Inf Syst Front (2018) 20:195–207
DOI 10.1007/s10796-016-9724-0

A comparative analysis of semi-supervised learning: The case of article selection for medical systematic reviews

Jun Liu[1] · **Prem Timsina**[1] · **Omar El-Gayar**[1]

Published online: 22 November 2016

Abstract While systematic reviews 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 have 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 systematic reviews. However, this technique holds very little promise for creating new reviews because training data is rarely available when it comes to systematic creation. In this research we assess and compare the applicability of semi-supervised learning to overcome this labeling bottleneck and support the creation of systematic reviews. The results indicated that semi-supervised learning could significantly reduce the human effort and is a viable technique for automating medical systematic review creation with a small-sized training dataset.

Keywords Medical systematic reviews · Semi-supervised learning · Active learning · Self-training · Text mining · Text analytics

1 Introduction

The healthcare industry is undergoing dramatic transformation in promoting evidence-based medicine (EBM). It has been adopting the practice of generating evidence from experimental or quasi-experimental studies to inform clinicians and patients. Big data will play an important role in this transformation. Nowadays, most physicians struggle to stay current with the latest evidence guiding clinical practice (Murdoch and Detsky 2013). The digitization of medical literature has greatly improved access; however, the sheer number of studies makes knowledge translation difficult. Given a medical problem, even if a clinician had access to all relevant studies, sorting through this huge amount of information to gather relevant evidence and develop a reasonable treatment approach is a daunting task. Our research focuses on analytic techniques that address this big data challenge. More specifically, we investigate the use of semi-supervised learning to automate the process of selecting relevant studies/articles that should be included in a systematic review when a training dataset is not readily available.

Systematic reviews are a cornerstone of evidence-based medicine (Tsafnat et al. 2014). With the increasingly rapid pace by which medical knowledge is created, practitioners are challenged to keep pace with state-of-the-art medical evidence and incorporate such evidence into practice. Systematic reviews respond to this issue by recognizing, appraising, and synthesizing research-based evidence from multiple sources and translating it into practical guidelines. Each systematic review focuses on a particular research question and tries to synthesize and appraise all high quality research evidence relevant to that question in order to answer it. For example, the systematic review, "*Screening for Cognitive Impairment in Older Adults: A Systematic Review for the U.S. Preventive Services Task Force*" (Lin et al. 2013), aims to answer the question about the accuracy of brief cognitive screening instruments are in diagnosing cognatic impairment. After analyzing relevant studies, Lin et al. conclude that instruments to screen for cognitive impairment can adequately detect dementia, but there is no empirical evidence that screening improves decision making.

✉ Jun Liu
jun.liu@dsu.edu

[1] Dakota State University, Madison, SD, USA

Developing a medical systematic review is a demanding, rigorous, and resource-intensive process. The current workflow for creating systematic reviews is largely a manual process. It consists of 1) performing keyword search to identify potentially relevant articles, 2) performing article triage to identify articles for inclusion, and 3) finally, summarizing the selected studies via meta-analysis or other review methods. Within the workflow, article triage - identifying articles for inclusion in a systemic review - is particularly resource intensive. Specifically, articles are triaged in two steps (Shojania et al. 2007). The first step is called "abstract triage", where scientists often manually review the title and abstract of a large number of articles to identify "relevant" ones that can be potentially included in a systematic review. The second step, often referred to as "full-text triage", involves full text inspections of the articles selected in abstract triage to identify those that satisfy the inclusion criteria and will be included in a systematic review. The growing number of published studies imposes a significant screening workload on reviewers. An initial search by querying databases such as Medline, Cochrane and Embase often returns thousands or tens of thousands of articles given a review problem. For example, Lin et al. (2013) retrieved 16,179 articles based on keywords such as "cognitive impairment" and "cognitive impairment and older adults". The abstract triage process, where two scientists manually reviewed the title/abstract of the 16,179 articles, resulted in 1,190 articles. Finally, 253 articles were included in the systematic review after full-text triage of the 1,190 articles. Developing a systematic review 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 and Olkin 1999; McGowan and Sampson 2005). A large majority of the time and funds are spent on identifying "relevant" studies for inclusion in the review.

In that regard, various text mining methods have been proposed to automate the article screening for systematic reviews (Bekhuis and Demner-Fushman 2012; Shemilt et al. 2013; Adeva et al. 2014). These text mining methods have been proved to be very helpful during "abstract triage". The process of text mining in abstract triage starts from using the abstracts of thousands or tens of thousands of articles retrieved from medical databases as the corpus. Each document (i.e., a text file including the article tile and abstract) in the corpus is then pre-processed and represented by a vector of weights m features $d_j = (w_{1j}, w_{2j}, \ldots \ldots, w_{mj})$, where m is the number of features and w_i is the weight of the i^{th} features. Then, in almost all existing research that leverages text mining for abstract triage, supervised learning has been used to classify the documents into "relevant" and "irrelevant". Supervised learning assumes a readily available training dataset. For instance, Cohen et al. (2006) proposed a perceptron-based classifier that helps automatically identify relevant articles. The corpus used in the study includes 24 datasets on different medical topics collected by scientists at the Oregon Evidence-based Practice Center for the Drug Effectiveness Review Project (DERP). DERP scientists labeled each article in the datasets as "relevant" or "irrelevant" based on the abstracts alone. Only the articles labeled as "relevant" would be further reviewed in the full-text triage stage. Similarly, Adeva et al. (2014) used a dataset called Internet-Based Randomized Control Trial (IBRCT) mapping to compare various supervised learning algorithms for article selection. This IBRCT dataset consists of 1941 articles that were read and classified by a committee of experts into 510 relevant and 1431 irrelevant instances. Supervised learning relies on a large training dataset, which can be problematic in this context when we create a new systematic review, training data is rarely available. Cohen et al. (2006) admitted the problem and focused on predicting which new articles are most likely to include evidence warranting inclusion in a review update. According to Cohen et al. (Cohen et al. 2009), the procedures for creating and updating systematic reviews (SRs) are similar; however, one important difference is that an SR update already has a collection of included/excluded article judgments that are based on previous reviews. Due to the lack of considerable amounts of training data, supervised learning methods proposed in exiting research hold very little promise for systematic review creation. Given a medical problem, a keyword search often return thousands or tens thousands of articles. Labeling these articles to create a sufficiently large training dataset is difficult, laborious and time-consuming. Scientists can afford to create a small-sized training set. However, it is known that supervised learning with a small-sized training dataset often leads to an overly simple prediction function that may not be rich enough to capture the true underlying relationship.

In recent years, semi-supervised has received considerable attention in the area of data mining due to its potential for reducing the effort of labeling data. Semi-supervised learning falls between supervised and unsupervised learning techniques. It refers to the method of using a large unlabeled data set together a given labeled dataset during the training process (Wang et al. 2015). It is motivated by the fact that in many settings, unlabeled data is plentiful but labeled data is limited or expensive. Generally speaking, labeling data for a specific problem involves the input of a skilled human expert, or the execution of a physical experiment, both represent costly endeavors. Examples include areas ranging from enterprise document search to research in the humanities (i.e., history) to journalism (i.e., looking through past news stories to identify relevant past items on the same topic; on the consumer side, showing the reader relevant related stories to the article currently being read) as well as creating medical systematic reviews. When it comes to creating a new systematic review, labeled training data (i.e., articles that have been reviewed by

human experts) is mostly not readily available and is costly to obtain, requiring a manual review of thousands or even hundreds of thousands of articles. The goal of this research therefore is to assess and compare the performance of select semi-supervised learning methods for article selection for medical systematic reviews. More specifically, we plan to explore the ability of semi-supervised learning to overcome the labeling bottleneck and automate systematic review creation with a small-sized training dataset that includes, say, one or two hundred labeled articles. We perform comparative studies of various semi-supervised learning methods and identify the techniques suited for systematic review creation. To our knowledge, the proposed research is one of the first that conducts a comprehensive comparative analysis on the feasibility of using semi-supervised learning to address the small-sized training dataset problem that hampers the use of classification algorithms for medical systematic review creation. Further, the research provides insights into the relative performance of various semi-supervised learning that can be applicable to other domains such as noted earlier.

The following section presents related work and concludes with a list of research objectives. Section 3 presents various semi-supervised learning techniques utilized in this study, followed by a presentation and discussion of the results of various experiments in Section 4. Section 5 concludes the paper highlighting contributions, limitations and directions for future research.

2 Related work and research objectives

Nowadays, there are public databases such as a global network of Cochrane entities and a North American network of AHRQ-funded Evidence-based Practice Centers that enable scientists to access up-to-date medical research findings. Even so, developing a systematic review is slow. The average time to complete a systematic review is 2.4 years with a reported maximum of 9 years (Bekhuis and Demner-Fushman 2012). A bottleneck occurs during "abstract triage", where scientists screen the title and abstract of thousands or tens of thousands of articles for inclusion in a systematic review. Hence, most of the existing research has focused on automating abstract triage using supervised learning methods (e.g., Cohen et al. 2009; Frunza et al. 2010; Bekhuis and Demner-Fushman 2012; Shemilt et al. 2013). Cohen et al. (2006), in a National Institute of Health (NIH) supported project, developed- a perceptron-based classifiers to identify journal articles for inclusion in systematic review update, based on the title and abstract of the articles. In another study, Frunza et al. (2010) applied naïve Bayes to a dataset of 47,274 manually labeled article abstracts. They obtained very high recall values (up to 99 %) and moderately high precision of 63 %. In a recent study, Timsina et al. (2015) proposed a supervised learning based text mining method that employs the soft-margin polynomial Support Vector Machine (SVM) as a classifier, exploits Unified Medical Language Systems (UMLS) for medical terms extraction, and uses SMOTE sampling to resolve class imbalance issues that are ubiquitous for medical review datasets. There are also studies that focus on comparing multiple algorithms that can be used to classify articles for systematic reviews. For instance, Bekhuis and Demner-Fushman (2012) compared different supervised learning algorithms including K-nearest neighbor (K-NN), naïve Bayes, complement naïve Bayes (cNB), and evolutionary SVM (EvoSVM) for "abstract triage". The authors demonstrated that based on text mining techniques, the number of documents that need to be further manually screen 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). Timsina et al. (2015) compared different supervised algorithms including SVM, naïve Bayes, perceptron, etc., exploited Unified Medical Language Systems (UMLS) for medical terms extraction, and examined various techniques to resolve class imbalance that is common for systematic review datasets. Through an empirical study, they demonstrated that SVM with polynomial kernel achieves better classification performance than the other algorithms, and the performance of the classifier can be further improved by using UMLS to identify medical terms in articles and applying re-sampling methods to resolve the class imbalance issue. Adeva et al.'s (Adeva et al. 2014) conducted experiments that involved multiple classification supervised learning 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 produced the highest F-measure when applied to the titles/abstracts. All these studies developed supervised learning classifiers based on large training datasets with manually designated labels. As discussed previously, a conspicuous problem with the supervised learning based approach to article selection is that supervised learning, to be effective, requires large amounts of training data, which is often not readily available in most circumstances when we create a new systematic review. It is time-consuming and resource-intensive for scientists to screen thousands of articles (even just the title/abstract of the article) to create a large enough training dataset. In view of the problem, Cohen et al. (2006) suggested to focus on updating a review, where a reviewer already has a set of relevant documents in the form of the studies included in the original review.

Is it possible to develop a new systematic review without asking scientists to manually review thousands of articles? There are a few studies that attempted to provide feasible

solutions to the problem. Cohen et al. (2009) investigated whether a topic-specific automated document ranking system for systematic reviews (SRs) can be improved using a hybrid approach, combining topic-specific training data with data from other SR topics. The authors found that when topic-specific training data are scarce, leveraging training data previously used for developing systematic reviews for other related topics can significantly enhance the classification performance. There is also research that focuses on prioritizing the order in which citations (including titles, articles, keywords, etc.) will be screened. Thomas et al. (2011) suggested a possible method called "term recognition", which works by treating the included titles and abstracts as one big (and growing) document. This method can start with a relatively small number labeled articles. Each time another article is marked as "included", its text is added to the previously included titles and abstracts. The key terms from this string of text are then identified, and a search is carried out on the remaining titles and abstracts. The search is weighted by the significance attached to each term and the results ordered in terms of relevance. Thus, rather than viewing the documents in no particular order, those most similar to the studies already included are reviewed first. Unfortunately, no empirical results were presented on this "term recognition" method.

Overall, the findings of extant research indicate that supervised learning shows enough promise for automate the article selection process for systematic reviews if sufficient training instances are available. This is however a big "if" since developing a sufficiently large training set often requires screening the title/abstract of thousands of articles. Extensively studied in machine learning and applied to text classification, semi-supervised learning has been proved to effective in case of a small-sized training dataset (e.g., Song et al. (2011); Jin et al. (2011)). Nonetheless, little research to date has examined if semi-supervised learning can help truncate the costly and laborious article screening process for systematic reviews by requiring a small percentage of labeled instances. This leads us to the following research objectives: 1) Assess and compare the classification performance of various semi-supervised learning algorithms for systematic review article selection, and 2) determine if classification performance can be improved using wrapper methods such as "self-training" and "active-learning" with the best performing algorithm.

3 Article classification

We conducted three experiments using three systematic review datasets. Before we describe our experiments in detail, we first describe the data sources, the semi-supervised methods, and the evaluation metrics for article classification used in our research.

3.1 Datasets and data processing

We used three systematic review datasets on drug topics including AtypicalAntipsychotics (AT), NSAID, and Estrogens (ESTRO) collected by AHRQ's Evidence-based Practice Center (EPC) at Oregon Health and Science University in our research. These three systematic review datasets were also used in (Cohen et al. 2006). Table 1 shows an overview of the datasets. As discussed above, imbalanced class distributions are the norm for article selection in systematic reviews. As shown in Table 1, there are much more irrelevant articles than the relevant ones in all three datasets. Among the three dataset, Estrogens (ESTRO) has the most serious class imbalance problem with 29 % of articles labeled as "relevant".

We used the MEDLINE records for each of the articles in the above three datasets to generate the feature set as inputs to our classification technique. The feature set includes the features extracted from the title and abstract as well as the article's Medical Subject Headings (MeSH) and MEDLINE publication type. Following (Timsina et al. 2015), we used the Unified Medical Language System (UMLS) implemented within the software tool MetaMap version 4 to extract terms and use them as features. For instance, given the sentence "the objective of this study was to examine the relationships of serum and dietary magnesium (Mg) with prevalent cardiovascular disease", the MetaMap extracts the ULMS terms including "Study Objective", "Relationship", "Serum", "Dietary Magnesium", "Cardiovascular ", and "Disease prevalence" from the sentence. (Timsina et al. 2015) proved that while the majority of existing research used the "bag-of-words" approach in systematic review article screening, the automatically extracted Unified Medical Language System (UMLS) terms help boost classification performance. We then used the term frequency inverse document frequency (tf-idf) technique (Robertson 2004) to assign a weight to each UMLS term. Each document was represented by a vector consisting of the TF-IDF weights of the UMLS terms. TF-IDF of a term increases when the term appears more often in a document, but it is offset by the count of the term in the whole dataset, which mitigates for the fact that some words such as "patient" are generally more common than others in medical documents.

Table 1 Overview of datasets

Dataset	Total number of articles	Number of articles labeled as relevant	Number of articles labeled as irrelevant	Ratio—relevant vs. irrelevant
Antihistamines (AT)	1120	757	363	0.48
Estrogens (ESTRO)	370	289	81	0.28
NSAID	393	305	88	0.29

3.2 Semi-supervised learning methods

We investigate the following semi-supervised learning methods.

Label spreading (Zhou et al. 2004) Label Spreading assumes that geometrically closer data points tend to be similar. There are two general ideas related to label spreading: 1) the labeled examples act as sources that push out labels to unlabeled data, and 2) an example propagates its label to its neighboring examples according to their proximity to it. Label spreading proposed is a graph based semi-supervised learning technique that spreads the label information from a labeled data point to an unlabeled data point based on the affinity of the data points. Due to the smoothness constraints, reliable labels should reinforce each other, resulting in higher node weights, whereas labels showing inconsistencies tend to cancel out, resulting in lower node weights.

Label propagation (Zhu and Ghahramani 2002) Label propagation is similar to Label spreading in that both algorithms are graph-based, and both attempt to propagate a node's label to its neighboring nodes according to their proximity. The major difference between label propagation and label spreading is that label propagation uses the raw similarity matrix constructed from the data with no modifications, while label spreading iterates on a modified version of the original graph and normalizes the edge weights by computing the normalized graph Laplacian.

Semi-supervised support vector machine (S3VM) (Bennett and Demiriz 1999) S3VM, an extension of standard support vector machine with unlabeled samples, is another widely used semi-supervised learning technique. The goal of an S3VM classifier is to find a labeling of the unlabeled samples, so that a linear boundary has the maximum margin on both the original labeled samples and the (now labeled) unlabeled samples. The obtained decision boundary has the smallest generalization error bound on unlabeled samples. The main problem is that this the objective function is non-convex, which make optimization difficult (Zhu 2005).

We selected the above three semi-supervised learning algorithms because they have been widely used, and we have reliable implementations of them. Scikit-learn, a well-known machine learning toolkit, includes implementations of label spreading and label propagation. We used the S3VM implementation developed by (Gieseke et al. 2014).

In our research, we also considered two wrapper methods for semi-supervised learning: **Self-training** and **Active Learning**. They are wrapper methods because they "wrap" some existing classifiers. In self-training, an existing classifier (such as SVM) is first trained with the small amount of labeled data. The classifier is then used to classify the unlabeled data. Typically the most confident unlabeled points, together with their predicted labels, are added to the training set. The classifier is re-trained and the procedure repeated.

Active learning is a special type of semi-supervised learning. Active learning resembles self-training in that it also attempts to overcome the labeling bottleneck by identifying the most informative set of unlabeled instances based on some existing classifiers. It differs from self-training in that after selecting the most confident unlabeled samples, it requests an oracle (e.g., a human expert) to assign their labels. Active learning is also an iterative process in which it first train a classifier 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.

3.3 Evaluation

We evaluated article classification performance using the classical precision, recall, and F1 metrics. The formulas for computing recall, precision, and F1 are shown in Table 2. TP represents the number of True Positives, i.e., positive samples that were correctly classified. 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. 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 represents the harmonic mean of recall and precision. We did not use other widely-used metrics for classification such as accuracy or AUC (area under ROC curve) because 1) when the class distribution is imbalanced, the evaluation based on accuracy breaks down, and 2) classification accuracy assumes equal misclassification costs (for false positive and false negative errors). However, for systematic review article classification, the cost of false negative is high, since we need to guarantee that our classification technique should identify most, if not all, of the articles that should be included in a systematic review. Thus, a high recall is necessary for any classification technique to be useful.

Table 2 Evaluation metrics

Evaluation Metric	Formula
Recall	TP/(TP + FN)
Precision	TP/(TP + FP)
F1	(2*recall*precision)/(recall + precision)

4 Experiments

We conducted three experiments to evaluate the effectiveness of the various semi-supervised learning methods for article selection for systematic reviews. The datasets we used in the experiments are the three datasets we described in Table 1. The objectives of the experiments include:

Experiment 1: Assess and compare the classification performance (w.r.t. recall, precision, and F1) of various semi-supervised learning algorithms for systematic review article selection.

Experiment 2: Determine if classification performance can be improved using "self-training" with the best performing semi-supervised learning algorithm identified in Experiment 1.

Experiment 3: Determine if classification performance can be improved using "active learning" with the best performing semi-supervised learning algorithm identified in Experiment 1.

4.1 Experiment 1 – comparing different semi-supervised learning algorithms

In Experiment 1, we evaluated the effectiveness of three generic semi-supervised learning algorithms including label spreading, label propagation, and S3VM. We compared the performance of these semi-supervised learning algorithms with standard supervised SVM with polynomial kernel. SVM with polynomial kernel has been proved to achieve better performance than others in a recent study (Timsina et al. 2015) that compares a variety of supervised learning algorithms for article selection for systematic reviews.

4.1.1 Experiment design

We started with 5 % labeled articles as seeds or initial training instances. We conducted stratified sampling to make sure that there are 5 % of the positive instances and 5 % of the negative instances in the seeds. Using the 5 % seeds (i.e., initial labeled instances) as the training set and the rest 95 % samples as the test set, we conducted semi-supervised learning using the three different algorithms. Since the seeds were randomly sampled, this random sampling could have a substantial effect on the performance of the classifiers. Hence, for each algorithm, we conducted 50 trials to ensure the reliability of the results. We started with label spreading. In each trial, we first randomly selected 5 % seeds including 5 % of the positive instances and 5 % of the negative instances and then performed learning. We then averaged the results of 50 trials to generate the final results for the label spreading algorithm with 5 % seeds. This approach is consistent with an earlier approach used in literature (Zhu and Ghahramani 2002). For the other algorithms including label propagation, S3VM, supervised SVM, we did not re-select the seeds. Rather, we used the 5 % seeds that were selected in the 50 trials for label spreading to ensure that we compared the different algorithms using the same training and test sets. After getting the results with 5 % seeds, we increased the number of seeds to 10, 15, 20, 25, and 30 %. For each number of seeds, we again conducted 50 trials and reported the average.

4.1.2 Results and findings

Table 3 shows the results of Experiment 1 with the largest recall, precision and F1 scores for each dataset with a specific number of seeds being highlighted.

Among the three measures including recall, precision and F1, recall is probably the most important one in this context. Any automated system for identifying relevant articles must maintain a very high level of recall since ideally, a systematic review should include all articles that provide high quality evidence relevant to a topic. Any system with a low recall would be of little use (Matwin et al. 2010). Cohen et al. (2006) even assumed that a recall of about 0.95 is required for a classification system to identify an adequate fraction of the positive papers. Label spreading consistently achieved higher recall than the other algorithms across all three datasets. When applied to the dataset AT, label spreading obtained around 90 % recall with over 10 % seeds. For the dataset ESTRO, label spreading produced recall of 83.32 % with 10 % seeds and raised recall to 90 % with 20 % seeds and to 94.36 % with 30 % seeds. It also produced recall of 90.46 % with 20 % seeds and of 91.23 % with 30 % seeds for the dataset NSAID. Label propagation also achieved relatively high recall for all three datasets, but label spreading consistently achieved higher recall than label propagation. S3VM and SVM produced lower recall results than the two graph-based algorithms including label spreading and label propagation. S3VM produced higher recall than standard supervised SVM in all three datasets. It, however, failed to produce a recall that is high enough to make it a feasible method for article selection with a small-sized training set. The highest recall values yielded by S3VM for the three datasets include 81.81 % for AT, 85.52 % for ESTRO, and 84.69 % for NSAID.

Precision is still essential in this context, but it is only meaningful when high recall has been achieved. A higher precision means that the articles that are classified as "relevant" are indeed relevant, which means that a smaller number of articles needed to be manually reviewed. In this area, F1 is not as important a measure as it is in other contexts. F1 represents the harmonic mean of precision and recall. It hence assumes equal misclassification costs for false positive and false negative errors, but in the context of article selection

Table 3 Experiment 1 results

Data-set	Seed	Label Spreading			Label Propagation			S3VM			SVM		
		Recall (%)	Preci-sion (%)	F1 (%)	Recall (%)	Preci-sion (%)	F1 (%)	Recall (%)	Preci-sion (%)	F1 (%)	Recall (%)	Preci-sion (%)	F1 (%)
AT	5 %	79.87	39.52	52.88	76.64	34.17	49.14	80.90	36.72	50.24	45.18	40.98	46.06
	10 %	87.85	39.75	54.73	81.79	35.18	50.46	81.58	37.61	51.19	72.36	44.40	54.82
	15 %	89.43	39.02	54.33	84.73	34.27	49.90	81.81	38.30	51.86	72.59	45.25	55.67
	20 %	88.59	38.31	53.49	86.35	35.37	50.63	81.56	38.99	52.43	73.80	46.54	56.97
	25 %	89.61	37.25	52.62	88.32	35.70	50.87	81.20	39.51	52.83	74.52	47.69	58.04
	30 %	91.34	35.91	51.55	88.75	34.76	50.05	80.87	39.86	53.07	75.14	47.76	58.28
ESTRO	5 %	74.99	28.45	41.01	56.29	25.48	35.08	63.09	28.11	38.89	60.48	21.14	31.33
	10 %	83.32	29.98	43.88	81.36	24.78	37.99	69.56	30.45	42.35	80.72	26.54	39.94
	15 %	87.88	30.41	45.00	86.96	27.53	41.82	78.81	32.99	46.49	81.94	28.74	42.55
	20 %	90.00	30.11	44.96	86.74	28.94	43.40	81.86	33.52	47.53	83.32	29.89	44.00
	25 %	92.15	29.17	44.17	88.13	28.28	42.82	83.98	33.67	48.03	83.33	37.34	51.57
	30 %	94.36	28.38	43.50	89.02	28.06	42.67	85.52	33.58	48.17	83.36	36.34	51.01
NSAID	5 %	81.43	28.94	42.51	87.71	26.64	36.27	73.63	29.37	41.73	55.63	28.69	37.85
	10 %	86.45	29.59	43.92	86.55	25.66	39.68	77.47	30.23	43.21	76.35	30.98	44.07
	15 %	89.38	29.54	44.25	86.61	28.52	42.67	80.02	31.03	44.44	78.88	30.33	43.82
	20 %	90.46	29.91	44.96	86.35	28.84	43.06	82.11	31.49	45.25	80.64	31.84	45.65
	25 %	90.22	29.47	44.42	87.64	28.65	42.84	83.59	31.72	45.72	78.41	40.12	53.08
	30 %	91.23	29.16	44.20	90.32	28.47	42.81	84.69	31.79	45.98	78.79	43.28	55.87

for systematic reviews, an error of missing a relevant article (i.e., a false negative error) can be more expensive than an error of selecting an irrelevant article (i.e., a false positive error). After all, the articles selected by machine learning methods still need to be manually verified. Compared with label spreading, S3VM produced similar F1 scores to label spreading for two datasets (AT and NSAID) and higher F1 scores (46.49 % vs.45.00 % with 15 % seeds, and 48.17 % vs. 43.50 % with 30 % seeds) for ESTRO with >10 % seeds. Supervised SVM performed even better than S3VM in terms of precision and F1. For the dataset AT, it yielded over 47 % precision with 30 % seeds, roughly 10 % higher than those obtained by S3VM and label spreading. For the other two datasets (ESTRO and NSAID), the precision results and subsequently F1 scores obtained by supervised SVM underwent a jump between 20 % seeds and 25 % seeds, indicating that a supervised learning algorithm such as SVM requires a certain number of training instances (more than 20 % in this case) to take effect. Such a jump, however, did not occur to SVM's recall results. Even with 30 % seeds, SVM produced low recall results (75.14 % for AT, 83.36 % for ESTRO, and 78.79 % for NSAID).

In summary, the two graph-based semi-supervised methods, label spreading and label propagation, produced higher recall results than S3VM and SVM, while S3VM and SVM (with more than 20 % seeds) produced similar or higher precision results. It appeared that the graph-based methods and the SVM-based algorithms have both advantages and disadvantages. Further analysis showed that label spreading and label propagation produced a significantly larger number of true positives than S3VM and SVM, which means label spreading and label propagation were able to identify some positive instances (i.e., relevant articles) that were missed by S3VM and SVM. With a significantly larger number of true positives, label spreading and label propagation achieved higher recall values. On the other hand, label spreading and label propagation also made a significantly larger number of false positive errors than S3VM and SVM. A false positive error means that a negative instance (i.e., an irrelevant article) was falsely classified as positive (i.e., relevant). As a result, overall, label spreading and label propagation yielded a lower level of precision than S3VM and SVM. In the context of systematic reviews, high recall is a prioritized criterion for effective article classification algorithms. Precision is useful only when a high level of recall is obtained. We hence believe that in this context, the graph-based algorithms are preferred to the SVM based algorithms. Between the two graph-based methods, label spreading performed better than label propagation with respect to both recall and precision. A plausible reason can be label spreading minimizes a loss function that has regularization properties, as such it is often more robust to noise. Hence, among the three semi-supervised learning methods we investigated, label spreading appeared to be the optimal method for dealing with article selection for

systematic reviews with limited labeled instances. Moreover, the results of experiment 1 indicate that semi-supervised learning, more specifically label spreading, could be a viable method for article selection with limited labeled instances. In this context of systematic review article selection, a critical requirement for an automated classification algorithm to be feasible is that it must achieve a very high level of recall. Label spreading obtained high recall in all three datasets. It achieved over 90 % recall for AT and NSAID and about 95 % recall for ESTRO. It is noteworthy that compared with standardized supervised SVM, label spreading produced lower precision results. Although not as critical as recall in this context, lower precision signifies more false positive errors, which means that more irrelevant articles would be manually reviewed. We hence conducted the next two experiments, Experiment 2 and Experiment 3, to explore methods for further enhancing classification performance.

4.2 Experiment 2 – enhancing classification performance with self-training

The goal of this experiment is to investigate if combining label spreading with self-training and supervised SVM can improve precision while maintaining or even enhancing recall, thus helping further reduce workload for systematic review article selection. In Experiment 1, label spreading achieved recall of about 95 % for ESTRO and of over 90 % for the other two datasets. However, if we follow Cohen et al.'s requirement that a recall close to 95 % is imperative for classification algorithms, further improving recall is still necessary.

Self-training is a semi-supervised method that can be used to increment the training set. Given an initial training dataset, self-training relies on an existing algorithm to label some of the most confident unlabeled instances. It then adds the newly labeled instances to the training dataset and re-train the algorithm. This process can be iterated over the remaining unlabeled data. Supervised learning algorithms such as SVM have often been used in self-training to identify the most confident instances. In this experiment, we used label spreading, a semi-supervised algorithm, to select the optimal unlabeled instances. As shown in Experiment 1, label spreading produced much higher recall and identified more true positives than SVM with a small-sized training dataset.

4.2.1 Experiment design

We used different numbers of seeds (i.e., initially labeled articles) ranging from 5 to 30 %. Again, to alleviate the effect of random sampling, for each seed number, we conducted 50 trials. Using the seeds as the initial training dataset, we performed label spreading to classify the unlabeled instances. Label spreading computed a weight for each unlabeled instance. An unlabeled instance with a higher weight was considered more likely to be positive. We ranked the unlabeled instances according to the weights. We then selected a few top instances and a few bottom ones and incorporated them with their predicted labels into the training set. This completed one iteration of self-training. The incremented training dataset was used to re-train the label spreading algorithm in the next iteration. Among the three datasets, ESTRO and NSAID have similar numbers of positive instances and negative instances. We tested different numbers of iterations (from 4 to 12) for these two datasets, and in each iteration, we also tried to select from 10 (including top 5 and bottom 5 instances) to 20 instances (including top 10 and bottom 10 instances). It appeared that 9 iterations of self-training with top 8 and bottom 8 instances selected in each iteration produced the best performance for ESTRO and NSAID. The dataset AT has a much larger number of positive and negative instances. We hence conducted 18 iterations of self-training with top 8 and bottom 8 instances selected in each iteration to make sure that relatively similar percentages of new instances would be labeled and added to the training set across all three datasets. Table 3 below includes a column called "Final Train", which shows the final sizes of the incremented training datasets after the iterative self-training process. For instance, for the dataset AT, the initial training set included just the 5 % seeds. 18 iterations of self-training added 40.44 % instances to the training set, which resulted in a final training dataset that included 45.44 % (40.44 % new instances + 5 % seeds) of the total instances. With the final training set, we trained a supervised SVM classifier and classified the remaining unlabeled instances. Our self-training based method combines both semi-supervised learning and supervised learning. We were aware that existing studies such as (Cohen et al. 2006; Bekhuis and Demner-Fushman 2012) had shown a tendency for recall to decline when precision increases. Since Experiment 1 results showed that supervised SVM achieved lower recall but higher precision than label spreading, we decided to use SVM to training a portion of the unlabeled instances, which could potentially enhance precision but lower recall. We attempted to remedy this by using self-training to increment the training dataset. Our strategy hence included using self-training to increment the training set, in order to maintain a high level of recall, and using the incremented training set to train a supervised SVM learner, in order to enhance precision.

4.2.2 Results and findings

We compared the performance of self-training with that of using label spreading alone and of using SVM. Table 4 shows the results, with the largest recall, precision and F1 scores for each dataset with a specific number of seeds being highlighted.

Table 4 Experiment 2 results

Dataset	Seed	Self-training				Label Spreading			SVM		
		Final Train*	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
AT	5 %	45.44 %	80.26 %	45.26 %	57.88 %	79.87 %	39.52 %	52.88 %	45.18 %	46.98 %	46.06 %
	10 %	52.84 %	85.58 %	45.08 %	58.81 %	84.85 %	39.75 %	54.73 %	72.36 %	44.40 %	55.03 %
	15 %	60.24 %	85.42 %	45.42 %	59.31 %	88.63 %	39.02 %	54.33 %	72.59 %	45.25 %	55.75 %
	20 %	67.64 %	87.33 %	45.33 %	59.68 %	88.59 %	38.31 %	53.49 %	73.80 %	46.54 %	57.08 %
	25 %	75.03 %	88.89 %	44.89 %	59.65 %	89.61 %	37.25 %	52.62 %	74.52 %	47.69 %	58.16 %
	30 %	82.43 %	90.17 %	45.34 %	60.34 %	91.34 %	35.91 %	51.55 %	75.14 %	47.76 %	58.40 %
ESTRO	5 %	54.67 %	86.74 %	29.35 %	43.86 %	74.99 %	28.45 %	41.01 %	60.48 %	21.14 %	31.33 %
	10 %	59.86 %	89.22 %	30.93 %	45.94 %	83.32 %	29.98 %	43.88 %	80.72 %	26.54 %	39.94 %
	15 %	64.71 %	90.87 %	35.64 %	51.20 %	87.88 %	30.41 %	45.00 %	81.94 %	28.74 %	42.55 %
	20 %	69.90 %	89.53 %	36.44 %	51.80 %	90.00 %	30.11 %	44.96 %	83.33 %	29.89 %	44.00 %
	25 %	74.74 %	92.16 %	38.38 %	54.24 %	92.15 %	29.17 %	44.17 %	83.32 %	37.34 %	51.57 %
	30 %	79.93 %	93.75 %	38.66 %	54.74 %	94.36 %	28.38 %	43.50 %	83.36 %	36.74 %	51.01 %
NSAID	5 %	52.13 %	82.60 %	30.35 %	44.40 %	81.43 %	28.94 %	42.51 %	55.63 %	29.37 %	38.44 %
	10 %	57.38 %	86.37 %	30.93 %	45.41 %	86.45 %	29.59 %	43.92 %	76.35 %	30.98 %	44.07 %
	15 %	62.30 %	86.17 %	32.64 %	47.35 %	89.38 %	29.54 %	44.25 %	78.88 %	30.33 %	43.82 %
	20 %	67.21 %	89.11 %	38.44 %	53.71 %	90.46 %	29.91 %	44.96 %	80.64 %	31.84 %	45.65 %
	25 %	72.13 %	89.56 %	43.38 %	58.45 %	90.22 %	29.47 %	44.43 %	78.41 %	40.12 %	53.08 %
	30 %	77.38 %	90.78 %	43.66 %	58.97 %	91.23 %	29.16 %	44.19 %	78.79 %	43.28 %	55.87 %

Obviously, our strategy of combining semi-supervised learning and supervised learning has been proved to be effective in enhancing precision. Compared with using label spreading alone, our self-training method produced significantly higher precision for all three datasets. For instance, for the dataset AT and ESTRO, self-training with 30 % seeds produced precision that is about 10 % higher than the precision obtained by label spreading alone. For the dataset NSAID, self-training with 30 % seeds produced precision of 43.66 %, while label spreading with the same seeds produced precision of only 29.16 %. Self-training also yielded very comparable precision results to SVM. Our strategy was also effective in maintaining a high level of recall. It worked especial well with a small number of seeds. For the dataset AT, with 5 and 10 % seeds, self-training achieved higher recall (80.26 % vs. 79.87 % for 5 % seeds and 85.58 % vs. 84.85) than label spreading alone. For ESTRO with 5, 10, and 15 % seeds and for NSAID with 5 % seeds, self-training also yielded slightly higher recall. When the number of seeds got larger, self-training obtained slightly lower recall than label spreading alone.

To summarize, in Experiment 2, we aimed to enhance precision while maintaining or, better, improving recall. We used self-training with label spreading to identify the most confident unlabeled instances. These instances with their predicted labels were incorporated into the training dataset, and with the incremented training set, we employed SVM to classify the remaining unlabeled instances. The self-training based method succeeded in enhancing precision and maintaining a high level of recall. It, however, failed to further enhance recall. A reason could be that even if we chose to add the most confident instances in self-training, some instances were still misclassified. In Experiment 2, across the three datasets, we labeled 1800 unlabeled instances as positive. We made 177 (or 9.83 %) false positive errors. Our self-training method was much more effective in identifying negative instances, probably because our datasets are imbalanced, i.e., there are far fewer "relevant" than "irrelevant" instances in all three datasets. Among 1800 instances labeled as negative in the self-training process, only 28 (1.56 %) were misclassified. A serious limitation of self-training is that these misclassified instances were treated as truth and were used to classify other unlabeled instances. The impact of these misclassified instances could snowball as the self-training process proceeded. We hence continued to explore the effectiveness of active learning. We expected that with human labeled instances incorporated into the training dataset, we could enhance both recall and precision.

4.3 Experiment 3 – enhancing classification performance with active learning

Active learning approach has received considerable attention due to its potential for achieving greater classification accuracy in applications where unlabeled data may be abundant or easily obtained, but labels are difficult, time-consuming, or

expensive to obtain (Settles 2010). Active learning is similar to self-training in that the learner is responsible for acquiring training samples. The main difference of active learning from self-training is that in active learning, after an optimal set of unlabeled instances were identified, human experts need to label these instances. In this experiment, we wanted to investigate whether active learning based on label spreading can further enhance the performance of article classification, as compared with the fully automated approaches such as the self-training method described above.

4.3.1 Experiment design

For each dataset, we again used different numbers of seeds. Again, to alleviate the effect of random sampling, given a specific number of seeds, we conducted 50 trials and took the average of the results. In each trial, we performed active learning iteratively. We conducted 9 iterations of active learning for the datasets NSAID and ESTRO and 17 iterations for the dataset AT. In each iteration, we added 6 articles predicted by the algorithm to be negative and another 6 articles predicted to be positive to the labeled set. We conducted multiple tests to identify these optimum parameters such as the number of iterations and the number of instances added to the training set. As discussed previously, there are more negative instance 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). Our datasets indeed included much fewer "relevant" articles than "irrelevant" ones. The Experiment 2 results showed that label spreading is effective in identify negative instances, with only misclassified 1.56 % negative instances. Thus, in our active learning method, we added the instances predicted by the label spreading algorithm to be negative into the labeled set without asking human experts to annotate them. Positive articles, on the other hand, are fewer, and label spreading identified them with a higher misclassification rate in Experiment 2. In real practice, it is necessary for human experts to label the articles that were predicted to be positive, before adding them to the training dataset. In our experiment, since the actual label of each instance is available in our datasets, we simply added the instances with their correct labels to the training dataset. Like in Experiment 2, we used active learning to increment the training dataset iteratively. With the final incremented training set, we trained a SVM classifier, which was then used to classify the remaining unlabeled instance.

The sizes of the final training datasets after the iterative active learning process are shown in the column "Total Article Read" in Table 5 below. Each final training dataset after active learning included the initial seeds and the newly labeled instances. In real practice, both the seeds and the instances labeled during active learning represent manually reviewed instances. We used the self-training method described in section 4.2 and supervised SVM as the benchmark methods. We conducted self-training and supervised SVM classification with an initial training dataset that included the same number of instances as in the training set obtained by active learning. For instance, for the dataset AT with 5 % seeds, the augmented training dataset after active learning encompassed 26.43 % of the instances, which included 5 % seeds plus 21.43 % newly labeled articles – these are articles supposedly reviewed by human experts. When we conducted self-training using the method described in section 4.2 for comparison, we also created an initial training set that contained 26.43 % instances (including 5 % seeds and another 21.43 % stratified samples). By doing this, we made sure that we compared actively learning and self-training based on an equal number of supposedly manually reviewed articles. We conducted supervised SVM classification using the same initial training set prepared for self-training.

4.3.2 Results and findings

We compared the active learning method with supervised SVM and the self-training method described in section 4.2. Table 5 shows the comparison results.

Table 5 shows that the active learning method produced considerably better recall and precision than both self-training and supervised SVM. It worked well even with a small number of seeds. For instance, with 10 % seeds (around 31 % of total instances read), the active learning method produced recall of 91.50 % for AT, of 95.87 % for ESTRO and of 92.94 % for NSAID. We also included SVM classification results with 70 % training datasets in Table 5. Active learning even outperformed SVM with 70 % training sets. They have comparable precision results. However, SVM, even with a large training set, still made quite some false negative errors and produced a level of recall, which indicates that SVM needs to be adapted or extended to be useful for systematic review article selection. Another contributing factor that led active learning to outperform SVM could be that in each iteration of active learning, we selected roughly an equal number of positive vs. negative instances. In other words, the proposed active learning method implicitly performed under-sampling. Timsina et al. (2015) proved that since a typical systematic review dataset includes much fewer relevant articles than irrelevant ones, employment of re-sampling methods dealing with class imbalance such as under-sampling can significantly improve the performance of machine learning classifiers. Active learning appears to work especially well for medical systematic review datasets that normally have class imbalance. In active learning, label spreading is first used to identify the same number of articles that are most likely to be relevant vs. irrelevant, and human experts then manually verify the relevance of these articles. Due to class imbalance typical to medical review datasets,

Table 5 Experiment 3 results

Dataset	Seed	Total Article Read	Active Learning			Self-training			SVM		
			Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
AT	5 %	26.43 %	89.50 %	50.54 %	64.60 %	88.98 %	44.92 %	59.70 %	74.52 %	47.69 %	58.16 %
	10 %	31.43 %	91.50 %	49.52 %	64.26 %	91.17 %	45.68 %	60.86 %	75.14 %	47.76 %	58.40 %
	15 %	36.43 %	90.40 %	51.12 %	65.31 %	90.17 %	45.34 %	60.34 %	75.87 %	47.78 %	58.63 %
	20 %	41.43 %	91.42 %	51.53 %	65.91 %	90.76 %	46.29 %	61.31 %	75.76 %	47.66 %	58.51 %
	25 %	46.43 %	92.74 %	52.26 %	66.85 %	91.37 %	46.49 %	61.62 %	76.88 %	48.22 %	59.27 %
	30 %	51.43 %	93.18 %	52.81 %	67.41 %	91.92 %	46.38 %	61.65 %	78.63 %	47.73 %	59.40 %
	70 %								89.95 %	54.62 %	67.97 %
ESTRO	5 %	27.84 %	93.89 %	41.82 %	57.87 %	92.87 %	38.60 %	54.53 %	86.36 %	37.34 %	52.14 %
	10 %	32.70 %	95.87 %	42.49 %	58.88 %	93.97 %	38.97 %	55.09 %	83.32 %	36.74 %	50.99 %
	15 %	37.84 %	96.33 %	42.67 %	59.14 %	94.22 %	38.20 %	54.36 %	83.48 %	36.69 %	50.98 %
	20 %	42.70 %	96.59 %	41.68 %	58.24 %	94.69 %	39.14 %	55.39 %	85.66 %	36.93 %	51.61 %
	25 %	47.84 %	97.56 %	41.95 %	58.67 %	95.16 %	39.18 %	55.50 %	84.36 %	37.19 %	51.62 %
	30 %	52.70 %	98.06 %	42.77 %	59.56 %	95.64 %	39.17 %	55.57 %	82.85 %	39.88 %	53.84 %
	70 %								93.38 %	43.43 %	59.29 %
NSAID	5 %	26.46 %	91.60 %	48.02 %	63.01 %	89.76 %	44.02 %	59.07 %	78.41 %	40.12 %	53.08 %
	10 %	31.30 %	92.94 %	49.07 %	64.23 %	90.77 %	44.53 %	59.75 %	78.79 %	43.28 %	55.87 %
	15 %	36.39 %	93.34 %	49.94 %	65.91 %	91.14 %	45.16 %	60.40 %	77.97 %	44.68 %	56.81 %
	20 %	41.48 %	94.01 %	46.53 %	62.25 %	91.51 %	45.61 %	60.87 %	77.10 %	44.97 %	56.81 %
	25 %	46.31 %	94.44 %	50.32 %	65.66 %	91.87 %	45.37 %	60.74 %	78.40 %	43.17 %	55.68 %
	30 %	51.40 %	94.90 %	50.14 %	65.61 %	92.24 %	46.53 %	61.85 %	79.02 %	43.85 %	56.40 %
	70 %								90.48 %	51.51 %	65.65 %

i.e., a medical review dataset often includes a small number of relevant articles and a large number of irrelevant articles, the incremented training dataset resulting from several iterations of active learning includes a large portion of all the relevant articles and a relatively small portion of all the irrelevant articles. As a result, with the same number of articles being manually reviewed, active learning helps identify more relevant articles and achieve higher recall than supervised learning with random samples.

In summary, we conducted three experiments, each of which shed some light on the use of semi-supervised learning in selecting articles for systematic reviews. The Experiment 1 results showed that given a small-sized training dataset, semi-supervised methods, especially label spreading, achieved a high level of recall, which makes them viable methods for reducing workload for systematic review article selection. However, using label spreading alone resulted in low precision. To improve precision while maintaining or better enhancing recall, we proposed a self-training based method that combines semi-supervised learning (with label spreading based self-training) and supervised learning (with SVM). The Experiment 2 results showed that the proposed self-training based method significantly enhanced precision while maintaining a high level of recall. It worked especially well with small training sets (5 % or 10 % seeds). Next, we explored the feasibility of using active learning to further enhance both recall and precision. The Experiment 3 results showed that active learning produced a very high level of recall that meets Cohen et al.'s 95 % recall requirement, suggesting that the active learning method is a highly feasible method for systematic review article selection with small-sized training datasets. However, active learning requires human expert to be continuously engaged to produce optimum results. If experts' engagement is not available, with an initial small-sized training set, self-training provides a feasible alternative. It is automatic, though the classification performance of self-training is inferior to that of active learning.

5 Conclusion

Developing a medical systematic review involves a group of scientists evaluating thousands or even hundreds of thousands of articles in order to identify the relevant ones that need to be included in a review, which hence poses a Big Data challenge. This paper presents a comprehensive study assessing and comparing the applicability of using semi-supervised learning in addressing the challenge. We examined several different semi-supervised methods and identified label spreading as an algorithm that produced high recall that is necessary for

systematic review article selection. We also demonstrated that the performance of label spreading could be further enhanced when it is combined with self-training and active learning.

In prior research, supervised-learning has been used as the de-facto standard method for article classification for systematic reviews. Supervised learning, however, relies on a large training dataset that in real practice is extremely costly and time-consuming to obtain. We proposed to use semi-supervised learning methods such as label spreading, self-training, and active learning to classify articles based on a small-sized training dataset. The use of semi-supervised learning in selecting systematic review articles has so far been largely ignored in literature.

Moreover, the experiences and lessons learned from our research are expected to inform the literature regarding the efficacy of the proposed techniques and the further development and refinement of these techniques not just in the context of medical systematic reviews but in other domains such as enterprise document searching, research in the humanities (i.e., history) to journalism (i.e., looking through past news stories to identify relevant past items on the same topic; on the consumer side, showing the reader relevant related stories to the article currently being read).

With respect to medical systematic reviews, this research has the potential to optimize systematic creation and contribute to the adoption of evidence-based medicine. Currently, laborious efforts for selecting articles for systematic reviews 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 when they create a new systematic review. This research provides direct impact in the availability of best medical evidence and consequently, may contribute to improving the health and wellbeing of society.

As for limitations and directions for future research, we note the following: First, the viability of semi-supervised learning and wrapper methods was demonstrated using three data sets from the medical domain. Future research can further explore the generalizability of the results to other data sets from the medical and other domains. Second, with respect to medical systematic reviews, this research focused on the first step in conducting systematic reviews, namely, abstract triage. This approach can be extended to assess applicability to full-text triage leveraging existing and emerging to analyze not only the abstracts of tens of thousands of articles but also the full text of the articles. Last but not least, future research may investigate means for deploying the proposed approach in a manner that simplifies and automates (or semi-automate) the update of systematic reviews on a frequent basis as new literature is added to the existing knowledge repository.

References

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

Allen, I., & Olkin, I. (1999). Estimating time to conduct a meta-analysis from number of citations retrieved. *JAMA, 282*(7), 634–635.

Bekhuis, T., & Demner-Fushman, D. (2012). Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artificial Intelligence in Medicine, 55*, 197–207.

Bennett, K. and Demiriz, A. (1999). Semi-supervised support vector machines. *Advances in Neural Information processing systems*: 368–374.

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association, 13*(2), 206–219.

Cohen, A. M., Ambert, K., & McDonagh, M. (2009). Cross-topic learning for work prioritization in systematic review creation and update. *Journal of the American Medical Informatics Association, 16*(5), 690–704.

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: 303–311.

Gieseke, F., Airola, A., Pahikkala, T., & Kramer, O. (2014). Fast and simple gradient-based optimization for semi-supervised support vector machines. *Neurocomputing, 123*, 23–32.

Jin, Y., Huang, C., & Zhao, L. (2011). A semi-supervised learning algorithm based on modified self-training SVM. *Journal of Computers, 6*(7), 1438–1443.

Lin, J. S., O'Connor, E., Rossom, R. C., Perdue, L. A., & Eckstrom, E. (2013). Screening for cognitive impairment in older adults: a systematic review for the U.S. preventive services task force. *Annals of Internal Medicine, 159*(9), 601–612.

Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., & O'Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association, 17*(4), 446–453.

McGowan, J., & Sampson, M. (2005). Systematic reviews need systematic searchers. *Journal of the Medical Library Association, 93*(1), 74–80.

Murdoch, T., & Detsky, A. (2013). The inevitable application of big data to health care. *JAMA, 309*(13), 1351–1352.

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

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

Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., Kelly, M. P., & 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, 5*(1), 31–49.

Shojania, K. G., Sampson, M., Ansari, M. T. and Garritty, C. (2007). Updating Systematic Reviews. Publication No. AHRQ 07–0087, Rockville, MD, Agency for Healthcare Research and Quality.

Song, M., Yu, H. and Han, W. S. (2011). Combining active learning and semi-supervised learning techniques to extract protein interaction sentences. *BMC bioinformatics 12*.

Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods, 2*(1), 1–14.

Timsina, P., Liu, J. and El-Gayar, O. (2015). Advanced analytics for the automation of medical systematic reviews. *Information Systems Frontiers* (A Special Issue on Big Data and Analytics in Healthcare): 1–16.

Tsafnat, G., Glasziou, P., Choong, M., Dunn, A., Galgani, F., & Coiera, E. (2014). Systematic review automation technologies. *Systematic Reviews, 3*, 74.

Wang, S., Li, D., Petrick, N., Sahiner, B., Linguraru, M. G., & Summersa, R. M. (2015). Optimizing area under the ROC curve using semi-supervised learning. *Pattern Recognition, 48*(1), 276–287.

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

Zhu, X. (2005). Semi-supervised learning literature survey. TR-1530, University of Wisconsin-Madison, Department of Computer Science.

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

Jun Liu is an assistant professor in information systems in the College of Business & Information System, Dakota State University. He obtained his Ph.D. and M.Sc. in Management Information Systems from the Eller College of Management, University of Arizona. His research interests include data and text mining, social network analysis, data provenance, examining user collaboration in open source environments such as Wikipedia, and using technology to support business intelligence and decision-making. He has published several research papers in internationally refereed journals such as Information SystemsFrontiers, ACM Transactions on Management Information Systems, Journal of Data Semantics, Journal of Computing Science and Engineering, International Journal of Intelligent Information Technologies, Lecture Notes in Computer Science, etc. and has presented several papers at several international conferences.

Prem Timsina obtained his doctorate from Dakota State University. His research interests include machine learning, big data, data and text mining, health analytics, and leveraging data analytics to support business intelligence and decision-making. He has published several articles in international journals like Information System Frontiers, International Journal of Medical Informatics, and given various talks in conferences like Americas Conference on Information Systems, and Hawaii International Conference on System Sciences.

Omar El-Gayar, Ph.D. is the Dean for the College of Information Technology, United Arab Emirates University (UAEU). Prior to joining the UAEU, Dr. El-Gayar served as a Professor of Information Systems and Dean of Graduate Studies and Research, Dakota State University. His research interests include: analytics, business intelligence, and decision support with applications in problem domain areas such as healthcare, environmental management, and security planning and management. His inter-disciplinary educational background and training is in information technology, computer science, economics, and operations research. Dr. El-Gayar's industry experience includes working as an analyst, modeler, andprogrammer. His numerous publications appear in various information technology related fields. He is a member of AIS, ACM, INFORMS, and DSI.