# Active Learning for the Automation of Medical Systematic Review Creation

*Full Paper*

**Prem Timsina**
Dakota State University
ptimsina@pluto.dsu.edu

**Omar El-Gayar**
Dakota State University
omar.el-gayar@dsu.edu

**Jun Liu**
Dakota State University
jun.liu@dsu.edu

## Abstract

While systematic reviews (SRs) are positioned as an essential element of modern evidence-based medical practice, the creation of these reviews is resource intensive. To mitigate this problem there has been some attempts to leverage supervised machine learning to automate the article triage procedure. This approach has been proved to be helpful for updating existing SRs. However, this technique holds very little promise for creating new SRs because training data is rarely available when it comes to SR creation. In this research we propose an active machine learning approach to overcome this labeling bottleneck and develop a classifier for supporting the creation of systematic reviews. The results indicate that active learning based sample selection could significantly reduce the human effort and is viable technique for automating medical systematic review creation with very few training dataset.

## Introduction

Evidence Based Medicine (EBM) refers to the application of state-of-the-art medical evidence to improve the quality and reduce the cost of medical care (Cohen et al. 2010). Although the classical vision of EBM required physicians to directly search the relevant medical research for evidence applicable to their patients, the modern conception of EBM heavily relies on synthesis of research findings in the form of evidence reports commonly referred to as a systematic reviews (SR). According to Higgins and Green (2011), "*A systematic review is a high-level overview of primary research on a particular research question that tries to identify, select, synthesize and appraise all high quality research evidence relevant to that question in order to answer it*". Currently, SRs form a key resource for informing medical practice. With the increasingly rapid pace by which medical knowledge is created, researchers, practitioners and policy makers are challenged to keep pace with state-of-the-art medical evidence and incorporate such evidence into practice. SRs respond to this issue by recognizing, appraising, and synthesizing research-based evidence from multiple sources and presenting it in an accessible format (Mulrow 1994).

While SRs are positioned as an essential element of modern evidence-based medical practice, developing these reviews is resource intensive. Surprisingly, the current workflow for creating SRs is largely a manual process. An initial search by querying databases such as Medline, Cochrane and Embase often returns thousands of articles given a medical topic. To select the articles that should be included in a review or not, scientists need to review the articles in two steps. The first step is called the abstract triage, where scientists identify "relevant" articles that can be potentially included in a SR based on the title and abstract of the articles. This phase of screening articles usually requires a significant amount of time and effort as it involves a group of scientists evaluating thousands of articles in order to find the relevant instances returned by the database queries. The second step referred to full-text triage, involves full text inspections of those relevant ones selected in the title/abstract triage to determine the articles that satisfy the inclusion criteria and will be included in a systematic review (Shojania et al. 2007). Due to the manual workflow of selecting articles for SRs, developing SRs requires a significant investment in time (1,139 expert hours on average) and funds (up to a quarter of a million dollars) from a dedicated and qualified research team (Allen et al. 1999; McGowan et al. 2005).

Nowadays, medical knowledge base is growing at an astounding pace. Reports of new clinical trials are being published at the rate of over 20,000 per year (Shojania et al. 2007). This creates an enormous challenge for scientists trying to create systematic reviews to keep pace with the developments in the medical field. Cochrane Collaboration estimates that at least 10,000 new SRs are needed to cover most of the healthcare problems (Higgins et al. 2011). Unfortunately, fewer than half of this number has been published even after ten years of focused effort by the EBM community (Higgins et al. 2011).

The unavailability of new medical evidence is hugely attributed to the expert effort needed for creation of new Systematic Review. To mitigate this problem there has been some attempts to leverage supervised machine learning to automate the article triage procedure. This approach has been proved to be helpful for updating existing SRs. Most of existing research assumes readily available training data and focus on updating reviews by applying supervised learning to new medical evidence. This supervised learning based approach, however, holds very little promise for creating new SRs. This is because supervised machine learning needs considerable amount of training data; nevertheless, training data is rarely available when it comes to SR creation. Also, creating training data with labels indicating when an article should be included or excluded in a SR is difficult, laborious and time-consuming.

Active learning approach has received considerable attention due to its potential for achieving greater accuracy with fewer training labels. Active learning attempts to overcome the labeling bottleneck by identifying the most informative set of unlabeled instances and let them be labeled by an oracle (e.g., a human annotator). Active learning is an iterative process in which it first train a machine learning algorithm with few training instances, based on the training results, it selects an optimal set of unlabeled instances and queries an oracle for manual labeling, and then it re-train the algorithm based on the incremented training data. In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data.

The overarching goal of our research is to develop an advanced analytics approach to automatically identify relevant articles that could be used to create SRs, based on the title and abstract of articles. More specifically, given the fact that when it comes to creating a new SR, labeled training data (i.e., articles that have been reviewed by human experts to be included in or excluded from a systematic review) is mostly not readily available and is difficult and time-consuming to obtain, we plan to adopt active learning to overcome this labeling bottleneck and develop data mining models that classify articles for inclusion or exclusion, thus helping automate SR creation with only a few labeled instances. To our knowledge, the proposed research is one of the first that attempts to address the small-sized training dataset problem that hampers the use of classification algorithms in SR creation.

## Related Work

There have been some attempts in literature to leverage analytics to automate SR generation procedure (Ananiadou et al. 2009; Bekhuis et al. 2012; Cohen et al. (2006) ; Frunza et al. 2010; Shemilt et al. 2013). One of the most significant research done in this area is one conducted by Cohen et al. (2006b). In this National Institute of Health (NIH) supported project, Cohen et al, used the perceptron algorithm to identify journal articles for inclusion in systematic review update, based on the title and abstract of the articles. While the research is formidable in the automation of SR creation, the techniques used 50% of data as training set. This is very costly.

There are also studies that focus on comparing different algorithms that can be used to classify articles for systematic reviews. For instance, Bekhuis & Demner-Fushman (2012), 2012 compared algorithms including K-nearest neighbor (K-NN), naïve Bayes, complement naïve Bayes (cNB), and evolutionary SVM (EvoSVM) (implemented in the RapidMiner) and used information gain as their feature selection method to select features from titles or titles and abstract, and full citation with metadata. The authors demonstrated that based on text mining techniques, the number of documents that need further manually screening was reduced by up to 46%, and among the three algorithms, EvoSVM achieved the highest recall (100% for both datasets) and relatively low precisions (13.11% for the ameloblastoma dataset and 10.69% for the influenza dataset). Adeva et al.'s (Adeva et al. 2014) research is probably the most comprehensive so far. They conducted experiments that involved multiple classification algorithms (including naïve Bayes, k-Nearest neighbor, Support vector machines, and Rocchio) combined with several feature selection methods (including TF, DF, IDF, etc.), and applied to different parts of the given

articles (including titles alone, abstracts alone and both titles and abstract). SVM has the highest F-measure when using both titles and abstracts.

On the other hand, there is existing research, though not in the area of SRs, that demonstrates the possibility active learning in the case of rare training instances. For example, Tong et al. (2002) used SVM active learning for text classification. Authors showed that employing active learning method significantly reduces the need for labeled training instances in both the standard inductive and transductive settings. Using version space approach they proved that active learning, which access to a pool of unlabeled instances and request the labels for some number of them, is superior than using a randomly selected training set. In another study, Kim et al.(2006) incorporated Maximal Marginal Relevance (MMR) based active learning into a biomedical named-entity recognition system. Their experimental results indicated that active learning based sample selection could significantly reduce the human effort.

Overall, extant research focuses on applying supervised learning to article selection for a SR, assuming the existence of a large number of labeled training examples. Supervised learning is practical for article selection in SR updates, but less feasible for SR creations that often start with zero or few labeled articles. Again, existing literature in SRs automation indicated an active learning system that is carefully designed is possible in principle, and is an interesting area for future research for SR creations (Cohen et al. 2009). Existing research on active learning also show promise to consider the use of it in text classification with few labeled examples. We hence propose to apply active learning to systematic review creation.

## Research Gap

Our literature review indicates that 1) SRs creation procedure is very resource and time intensive, and 2) the generation of a training dataset for article classification requires significant human efforts. This leads us to the following overarching research question:

## Methodology

Table 1 shows the proposed process for creating systematic Reviews. We used Systematic Review data corpus provided by Oregon Health and Science University (2006b) as our data sources. Our analytics approach to identifying relevant articles for systematic reviews includes two major steps: 1) active learning for creating a training dataset incrementally, and 2) soft-margin based support vector machine for creating our final model. We conduct experiments to evaluate the effectiveness of the proposed approach. We start with 2-10% of labeled articles as seeds or initial training instances. Using the seeds, we conduct active learning to identify a few training articles that are informative for learning and need to be reviewed by human experts. The outcome of active learning, the incremented training dataset, is then fed into the soft-margin SVM algorithm, which produces a classification model. In following sub-sections, we describe the data source, and each component of our model in detail.

**Table 1: Proposed process for creating systematic Reviews**

1. Manually create an initial training set X comprised of (2% to 10%) of the data corpus as a labeled set for seeds.
2. Employ active learning to expand the training data set.
    a. Use X as a training set.
    b. Use the label-spreading algorithm to classify each unlabeled article and record statistical confidence, which is a probability measure, for each classification
    c. Identify the top 6 articles (with highest statistical confidence) classified to be positive (i.e., inclusion in an SR) and the top 6 articles classified to be negative (i.e., exclusion from an SR).
    d. Add the top 6 negative articles to the labeled dataset.

```
        e. Ask human experts to label the top positive articles.
        f. Add the human labeled articles to the labeled dataset
        g.  Repeat steps "2.b. to 2.f." for 5 iterations.
  3. Apply supervised learning (using the developed training set)
        a. Apply soft-margin SVM to the training dataset obtained in
           step 2
        b. Test the proposed classifier (using a test data set)
```

### *Data Sets*

We used three systematic reviews on drug topics including AtypicalAntipsychotics (AT), NSAID, and Estrogens (ESTRO) performed by AHRQ's Evidence-based Practice Center (EPC) at Oregon Health and Science University as our datasets. These three systematic review datasets were also used in (2006b). The original datasets downloaded from (Cohen 2014) include the PubMed Unique Identifiers (PMID) of all the articles, whether included or excluded from the reviews, and the inclusion and exclusion decisions made by human researchers. Following (2006b), we focus on classifying the articles based on the title and abstract of the articles. We hence used Medline's Batch Entrez features to extract the title and abstract of all the articles based their PMIDs. Table 2 shows an overview of the datasets.

**Table 2: Overview of Data Corpus**

| Dataset | Total number of articles | Number of excluded articles | Number of included articles | Ratio—Included vs. Excluded |
|---|---|---|---|---|
| AtypicalAntipsychotics (AT) | 1120 | 751 | 361 | 1:2 |
| NSAID | 393 | 305 | 88 | 1:3.5 |
| Estrogens (ESTRO) | 370 | 289 | 81 | 1:3.6 |

### *Data Pre-processing*

We represented each article in our datasets using the standard vector space, aka, bag-of-word model (Lebanon et al. 2007). We created a vector for each article that includes the words in the title, abstract, Medline publication type, and Medical Subject Heading of the article. To create a vector for an article, we transformed all upper case words of the article into lower case words. We then split the abstract of the article into a sequence of tokens using "non-letter" character as the splitting point. We then filtered the English stop-words from each document. For creating a bag of words we used term frequency-inverse document frequency (tf-idf) technique (Robertson 2004). Basically, tf-idf is a numerical statistic that reveals the importance of a word in a document in a data corpus. The tf-idf value increases, as a word appears more often in a document; however, the tf-idf value is offset by the frequency of the word in the data corpus. This helps to mitigate for the fact that some words such as, "patient" are generally more common than other words in medical documents.

### *Active Learning:*

We propose to use active learning to create a training dataset with a few labeled articles. The key idea behind active learning is that a machine-learning algorithm can achieve greater accuracy with fewer training labels. An active learner identifies a set of unlabeled data instances that need to be labeled by a human expert. Active learning is well motivated in many modern machine learning problems, where unlabeled data may be abundant or easily obtained, but labels are difficult, time-consuming, or expensive to obtain (Settles 2010). In our research, we have used a modified version of active learning (Chapelle et al. 2006).

1. *Input: labeled data* $\{(x_i, y_i)\}_{i=1}^{i=l}$, *unlabeled data* $\{x_j\}_{j=l+1}^{l+u}$
2. *Initially, let* $L = \{(x_i, y_i)\}_{i=1}^{l}$, $U = \{(x_j)\}_{j=l+1}^{l+u}$ and L contains same amount of positive samples and negative samples

3. *Repeat:*
    a. *Train classifier f using L.* □
    b. *Apply f to the unlabeled instances U .*
    c. *Remove a subset S¹ and S² from U;* where, f(x) predicts *S¹€Negative Class and* f(x) predicts *S²€Positive Class*
    d. *Add S¹* to L
    e. *Ask human oracle to annotate S² resulting in S³*
    f. Add S³ to L where $\{S^3 \in S^2\}$

In our research, we conducted multiple experiments to identify the optimum parameters for active learning. We chose to use 5 iterations of active learning. In each iteration, we added 6 articles predicted by the algorithm as negative and another 6 articles predicted as positive to the labeled set. As discussed previously, there are more negative articles (articles that would not be included in an SR) than positive ones in a typical systematic review dataset; machine learning hence tends to achieving high accuracy on predicting the negative articles, as evidenced by existing research (Shemilt et al. 2013). Thus, we added those samples predicted by the algorithm as negative into labeled set without asking human experts to annotate. Positive articles (i.e., articles that need to be included in a SR), on the other hand, are often rare and machine learning often identifies them with low precision. It is hence necessary for human experts to review and label the articles and then add them to the training dataset. In our datasets, all of the articles are labeled. Thus in our experiments, there is no need for human labeling, but we count reviewing 30 articles (5 iterations, and 6 articles predicted to be positive in each iteration) as the workload for human experts.

## *Label Spreading Algorithm*

We propose to use the label-spreading algorithm (Zhou et al. 2004) as the algorithm for active learning. Our analysis with multiple algorithms including S3VM, graph-based, self-learning, and SVM showed that the label spreading algorithm outperforms other algorithms with small-sized training dataset. We describe the algorithm in following sub-sections (Zhou et al. 2004).

Given a point set $X = \{x_1,...,x_l,x_{l+1},...,x_n\} \subset R^m$ and a label set $L = \{1,...,c\}$, the first l points $x_i (i \le l)$ are labeled as $y_i \in L$ and the remaining points $x_u (l+1 \le u \le n)$ are unlabeled. The goal is to predict the label of the unlabeled points.

Let F denote the set of n × c matrices with nonnegative entries. A matrix $F = [F_1^T, \ldots, F_n^T]^T \in F$ corresponds to a classification on the dataset X by labeling each point $x_i$ as a label $y_i = \arg\max_{j \le c} F_{ij}$ . We can understand F as a vectorial function $F : X \rightarrow R^c$ which assigns a vector $F_i$ to each point $x_i$. Define a n×c matrix $Y \in F$ with $Y_{ij} = 1$ if $x_i$ is labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise. Clearly, Y is consistent with the initial labels according the decision rule. The algorithm is as follows (Zhou et al. 2004):

1. Form the affinity matrix W defined by $W_{ij} = \exp(-||x_i - x_j||^2/2\sigma^2)$ if i≠ j □and $W_{ii} = 0$. □

2. Construct the matrix $S = D^{-1/2}WD^{-1/2}$ in which D is a diagonal matrix with □its (i, i)-element

equal to the sum of the i-th row of W. □

3. Iterate F(t+1)=αSF(t)+(1−α)Y until convergence, where α is a parameter □in (0, 1). □

4. Let $F^*$ denote the limit of the sequence {F(t)}. Label each point $x_i$ as a label □$y = \text{argmax } F^*_{ij}$.

This algorithm can be understood intuitively in terms of spreading activation networks (Anderson 1983; Shrager et al. 1987) from experimental psychology.

### *Supervised Machine Learning Algorithm—Soft Margin SVM:*

We propose to use soft-margin linear SVM after active learning creates a sufficient number of labeled instances. The soft-margin SVM has shown to outperformed other supervised learning algorithm in case of medical document classification (citation after peer-review). In order to explain our soft-margin SVM, we describe the SVM algorithm first.

*SVM*: Existing studies such as (Bekhuis et al. 2012; Joachims 1998; Liu et al. 2002) have proved the effectiveness of SVM with a linear kernel in text classification in the process of medical systematic reviews. The optimization problem associated with the SVM is shown below.

$$min_{\mathbf{w},b} \frac{\mathbf{w^T w}}{2}$$

$$\text{subject to: } y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 \ (\forall \text{ data points } \mathbf{x_i}).$$

where for each data point ($x_i$, $y_i$), $y_i$ is either 1 or −1, indicating the class to which the point belongs. The two hyperplanes $w \cdot x - b = 1$ and $w \cdot x - b = -1$ are called support vectors that separate the data. SVM maximizes the distance (called "margin") between the support vectors.

*Soft-margin linear SVM*: We propose to use the soft-margin Support Vector Machine (SVM) with a linear kernel as a classifier. Soft-margin SVM is an extension of the standard "hard" margin SVM described above.

The "hard-margin" SVM sometimes does not work well since it does not allow data points in the margin. However, data is not often perfectly linearly separable, and it is necessary to allow some data points of one class to appear within the region bounded by the support vectors. Soft-margin SVM provides the flexibility by introducing a slack variable $\epsilon_i \geq 0$, and the optimization problem of soft-margin SVM becomes(Stanford 2014):

$$min_{\mathbf{w},b,\epsilon} \frac{\mathbf{w^T w}}{2} + C\sum_i \epsilon_i$$

$$\text{subject to: } y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \epsilon_i \text{ and } \epsilon_i \geq 0 \ (\forall \text{ data points } \mathbf{x_i}).$$

where $\epsilon_i$, the slack variable, represents the degree of error in classification.

### *Evaluation*

We evaluated the classification performance using four measures: precision, recall, F1-score and Work Saved a measure proposed in(Frunza et al. 2010). These measures are defined based on a confusion matrix as shown in Table 3. In our research, we treat the articles that were included in a review as positive samples and those that were excluded as negative samples. TP represents the number of True Positives, i.e., positive samples that were correctly classified by our SVM classifier. TN is the number of True Negatives, i.e., negative samples that were correctly classified, FP the number of False Positive, i.e., negative samples that were incorrectly classified as positive, and FN the number of False Negatives, i.e., positive samples incorrectly classified as negatives.

| **Table 3: Confusion Matrix** | | |
|---|---|---|
| | **Predicted Negative** | **Predicted Positive** |
| Actual Negative | True negative (TN) | False positive (FP) |
| Actual Positive | False negative (FN) | True positive (TP) |

The formulas for computing Recall, Precision, F1 and WSS are shown in Table 4. Recall refers to the rate of correctly classified positives among all positives and is equal to TP divided by the sum of TP and FN. Precision refers to the rate of correctly classified positives among all examples classified as positive and is equal to the ratio of TP to the sum of TP and FP. F1 means the harmonic mean of recall and precision. WSS defined as percentage of samples that met the initial search criteria that the human reviewers do not have to read because they have been correctly screened by the classifier.

| **Table 4: Evaluation metrics** | |
|---|---|
| **Evaluation Metric** | **Formula** |
| Recall | TP/ (TP+FN) |
| Precision | TP/(TP+FP) |
| F1 | (2*(recall*precision)/(recall + precision) |
| WSS | (TN + FN)/(TN + FN + TP + FP) – 1+ TP/(TP + FN) |
| WSS=Work Saved over Sampling | |

# Experiments and results

We conducted two experiments to evaluate the effectiveness of the proposed approach. In Experiment I, we evaluated the effectiveness of active learning for creating a training dataset as compared with random sampling. We intended to investigate that in order to create a certain number of labeled instances, how many articles needs to be reviewed by human experts, based on active learning vs. random sampling. Experiment 1 consisted three sub-experiments, each of which involved a different article set. Each sub-experiment included five steps, and in each step, we used different numbers of seeds (i.e., initially labeled articles). We used 2% positive examples in step 1, 4% positive examples in step 2, 6% in step 3, 8% in step 4, and 10% in step 5. We also randomly selected the same number of negative examples in each step. In each step, we first performed active learning based on the seeds to increment the number of positive examples by *k*, and counted the number of articles that human experts need to review and manually label. We then conducted random sampling. We selected articles randomly, stopped when we obtain *k* positive articles, and counted the number of articles for human labeling. We compared the number of articles for human labeling using active learning vs that using random sampling to determine the effectiveness of active learning in creating the training dataset. To confirm the reliability of the results, in each step, we conducted 20 trials of active learning and random sampling. Then, we averaged the results of 20 trials to generate the final results for each step. This approach is consistent with an earlier approach used in literature to comparing the results between active learning and random sample (Zhu et al. 2002). In Experiment 2, we focused on evaluating the performance of the proposed classifier with regard to the creation of systematic reviews. Experiment 2 is an continuation of Experiment 1. Given an article set, in Experiment 1, we selected different numbers of seeds (2%, 4%, 6%, 8% and 10%) and conduct 20 trials of activeness. Hence, given a certain number of seeds (e.g., 2% positive examples and the same number of negative ones), active learning returned 20 training datasets in 20 trails. We used these 20 training sets in Experiment 2 to fit 20 soft-margin SVM models. We obtained 20 precision, recall, F1 and WSS measures based on cross-validation and then took the average of the measures.

## *Experiment I: Effectiveness of active learning in creating a data set*

Table 5 shows the results of experiment I. We found that active learning approach was able to create considerable number of new positive examples. The number of new positive examples created via this approaches varies from 11.95 to 18.25 (average of 20 trials). The lowest number of positive examples created was in the sub-experiment involving the ESTOR dataset with 2% seeds (11.95 new positive

samples). The highest number of positive examples created was in the sub-experiment involving the NSAID dataset with 10% seeds (18.25 new positive samples).

For all datasets, we found that the number of positive examples created by active learning increases as the number of seeds increases. For example, for the NSAID dataset, with 2% seeds, the active learning approach was able to produce 13.10 positive examples; whereas with 10% labeled examples the same approach was able to produce 18.25 positive examples, which is as expected because as the number of seeds increases machine learning model becomes more accurate.

Experiment I also shows that if random sampling is done to select articles for human labeling, considerably more number of examples should be scanned as compared with the active learning approach. For an instance, for the ESTOR datasets with 4% labeled data, active learning approach was able to identify 13.50 (average of 20 trials) positive samples by scanning 30 articles. To identify same amount of positive samples (13.50 articles), 54.98 articles should be scanned through the random sampling approach. Overall, experiment I shows that the active learning approach reduces considerable amount of work for training dataset creation. For example, our lowest work reduction in training set creation was 47.43%. That is, even with the 2% labeled sample as seeds; we were able to optimize workload by about 50% (47.43%). Our result also attests to the findings of extent literature, which indicate that active learning based sample selection could significantly reduce the human effort(Kim et al. 2006).

**Table 5: Experiment I results**

| DATASET | Percentage Of Labeled Sample (Training Dataset, Or Seeds) **[Column A]** | Labeled Sample (seed) **[Column B]** | Average Number Of Positive Sample Created By Active Learning **[Column C]** | Sampling in Active Learning for Creation of Dataset in Column B | Average number of articles reviewed using Random Sampling for creating the Dataset in Column C | Percentage Improvement For Training Dataset Creation | Total Positive Sample After Active Learning **[Column B]** + **[Column C]** |
|---|---|---|---|---|---|---|---|
| AT | 2% | 7 | 13.50 | 30 | 44.23 | 47.43% | 20.5 |
| | 4% | 14 | 14.10 | 30 | 47.21 | 57.37% | 28.1 |
| | 6% | 21 | 14.50 | 30 | 47.45 | 58.17% | 35.5 |
| | 8% | 29 | 15.90 | 30 | 50.19 | 67.30% | 44.9 |
| | 10% | 36 | 16.50 | 30 | 53.87 | 79.57% | 52.5 |
| NSAID | 2% | 1 | 13.10 | 30 | 44.21 | 47.37% | 14.1 |
| | 4% | 3 | 15.65 | 30 | 50.22 | 67.40% | 18.65 |
| | 6% | 5 | 16.50 | 30 | 53.39 | 77.97% | 21.5 |
| | 8% | 7 | 17.75 | 30 | 56.34 | 87.80% | 24.75 |
| | 10% | 8 | 18.25 | 30 | 60.12 | 100.40% | 26.25 |
| ESTOR | 2% | 1 | 11.95 | 30 | 51.03 | 70.10% | 12.95 |
| | 4% | 3 | 13.50 | 30 | 54.98 | 83.27% | 16.5 |
| | 6% | 4 | 14.05 | 30 | 60.23 | 100.77% | 18.05 |
| | 8% | 6 | 15.51 | 30 | 62.04 | 106.80% | 21.51 |
| | 10% | 8 | 17.01 | 30 | 65.11 | 117.03% | 25.01 |

### *Experiment II: Performance of the proposed classifier*

After demonstrating the effectiveness of active learning for creating a training dataset, in this step we analyzed the effectiveness of our model and examined recall, precision, F1 score and work saved from our machine learning approach. The samples created in experiment I acts as the training dataset for this experiment. Here, we also analyzed how the measures varies when we change seed labeled set from 2%, 4%, 6%, 8% and 10%. Table 6 shows the experimental results of our model. Also, Figure 1 shows the line chart of how work reduction varies when we varies initial seeds (% labeled set) from 2% to 10%. In all datasets, we found that work reduction increases as we increase the initial seeds. Also, in terms of F1-score, as initial labeled set increases, F1-score also increases. For the ESTOR dataset F1 scores are 40.68% at 2% labeled set, 42.45 at 4% labeled set, 46.57 at 6% labeled set, 47.04 at 8% labeled set and 51.26 at 10 % labeled set. We can see the similar trends for the AT and NSAID datasets.

**Table 6: Experiment II results**

| Dataset | % Labeled (Seed) | No. Labeled (Seed) | N | TN | FP | FN | TP | Recall | Precision | F1 | WSS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AT | 2% | 7 | 20.5 | 159.5 | 590.5 | 41.1 | 314.9 | 88.45 | 34.94 | 50.02 | 6.59 |
| | 4% | 14 | 28.1 | 218.8 | 524.2 | 40.9 | 308.1 | 88.28 | 37.19 | 52.26 | 12.06 |
| | 6% | 21 | 35.5 | 238.2 | 497.8 | 43.6 | 298.4 | 87.25 | 37.63 | 52.51 | 13.39 |
| | 8% | 29 | 44.9 | 236.7 | 491.3 | 42.8 | 291.2 | 87.19 | 37.38 | 52.27 | 13.50 |
| | 10% | 36 | 52.5 | 287.4 | 433.6 | 46.5 | 280.5 | 85.78 | 39.45 | 53.99 | 17.64 |
| NSAID | 2% | 1 | 14.1 | 190.2 | 113.8 | 28.85 | 58.15 | 66.84 | 41.35 | 45.38 | 22.86 |
| | 4% | 3 | 18.65 | 186 | 116 | 25.35 | 59.65 | 70.18 | 40.57 | 45.42 | 24.79 |
| | 6% | 5 | 21.5 | 211.4 | 88.6 | 25.55 | 57.45 | 69.22 | 43.94 | 49.91 | 31.08 |
| | 8% | 7 | 24.75 | 183.45 | 114.55 | 17.05 | 63.95 | 78.95 | 38.28 | 50.12 | 31.85 |
| | 10% | 8 | 26.25 | 189.7 | 107.3 | 15.95 | 64.05 | 80.06 | 40.75 | 52.22 | 34.61 |
| ESTOR | 2% | 1 | 12.95 | 87.05 | 199.95 | 8.95 | 70.05 | 88.67 | 26.53 | 40.68 | 14.90 |
| | 4% | 3 | 16.5 | 111.1 | 173.9 | 9.45 | 67.55 | 87.73 | 27.97 | 42.45 | 21.02 |
| | 6% | 4 | 18.05 | 129.1 | 154.9 | 8.45 | 67.55 | 88.88 | 32.02 | 46.57 | 27.09 |
| | 8% | 6 | 21.51 | 137.8 | 144.2 | 9.25 | 64.75 | 87.50 | 32.56 | 47.04 | 28.81 |
| | 10% | 8 | 25.01 | 163.4 | 116.6 | 8.2 | 63.8 | 88.61 | 36.34 | 51.26 | 37.36 |

N= Number of positive sample for training, TN= True Negative, FP= False Positive, FN= False Negative, TP=True Positive, WSS= Work Saving by using machine learning approach

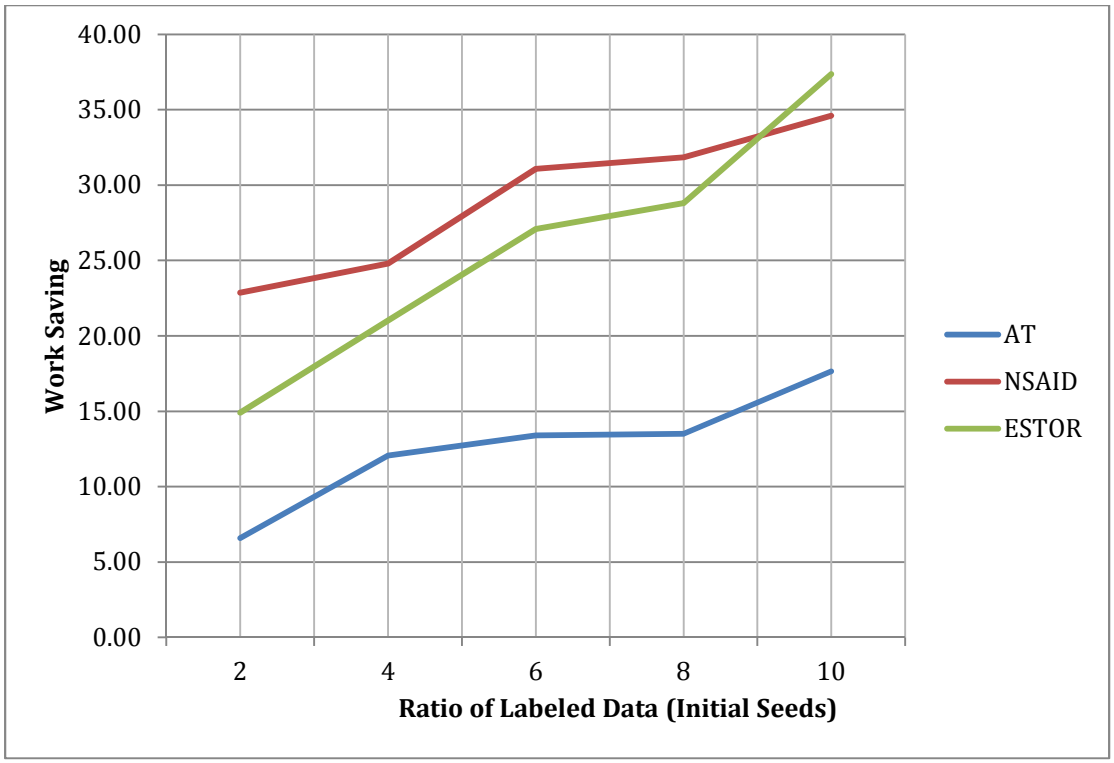


***Figure 1: Results Progression over Amount of Training Dataset***

## Conclusion

This research examined an automated method to classify relevant articles for inclusion or exclusion during the abstract triage stage for creating systematic reviews of medical research. We demonstrated that a less explored machine learning approach, namely active learning, is a viable technique for the problem where labeled articles are not available or very costly to obtain during SR creation.

From a theoretical perspective, this research explores the possibility of creating machine learning model with very few labeled instances. In prior research, supervised-learning has been used as the de-facto standard method for article classification for SRs, which however leaves the issue of a small-sized training dataset largely unaddressed. We propose to use active learning, which represent a novel approach that to our knowledge, has not been used in the area of SR creation. Among various active learning algorithms, we use label-spreading, which has shown exciting results in selecting articles during SR createion. After developing a training dataset that includes approximately 15%-20% positive instances, we use soft-margin SVM to create a classifier for creating systematic reviews. The experiences and lessons learned from this research are expected to inform the literature regarding the efficacy of the proposed techniques and the further development and refinement of these techniques.

From a practical and applied research perspective, this research is expected to result in a significant reduction in labor and other costs in connection with SR creation. Currently, laborious efforts for selecting articles for SRs preclude us from creating systematic reviews to keep pace with medical research advances, which subsequently impedes the translation of the latest medical evidence into healthcare practice. This research can help to automate the systematic review development process by significantly reducing the number of articles that scientists need to manually review, and has the potential to contribute to the adoption of evidence-based medicine. In summary, this research provides direct impact in the availability of best medical evidence and consequently, may contribute to improving the health and wellbeing of society.

The research can be further extended along a number of dimensions. First, the proposed approach can be further evaluated using additional datasets beyond the three datasets included in this research. Second,

future research may investigate means for deploying the proposed approach in a manner that simplifies and automates (or semi-automates) the creation of systematic review. Other integration and deployment possibilities include leveraging clinical trials documentation, e.g., from clinicaltrials.gov to further expedite the translation of medical research into practice.

In conclusion, this research attests to the potential of machine learning, text mining and big data analytics in supporting evidence-based medicine. It is a step towards closing the gap between research and practice in the quest toward providing higher quality healthcare outcomes at a reduced cost.

## REFERENCES


Adeva, G., Atxa, P., Carrillo, U., and Zengotitabengoa, A. 2014. " Automatic text classification to support systematic reviews in medicine," *Expert Systems with Applications* (41:4), pp 1498-1508.

Allen, I., and Olkin, I. 1999. "Estimating Time to Conduct a Meta□analysis From Number of Citations Retrieved," *JAMA* (282:7), pp 634-635.

Ananiadou, S., Procter, R., Rea, B., and Sasaki, Y. 2009. "Supporting Systematic Reviews using Text Mining," (3).

Anderson, J. R. 1983. "The architecture of cognition. Harvard Univ. press, Cambridge, MA,,").

Bekhuis, T., and Demner-Fushman, D. 2012. "Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers.," *Artificial intelligence in medicine* (55), pp 197-207.

Chapelle, O., Scholkopf, B., and Zien, A. 2006. *Semi-Supervised Learning*, (The MIT Press Cambridge, Massacgusetts.

Cohen, A. M., Adams, C. E., Davis, J. M., Yu, C., Yu, P. S., Meng, W., Duggan, L., Mcdonagh, M., and Smalheiser, N. R. 2010. "The Essential Role of Systematic Reviews , and the Need for Automated Text Mining Tools,"), pp 376-380.

Cohen, A. M., Ambert, K., and McDonagh, M. 2009. "Cross-topic learning for work prioritization in systematic review creation and update," *J Am Med Inform Assoc* (16:5) Sep-Oct, pp 690-704.

Cohen, A. M. C. 2014. "Systematic Drug Class Review Gold Standard Data."

Cohen, A. M. C., Ersh, W. R. H., Eterson, K. P., and En, P. O. I. N. Y. 2006a. "Reducing Workload in Systematic Review Preparation Using Automated Citation Classification,"), pp 206-219.

Cohen, A. M. C., Hersh, W. R., Peterson, K., and Yen, P.-Y. 2006b. "Reducing Workload in Systematic Review Preparation Using Automated Citation Classification," *JAMIA* (13:2), pp 206-219.

Cortes, C., and Vapnik, V. 1995. "Support-vector networks," *Machine learning* (20:3), pp 273-297.

Frunza, O., Inkpen, D., and Matwin, S. 2010. "Building Systematic Reviews Using Automatic Text Classification Techniques," *Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics*), pp 303-311.

Higgins, J., and Green, S. 2011. "Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]," *The Cochrane Collaboration*).

Joachims, T. 1998. "Text Categorization with Support Vector Machines : Learning with Many Relevant Features," *Universtat Dortmund*), pp 1-19.

Kim, S., Song, Y., Kim, K., Cha, J. W., and Lee, G. G. 2006. " Mmr-based active machine learning for bio named entity recognition," *In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics*), pp 69-72.

Lebanon, G., Mao, Y., and Dillon, J. 2007. "The Locally Weighted Bag of Words Framework for Document Representation," *The Journal of Machine Learning Research* (8).

Liu, H., Johnson, S. B., and Friedman, C. 2002. "Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS," *J Am Med Inform Assoc* (9:6) Nov-Dec, pp 621-636.

McGowan, J., and Sampson, M. 2005. "Systematic reviews need systematic searchers. ," *Journal of the Medical Library Association* (93:1), pp 74-80.

Mulrow, C. 1994. "Rationale for systematic reviews," *BMJ* (309), pp 597-599.

Robertson, S. 2004. "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of documentation* (60:5), pp 503-520.

Settles, B. 2010. "Active learning literature survey," *University of Wisconsin, Madison* (52:11), pp 55-66.

Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., Kelly, M. P., and Thomas, J. 2013. "Pinpointing needles in giant haystacks: use of text mining to reduce impractical

screening workload in extremely large scoping reviews," *Research Synthesis Methods*), pp n/a-n/a.

Shojania, K. G., Margaret Sampson, Ansari, M. T., and Garritty, C. 2007. "Updating Systematic Reviews," *AHRQ* (16).

Shrager, J., Hogg, T., and Huberman., B. A. 1987. "Observation of phase transitions in spreading activation networks," *Science* (236:1092–1094).

Stanford 2014. "Soft margin classification."

Tong, S., and Koller, D. 2002. "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research* (2), pp 45-66.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Sch¨olkopf, B. 2004. "Learning with Local and Global Consistency," *Max Planck Institute for Biological Cybernetics, 72076 Tuebingen, Germany*).

Zhu, X., and Ghahramani, Z. 2002. "Learning from labeled and unlabeled data with label propagation," *Technical Report CMU-CALD-02-107, Carnegie Mellon University*).