementary material.

### Gold standard

In this work, the presence or absence of the MEDLINE *RCT_PT* was used as a gold standard for training and for large-scale evaluation. While, as discussed above, these annotations are not 100% correct, taken at the scale of the millions of articles in MEDLINE, the *RCT_PT* annotations can be assumed to be correct with the addition of a small amount of noise. We used the MEDLINE *RCT_PT* annotation as the dependent variable for training our models, providing millions of samples on which to perform machine learning and large-scale evaluation across all human-related articles in MEDLINE.

Because the MEDLINE RCT_PT annotations do contain a small percentage of errors, a manual review of a random selection of articles was also performed. This was used to estimate the error rate in MEDLINE and also compared to the model predicted confidence. The manual review was based on the MEDLINE definition of an RCT (http://www.ncbi.nlm.nih.gov/mesh/68016449):

> *Work consisting of a clinical trial that involves at least one test treatment and one control treatment, concurrent enrollment and follow-up of the test- and control-treated groups, and in which the treatments to be administered are selected by a random process, such as the use of a random-numbers table.*

In the course of performing the manual review, it was found that the MEDLINE definition was not clear enough to assign ambiguous cases such as published studies that include both observational and randomized trials, or high level overview papers of large multicenter trials. Therefore, we constructed an extended RCT definition to give annotators additional guidance. Inter-rater agreement Cohen's $\kappa$ between the manual review and the MEDLINE *RCT_PT* was 0.72, which is considered substantial agreement. This is comparable to the best performance obtained in other inter-rater agreement students on MEDLINE annotation.[11] Additional detail is provided in the Supplementary material.

### Data preprocessing

The dataset was preprocessed to create a large number of feature types for each article. Only information available in the MEDLINE record was used to create feature types for machine learning. There may be one, several, or millions of features for

each type. The group of all features of a given type is termed a *feature set*. Feature types fit into three main categories: a) textual features extracted from the title and abstract, such as title n-grams of length 1–3; b) features extracted from the bibliographic fields of the citation, such as journal and author names; and c) features derived from annotated MEDLINE fields such as MeSH terms and publication types. In total approximately 45 million unique binary and numeric features were generated.

Note that for training, and some of the evaluations, the presence/absence of the MEDLINE publication type *RCT* was used as the gold standard to classify a positive RCT/negative RCT article. However, it is emphasized that the MEDLINE publication type *RCT* was not used as an input feature in any of our machine learning models, which in fact did not employ publication type information at all (see below).

As part of the feature set preprocessing, the statistical significance of every individual feature in the training set was evaluated using the chi-squared test (CHISQ), across all feature types for predicting whether an article was an RCT (details are given in the Supplementary material). This collection of features will be referred to as the *statistically significant feature* set. Predictive machine learning models built with this feature set will be compared with models built without first filtering the features by significance testing.

### Machine learning approach

For greater utility in using the results in real-world information retrieval applications, it was highly desirable that the machine learning model produce an accurately calibrated confidence estimate of whether each article was or was not an RCT. For example, over a large sample of articles predicted to be RCTs at 0.60 confidence, close to 60% of those articles on close examination should actually be RCTs. Resource constraints included reasonable time performance and fitting into the 64 GB memory space on the Linux CentOS release 6.4 computer that was available for building the models.

### Algorithm

After some initial experimentation with several machine learning approaches and implementations including the Weka (http://www.cs.waikato.ac.nz/ml/weka/) implementation of logistic regression, decision trees, K-nearest neighbors, and the SVMLight (http://svmlight.joachims.org/)[12] implementation of support vector machines (SVMs), the Liblinear[13] fast linear implementation of SVM was selected. This implementation of SVM is specifically optimized for solving binary linear SVM problems on large, sparse data and is much faster and requires less memory than more general SVM implementations. It was the only algorithm that could handle the large datasets (up to available memory) and converge to a solution within a reasonable amount of time on our equipment, which was operationally defined as a maximum of 24 h/run.

### Confidence prediction

While the Liblinear SVM implementation can generate predictive models from the large datasets, it does not produce a confidence prediction along with the classification of each document as RCT positive or negative. However, confidence predictions based on the signed margin distance that Liblinear provides were generated using an extension to the method of Rüping.[14] The signed-margin distance is the distance of a sample in feature space in front or behind the optimal separating hyperplane determined by the SVM. Applying Rüping's method converts this distance into a probabilistic confidence prediction. This method and our extensions are described in detail in the Supplementary material.

### Training data subsampling

Because of computer memory limitations, it was not practical to train the models on the entire 1987–2010 training set of over 5 million articles. Before beginning the model selection process, an initial study was conducted to determine how much of the training data was actually required to get maximum classifier performance. A $5 \times 2$ cross-validation was performed on the training data using all of the title and abstract derived features in the database and the article corpus was sampled at a number of increments representing training data fractions between 0.50% and 10.0% of the full training data set. Examining these initial results, it was clear that classifier performance is basically flat between 5% and 10% sampling of the training data (figure shown in the Supplementary material). To err on the side of using more than enough data, 7.5% sampling was selected for cross-validation and final model training for all of the feature set forward selection (see below) and final model building runs.

### Feature modeling

Most features were treated as binary in our model. This decision was made based on our prior experience with literature classification and the need for computational efficiency. In our past work, using the feature scalar instead of the binary value provided little if any gain.[8,15] Features such as MeSH terms are binary by nature, and, in short text samples such as titles and abstracts, important predictive features most often do not occur more than once. N-grams in the title and abstract were treated separately.

There is wide range of author counts in our dataset and for this feature it could be worth using the full scalar value. It was hypothesized that the relationship between being an RCT and the number of authors would be neither monotonic nor linear, and that various ranges of author count could have widely different predictive value. A method called recursive partitioning was used to split these count-based features into ranges and treat each range as a separate binary feature. Recursive partitioning (RP) uses a minimum description length approach.[16] Preliminary experiments using cross-validation on the 2010 training data showed that the RP approach resulted in slightly improved performance compared to simple author count normalization. We used RP to model the author count feature in all our subsequent experiments. Full details are provided in the Supplementary material.

### Feature set selection

With hundreds of millions of potentially predictive features, it was not feasible to select model features one feature at a time.

Instead the models were built using a forward selection process on each feature set. Starting with no feature sets in the model, the addition of each feature set not yet included in the model is tested for improvement using cross-validation and evaluation of area under the receiver operating characteristic curve (AUC) and F1 measures. The feature set that improves the model the most is added to the current model and the process repeated until no remaining feature sets improve the model. This is described in detail in the Supplementary material.

After the feature set selection step was complete, the final predictive model was trained using the chosen feature sets and the complete (subsampled) training set. In this way two final predictive models using forward selection were built, the main model that used citation data plus annotated metadata available in MEDLINE such as MeSH terms (the CITATION_PLUS_MESH model), and another model using only the features available in the publication citation (the CITATION_ONLY model). Each model was optimized for a different purpose. The CITATION_PLUS_MESH model was created to automatically annotate with RCT confidence all articles in MEDLINE that had been indexed and annotated by the National Library of Medicine. The CITATION_ONLY model is more flexible in that it can be applied to any article from a journal indexed in MEDLINE for which citation information (title, abstract, authors, journal, etc.) is available. The CITATION_ONLY model can also be applied to articles in the queue waiting for MEDLINE indexing. Together the CITATION_PLUS_MESH and CITATION_ONLY models are referred to as the FORWARD SELECTION models.

Several other models were built for comparing performance with our two main models. The statistically significant feature set was intersected with the features available using the same citation only and citation plus MEDLINE criteria used with the FORWARD SELECTION models to build two predictive models restricted to individually statistically significant features with these feature sets. These models were also trained with the 7.5% subsampled training data and are referred to as the CHISQ FEATURE FILTERING models. To test whether inclusion of Publication Type information would contribute to performance, models that included our final collection of feature sets plus the MEDLINE publication types (minus the *RCT_PT* of course) were evaluated. These are referred to as the FORWARD SELECTION PLUS PUBTYPES models. Lastly, it is important to show the relative impact of limiting predictive models to statistically significant features as opposed to simply limiting the number of features, and; therefore, citation only and citation plus MEDLINE models were created with the same number of features as the CHISQ FEATURE FILTERING models created using CHISQ statistically significant features. Features were sorted based on the absolute value of the SVM coefficient in the FORWARD SELECTION models and only the top N features were kept, where *N* was the number of features in the CHISQ filtered model. These are termed the POST HOC DROP LOW WEIGHTS models.

### Investigations to optimize the machine learning models

#### Model feature set selection and cross-validation

Our forward selection process found a collection of the seven best performing feature sets for the CITATION_ONLY model and ten feature sets in the CITATION_PLUS_MESH model. The selected feature sets are given in the Supplementary material. Final cross-validation AUC, average precision, F1, and mean squared error (MSE) on the training set are shown for the CITATION_PLUS_MESH model in Table 1, and for the CITATION_ONLY model in Table 2. Overall, both models performed with high accuracy achieving AUC's of approximately 0.97 with the CITATION_PLUS_MESH model being slightly better across all measures. The difference in AUC between the models is statistically significant at $\alpha = 0.05$. The AUC 95%

Table 1: Performance of the main classifier using the CITATION_PLUS_MESH_MODEL, showing 5x 2-way cross-validation performance on the entire training dataset, as well as performance on the held-out testing sets corresponding to human-related articles published in the years 2011 and 2012

| CITATION_PLUS_MESH_MODEL | | | |
|---|---|---|---|
| | DATASET | | |
| *MEASURE* | 1987–2010 cross-validation | 2011 | 2012 |
| AUC | 0.976 | 0.973 (0.9714, 0.9746) | 0.972 (0.9704, 0.9736) |
| AVERAGE_PRECISION | 0.877 | 0.873 | 0.870 |
| F1 | 0.820 | 0.822 | 0.824 |
| ACCURACY | 0.985 | 0.985 | 0.985 |
| MSE | */0.048 | 0.013/0.045 | 0.013/0.044 |

AUC = area under the receiver operating characteristic curve, AVERAGE_PRECISION = average precision at RCT positive rankings, F1 = balanced F-measure, harmonic mean of precision and recall, MSE = mean squared error of the confidence predictions with/without use of the enhanced Rüping confidence estimation method. The confidence estimation method was not used for the cross-validation model selection runs because of its increased run-time. The 95% confidence intervals for AUC on the 2011 and 2012 data sets are shown in parentheses next to these values.

Table 2: Performance of the final classifier using the CITATION_ONLY_MODEL, showing 5× 2-way cross-validation performance on the entire training dataset, as well as performance on the held-out testing sets corresponding to human-related articles published in the years 2011 and 2012

| CITATION_ONLY_MODEL | | | |
|---|---|---|---|
| | *DATASET* | | |
| *MEASURE* | 1987–2010 cross-validation | 2011 | 2012 |
| AUC | 0.969 | 0.966 (0.9642, 0.9678) | 0.965 (0.9632, 0.9668) |
| AVERAGE_PRECISION | 0.855 | 0.854 | 0.852 |
| F1 | 0.800 | 0.807 | 0.811 |
| ACCURACY | 0.984 | 0.984 | 0.984 |
| MSE | */0.052 | 0.014/0.048 | 0.014/0.048 |

AUC = area under the receiver operating characteristic curve, AVERAGE_PRECISION = average precision at RCT positive rankings, F1 = balanced F-measure, harmonic mean of precision and recall, MSE = mean squared error of the confidence predictions with/without use of the enhanced Rüping confidence estimation method. The confidence estimation method was not used for the cross-validation model selection runs because of its increased run-time. The 95% confidence intervals for AUC on the 2011 and 2012 data sets are shown in parentheses next to these values.

confidence intervals computed using the conservative Hanley/McNeil method do not overlap. These intervals are shown in parenthesis next to the AUC values in Tables 1 and 2.

Of particular note are the low values of MSE on the confidence predictions, showing the high accuracy of the enhanced Rüping method. On average, the squared error of the confidence predictions of both models is less than 0.015. Tables 1 and 2 also compare the MSE achieved by the enhanced Rüping method with a baseline method of simply mapping the signed margin value to a zero-centered sigmoid curve. The enhanced Rüping method is $>3\times$ as accurate than the simple sigmoid method.

### Comparison of final models with alternate feature selection methods

Table 3 compares the performance of our final models with that of three other models incorporating differing feature selection methods on the held-out 2011 data subset. These include the statistically significant feature set (CHISQ FEATURE FILTERING), adding MEDLINE publication type features to our final models (FORWARD SELECTION PLUS PUBTYPES), and dramatically reducing the feature set size of our final models by dropping low weight features (POST HOC DROP LOW WEIGHTS models).

The forward selection process greatly outperforms the CHISQ statistically significant feature selection method across all metrics. This is not simply due to the greater number of features available to the SVM machine learning method. The column POSTHOC DROP LOW WEIGHTS shows the performance of the forward selection method after limiting the final features to approximately the same number as the CHISQ method. This performance is almost identical to that of the forward selection method, with a factor of 300–400× fewer features.

The table also shows the results of adding MEDLINE publication type features to the final forward selection models. Performance on AUC increases a little, while average precision drops much more. The F1 and MSE also worsen. Overall, adding MEDLINE publication types to the final models decreases performance, validating our initial decision not to include these features in our model building process.

### Evaluation methods

A combination of large-scale gold standard based metric measures and small-scale manual review was used to evaluate the machine learning models and the resulting predictions. As noted above, using the MEDLINE *RCT_PT* as a gold standard, AUC and mean precision were used as the decision metrics for forward selection. The F1 was used as a measure of binary classification performance, and the MSE (mean squared error) as a measure of the accuracy of the predicted confidence value.

To evaluate the performance of the models as they would be used for a SR, a manual review of a subset of the predictions was also performed. For this, the topic-based search terms from four Cochrane reviews[17–20] were used to perform PubMed searches and collect articles. These topics were randomly chosen from a set that our group has used in previous work[4] that included PubMed search terms. The highly sensitive controlled trial search criteria of Dickersin *et al.*[21] were added to these terms. The topic search queries are given in Supplementary Table S2. Articles that did not have the MEDLINE publication type *RCT* assigned to them were specifically selected, as these are of the greatest interest, and then the CITATION_PLUS_MESH predictive model was run on these articles. The predicted confidence values were grouped into 0.10 wide intervals between 0.0 and 1.0, and 20 articles from each interval were randomly chosen for manual review.

Table 3: Performance comparisons between several alternate modeling approaches and the final classifier models

| | MODELING APPROACH | | | |
|---|---|---|---|---|
| MODEL | FORWARD SELECTION | CHISQ FEATURE FILTERING | POSTHOC DROP LOW WEIGHTS | FORWARD SELECTION PLUS PUBTYPES |
| AUC | | | | |
| CITATION_ONLY_MODEL | 0.966 | 0.948 | 0.967 | 0.969 |
| CITATION_PLUS_MESH_MODEL | 0.973 | 0.959 | 0.973 | 0.974 |
| AVERAGE_PRECISION | | | | |
| CITATION_ONLY_MODEL | 0.854 | 0.781 | 0.853 | 0.840 |
| CITATION_PLUS_MESH_MODEL | 0.873 | 0.826 | 0.873 | 0.856 |
| F1 | | | | |
| CITATION_ONLY_MODEL | 0.807 | 0.727 | 0.808 | 0.778 |
| CITATION_PLUS_MESH_MODEL | 0.822 | 0.771 | 0.823 | 0.794 |
| MSE | | | | |
| CITATION_ONLY_MODEL | 0.014 | 0.020 | 0.014 | 0.015 |
| CITATION_PLUS_MESH_MODEL | 0.013 | 0.016 | 0.013 | 0.014 |
| NUMBER OF FEATURES | | | | |
| CITATION_ONLY_MODEL | 44,114,421 | 102,023 | 102,262 | 44,114,572 |
| CITATION_PLUS_MESH_MODEL | 34,636,788 | 113,177 | 110,982 | 34,636,939 |

AUC = area under the receiver operating characteristic curve, AVERAGE_PRECISION = average precision at RCT positive rankings, F1 = balanced F-measure, harmonic mean of precision and recall, MSE = mean squared error of the confidence predictions.

The annotator was provided with only the PubMed identifiers for these 200 articles in randomized order and asked to assign to one of three categories: RCT, UNCERTAIN, or NOT_RCT.

To evaluate the possibility of errors in the assignment of the *RCT_PT* in MEDLINE, a random subset of 50 previously unseen articles assigned the RCT_PT published in the year 2014 were also manually reviewed. The CITATION_PLUS_MESH confidences were also computed for these articles and analyzed.

## RESULTS

### Performance of the optimized models in identifying RCTs

The high performance of both models under cross-validation was validated by the performance on the 2011 and 2012 held-out data sets. As shown in Tables 1 and 2, for both years, the AUC ($\geq$0.965) is extremely high, as is accuracy ($\geq$0.984) and *F*-score ($\geq$0.807). Overall performance is very nearly the same as cross-validation across all measures for both models on both datasets, demonstrating that using the cross-validation procedure resulted in accurate performance estimations for model selection.

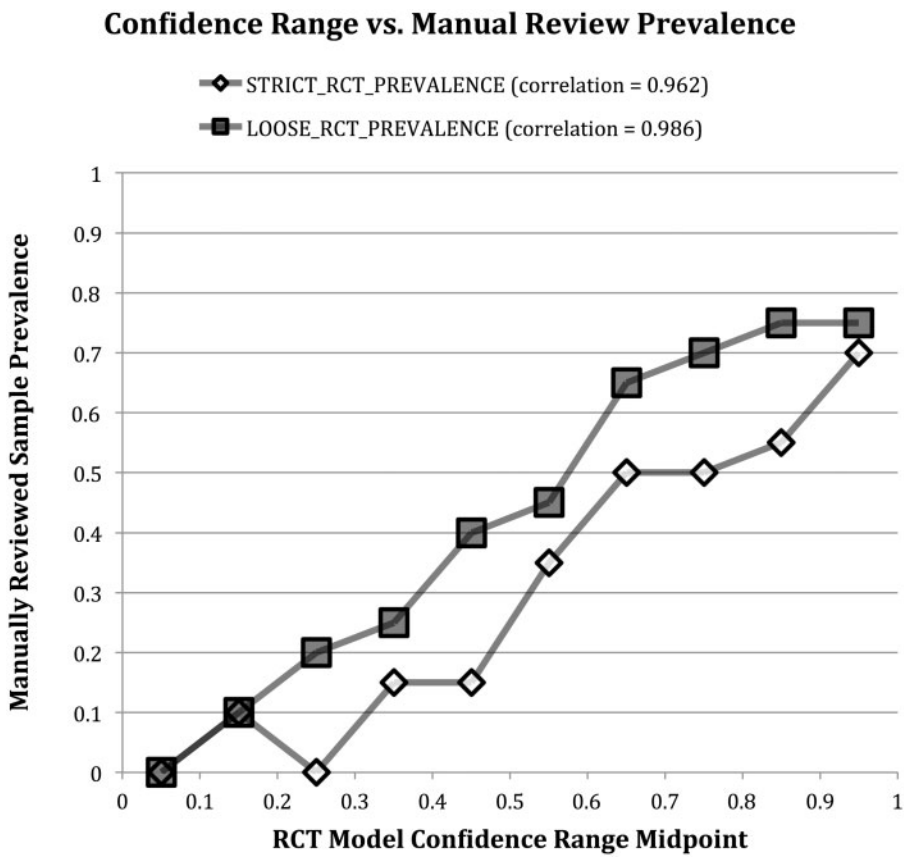
The differences in AUC performance between the forward selection CITATION_PLUS_MESH and CITATION_ONLY models are small but statistically significant for both 2011 and 2012 held-out data sets. The AUC 95% confidence intervals computed using the conservative Hanley/McNeil method do not overlap. The 95% confidence intervals are shown in parenthesis next to the AUC values in Tables 1 and 2.

For further analysis, the Supplementary material also includes tables showing the most strongly weighted positive and negative features in both models, and a distribution of the confidence predictions for MEDLINE articles indexed by publication type as either RCTs or non-RCTs in the held-out 2011 dataset.

### Analysis of cases in which the model disagreed with the MEDLINE RCT_PT

Figure 1 shows the result of our manual evaluation on the random sample of articles not assigned the MEDLINE *RCT_PT* from the four SR topics. The graph shows two lines. The first STRICT_RCT_PREVALENCE treats samples judged UNCERTAIN as NOT_RCT. The second line LOOSE_RCT_PREVALENCE, treats UNCERTAIN samples as RCT. Both lines show high correlation between the mid point of the prediction confidence range and the RCT prevalence of samples with confidence predictions fitting in that range. Both lines are essentially monotonic.

**Figure 1:** This graph shows the correspondence between the predicted RCT confidence centered at each 0.10 width range between 0.0 and 1.0, and the prevalence of articles determined to describe RCTs by manual review. Samples were chosen randomly across four searches corresponding to Cochrane topics where none of the chosen articles were tagged in MEDLINE with the "Randomized Controlled Trial" publication type. It can be seen that estimated prevalence is slightly below the predicted confidence. This is likely due to two reasons. First, in order to keep the manual review task modest, the binning that was used to group the confidence ranges, and the number of samples in each bin are somewhat coarse. Second, and more importantly, the manually reviewed samples do not represent a uniform random sample from MEDLINE. The samples were specifically chosen to not have the MEDLINE RCT_PT. Since all of these had been previously reviewed by MEDLINE annotators and not tagged with this publication type, it is reasonable to expect that these articles would have somewhat less than predicted chance of being RCTs. Still, for the articles with high predicted confidence, a large fraction of the articles were designated as RCTs by the reviewer.




The most extreme disagreements between our predictions and the annotations occurred where the CITATION_PLUS_MESH model predicted confidence is very high and the MEDLINE *RCT_PT* is absent, and also where the predicted confidence is very low and the MEDLINE publication type is present. Across the entire dataset, which includes articles meeting the human-related article search criteria (discussed above) from 1987 to 2012, 1811 articles with RCT confidence predictions $\geq$0.95 were found that did not have the MEDLINE *RCT_PT* assigned. Based on the manually reviewed RCT assignments shown in Figure 1, we expect about 70% (1268) of the 1811 high confidence prediction articles to correctly be RCTs. Using a less extreme threshold, it was found that 19 838 articles where the CITATION_PLUS_MESH model predicted confidence was $\geq$0.50 and the MEDLINE *RCT_PT* is missing. Again, based on Figure 1, we expect about 40% (7935) of these articles to be RCTs. The full 1987–2012 human related article dataset has 277 789 articles tagged with the MEDLINE *RCT_PT*, and so these 7935 articles represent a potential additional 3% correctly annotated RCT articles.

Examining the other type of disagreement, 20 523 articles were found with RCT confidence predictions $\leq$0.05 that did have the MEDLINE *RCT_PT* assigned. Using the confidence prediction and based on the results in Figure 1, it is estimated that 19 497 (95% of 20 523) articles may be assigned the *RCT_PT* incorrectly. This represents an error rate of about 7%.

The abstract and full text of the lowest scoring ten articles of this 20 523 article set were manually reviewed, all of which had CITATION_PLUS_MESH model predicted confidence $<0.0005$. After manual review, every one of these articles was determined to not be a specific RCT.

The random subset of 50 previously unseen articles published in 2014 annotated with the *RCT_PT* showed 14 articles where the independent annotator disagreed with the *RCT_PT* assignment, the annotator agreed with the other 36 assignments. Of these 14 articles, 11 had CITATION_PLUS_MESH confidence scores below 0.50, 5 had low confidence scores below 0.25, and 3 had very low confidence scores below 0.01. Reasons given by the independent annotator for designating the articles as not RCTs included: "not randomized," "not controlled," "treatment allocation blocked by clinic not patient," and "describes a proposed RCT." Based on these initial analyses, it is conservatively estimated that at least 5% of the articles having the MEDLINE *RCT_PT* are likely to not actually be RCTs and that these errors could be detected by our model.

Overall, the model could be used to identify approximately 7900 (3%) additional RCTs in MEDLINE. In addition, approximately 12 900 (5%) potential *RCT_PT* assignment errors could be identified. Combined, this represents about 20 000 (7%) potentially identifiable and fixable, *RCT_PT* errors.

## DISCUSSION

In this work, a machine learning model was built, which provides highly accurate confidence predictions about whether or not an article is an RCT using only information available in the MEDLINE record including the citation, the article abstract, and the assigned MeSH terms. While the CITATION_PLUS_MESH model is more accurate across all measures, it can of course only be applied to articles that have been annotated in MEDLINE. Therefore, a second model was created that is almost as good as the first model, but requires only the citation and abstract. Based on the manual evaluations, the current research also demonstrates that the automated retagging could be useful for reviewing RCTs in MEDLINE. First, the approach can identify articles not annotated in MEDLINE that are highly likely to actually be RCTs. Second, the approach can identify articles annotated as RCTs in MEDLINE that may not in fact be RCTs.

The machine learning model provides an RCT confidence prediction, rather than simply a binary assignment, which allows articles to be ranked by their RCT confidence. While it is expected that any article tagged with the MEDLINE *RCT_PT* will be considered by pipeline users as potentially an RCT, retagging provides additional information on all the articles both those tagged and not tagged with the MEDLINE publication type. Articles can be ordered by their RCT confidence values; this can be used for work prioritization and reading assignment by the SR team.

Using the CITATION_ONLY model, citations can be accurately ranked before they are tagged in MEDLINE and, therefore, do not otherwise have any assigned publication types. Since the citation information used in the model (title, authors, journal, and abstract) is essentially equivalent between databases, the CITATION_ONLY model should perform equally on articles published in MEDLINE-indexed journals identified by searching other databases. Further study is required to determine whether the performance of the models will be sufficient for articles indexed in other databases that are published in journals not included in MEDLINE. For our primary purpose of incorporating the RCT retagger in a meta-search engine, this will allow combining all citations retrieved by a search into a single ranked list, both those present in MEDLINE as well as those only found in other bibliographic databases. The most likely RCT articles, both those not tagged as such in MEDLINE or not indexed in MEDLINE, will be identified. Articles that are tagged with the MEDLINE *RCT_PT* but may not actually be RCTs may be ranked lower than articles with higher confidence.

Systematic reviewers can select a confidence cutoff threshold and decide to postpone or not review any articles that have either a low confidence, are not tagged with the *RCT_PT*, or both. Using our retagging models, users can decide on a per-topic basis what confidence cutoff level to apply in determining which articles to review for potential inclusion in an SR. For example, the acute cholecystitis literature search (see Table S2 in the Supplementary material) returned 3883 articles in our data set. About 20% (795) of these were annotated with the MEDLINE *RCT_PT*, the rest were not. Of the 3088 unannotated articles, 2751 had confidence predictions $\leq 0.10$. If the review team skipped just these 2751 articles, they would avoid the need to examine 85% of the unannotated articles. Conversely, for this topic search, 337 articles not annotated with the *RCT_PT* were found with confidence predictions $> 0.10$, and it is highly likely that searching this set of 337 articles will capture most, if not all, of the true RCTs residing within the 3088 unannotated articles.

As another direction of future work, the RCT confidence predictions could also be used as a means to review the application of the *RCT_PT* across all of MEDLINE. The manual evaluation of topic specific query results showed that there were a substantial number of articles not assigned the MEDLINE *RCT_PT* that had high RCT confidence scores. The random manual evaluation of articles assigned the MEDLINE *RCT_PT* showed articles that were not RCTs that were assigned the MEDLINE *RCT_PT* and a substantial number of these had very low confidence scores. This suggests that articles assigned the MEDLINE *RCT_PT* and having very low confidence scores may benefit from re-review. In general, articles with the greatest disagreement between the publication type and the retagger confidence prediction in either direction could be reviewed and possibly updated. Similar work has been performed manually by Weiland *et al.* [3]; the retagger opens the possibility of automating this process and repeating it periodically.

As far as we are aware, this is the first work demonstrating a highly accurate machine learned-based model for predicting the confidence of whether an article describes a RCT. Researchers have published work describing optimized search terms that can be used within an information retrieval system such as PubMed.[21–23] Several groups, including ourselves, have published work discussing the application of machine

learning techniques to improve the process of SR. Some machine learning models have been designed to identify articles likely to be included in a review given a prior set of reviewed documents,[8,24–28] while others identify specific high quality content (eg, articles, sentences) in terms of EBM criteria.[29–32] Interestingly, Bekhuis and Demner-Fushman have published work on using machine learning to screen nonrandomized studies for inclusion in SRs.[33,34] There has also been a significant amount of general Natural Language Processing (NLP) research (beyond machine learning) with applications to improving the SR process and supporting EBM in various ways (eg,[35–37]). There is also substantial literature discussing ways of improving the SR process independent of machine learning or NLP techniques (eg,[38–41]).

## CONCLUSION

In this work, highly accurate models were developed for identifying RCTs across a large subset of MEDLINE publications, both with and without MeSH and other MEDLINE annotation features. These models use a new enhancement to Rüping's method to create accurate confidence predictions from margin distances. These confidence predictions can be used to aid SR, and the method may assist both prospectively and retrospectively in quality control for manually assigned publication type tagging in MEDLINE and potentially other bibliographic databases. It has also been demonstrated that post hoc feature reduction can be used to significantly reduce the size and complexity of predictive models without diminishing performance.

The RCT confidence predictions described here have been made available to users as a free, public web service with a user query form front end at: http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/RCT_Tagger.cgi. Users can enter a PubMed query into the form, which returns a list of articles indexed in PubMed ranked by descending RCT confidence. Retrieved citations and corresponding RCT confidences can be downloaded in XML or BibTeX format. This service is part of our text mining pipeline to support SR and EBM.[42–44] This will offer significant savings of time and effort during the process of writing SRs.

## ACKNOWLEDGEMENTS

The authors wish to acknowledge the Cochrane Collaboration for providing the search queries used for the specific systematic review topics discussed in this manuscript. The authors also wish to thank Tracy Edinger N.D., for serving as the independent RCT manual reviewer.

## CONTRIBUTORS

The implementation and evaluation of the machine learning models were primarily performed by AMC and NRS. All authors contributed to the conception of this work and the preparation and review of the manuscript for publication.

## FUNDING

This work was supported by National Institutes of Health/National Library of Medicine grant number R01LM010817.

## COMPETING INTERESTS

None.

## SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

## REFERENCES

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ.* 1996;312(7023):71–72.
2. Haynes RB. What kind of evidence is it that evidence-based medicine advocates want health care providers and consumers to pay attention to? *BMC Health Serv Res.* 2002;2(1):3.
3. Wieland LS, Robinson KA, Dickersin K. Understanding why evidence from randomised clinical trials may not be retrieved from Medline: comparison of indexed and non-indexed records. *BMJ.* 2012;344:d7501.
4. Edinger T, Cohen AM. A large-scale analysis of the reasons given for excluding articles that are retrieved by literature search during systematic review. *AMIA Annu Symp Proc.* 2013;2013:379–387.
5. Cohen AM, Adams CE, Davis JM, *et al.* Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. *Proceedings of the 1st ACM International Health Informatics Symposium November, 2010; Arlington, Virginia USA.* 2010:376–380.
6. Harker J, Kleijnen J. What is a rapid review? A methodological exploration of rapid reviews in Health Technology Assessments. *Int J Evid Based Healthc.* 2012;10(4):397–410.
7. Khangura S, Konnyu K, Cushman R, Grimshaw J, Moher D. Evidence summaries: the evolution of a rapid review approach. *Syst Rev.* 2012;1:10.
8. Cohen AM. Optimizing feature representation for automated systematic review work prioritization. *AMIA Annu Symp Proc.* 2008;2008:121–125.
9. Tsertsvadze A, Maglione M, Chou R, *et al. Updating Comparative Effectiveness Reviews: Current Efforts in AHRQ's Effective Health Care Program. Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet].* Rockville (MD): Agency for Healthcare Research and Quality (US); 2008 [cited February 3, 2014]. http://www.ncbi.nlm.nih.gov/books/NBK66066/. Accessed February 3, 2014.
10. McDonagh MS, Jonas DE, Gartlehner G, *et al.* Methods for the drug effectiveness review project. *BMC Med Res Methodol.* 2012;12:140.
11. Funk ME, Reid CA. Indexing consistency in MEDLINE. *Bull Med Libr Assoc.* 1983;71(2):176–183.
12. Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning, April, 1998; Chemnitz, Germany. 1998;137–142.

13. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. *J Mach Learn Res.* 2008;9:1871–1874.
14. Rüping S. A simple method for estimating conditional probabilities for svms. *Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen*; 2004.
15. Cohen AM. An effective general purpose approach for automated biomedical document classification. *AMIA Annu Symp Proc.* 2006;2006:161–165.
16. Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann 1993;1022–1027.
17. Gurusamy KS, Davidson C, Gluud C, Davidson BR. Early versus delayed laparoscopic cholecystectomy for people with acute cholecystitis. *Cochrane Database Syst Rev.* 2013;6: CD005440.
18. Rösner S, Hackl-Herrwerth A, Leucht S, Vecchi S, Srisurapanont M, Soyka M. Opioid antagonists for alcohol dependence. *Cochrane Database Syst Rev.* 2010;(12): CD001867.
19. Anderson K, Norman RJ, Middleton P. Preconception lifestyle advice for people with subfertility. *Cochrane Database Syst Rev.* 2010;(4):CD008189.
20. Worthington HV, Clarkson JE, Bryan G, *et al*. Interventions for preventing oral mucositis for patients with cancer receiving treatment. *Cochrane Database Syst Rev.* 2011;(4): CD000978.
21. Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. *Int J Epidemiol.* 2002;31(1): 150–153.
22. Haynes RB, Wilczynski N. Finding the gold in MEDLINE: clinical queries. *ACP J Club.* 2005;142(1):A8–A9.
23. Wilczynski NL, McKibbon KA, Walter SD, Garg AX, *Haynes RB. MEDLINE clinical queries are robust when searching in recent publishing years*. J Am Med Inform Assoc [Internet]. September 27, 2012 [cited February 11, 2013]. http://jamia.bmj.com/content/early/2012/09/26/amiajnl-2012-001075. Accessed January 5, 2015.
24. Cohen AM, Ambert K, McDonagh M. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Med Inform Decis Mak.* 2012;12(1): 33.
25. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inf Assoc.* 2006;13(2): 206–219.
26. Cohen AM, Ambert K, McDonagh M. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. *AMIA Annu Symp Proc.* 2010;2010:121–125.
27. Frunza O, Inkpen D, Matwin S, Klement W, O'Blenis P. Exploiting the systematic review protocol for classification of medical abstracts. *Artif Intell Med.* 2011; 51(1):17–25.
28. Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. A new algorithm for reducing the workload of experts in performing systematic reviews. *J Am Med Inf Assoc.* 2010; 17(4):446–453.
29. Aphinyanaphongs Y, Statnikov A, Aliferis CF. A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. *J Am Med Inf Assoc.* 2006;13(4):446–455.
30. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. *J Am Med Inf Assoc.* 2005;12(2):207–216.
31. Demner-Fushman D, Seckman C, Fisher C, *et al*. A prototype system to support evidence-based practice. *AMIA Annu Symp Proc.* 2008;151–155.
32. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Inf Assoc.* 2009;16(1):25–31.
33. Bekhuis T, Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artif Intell Med.* 2012;55(3):197–207.
34. Bekhuis T, Tseytlin E, Mitchell KJ, Demner-Fushman D. Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PLoS ONE.* 2014;9(1):e86277.
35. Agarwal S, Yu H. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics.* 2009;25(23): 3174–3180.
36. Chung GY. Sentence retrieval for abstracts of randomized controlled trials. *BMC Med Inform Decis Mak.* 2009;9(1):10.
37. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. *J Biomed Inf [Internet].* 2008. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19022398.
38. Blake C, Pratt W. Collaborative information synthesis II: recommendations for information systems to support synthesis activities. *J Am Soc Inf Sci Technol.* 2006;57(14): 1888–1895.
39. Moher D, Tsertsvadze A, Tricco AC, *et al*. When and how to update systematic reviews. *Cochrane Database Syst Rev.* 2008; (1):MR000023.
40. Sampson M, Shojania KG, McGowan J, *et al*. Surveillance search techniques identified the need to update systematic reviews. *J Clin Epidemiol.* 2008;61(8):755–762.
41. Cooke A, Smith D, Booth A. Beyond PICO The SPIDER Tool for Qualitative Evidence Synthesis. *Qual Health Res.* 2012; 22(10):1435–1443.
42. Smalheiser NR, Lin C, Jia L, *et al*. Design and implementation of meta, a meta search engine for biomedical literature

retrieval intended for systematic reviewers. *Health Inf Sci Syst.* 2014;2(1): 1.
43. Jiang Y, Lin C, Meng W, Yu C, Cohen AM, Smalheiser NR. Rule-based deduplication of article records from bibliographic databases. *Database.* 2014;2014:bat086.
44. Shao W, Adams CE, Cohen AM, *et al.* Aggregator: a machine learning approach to identifying MEDLINE articles that derive from the same underlying clinical trial. *Methods.* 2014 http://www.sciencedirect.com/science/article/pii/S1046202314003661.

## AUTHOR AFFILIATIONS

[1]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239 USA

[2]Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 60612 USA

[3]Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60612 USA

[4]Division of Psychiatry, University of Nottingham, Nottingham, UK