Artificial Intelligence in Medicine 64 (2015) 89–103

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Automatic evidence quality prediction to support evidence-based decision making

Abeed Sarker[a,*], Diego Mollá[a], Cécile Paris[b]

[a] Department of Computing, Macquarie University, Sydney, NSW 2109, Australia
[b] Commonwealth Scientific and Industrial Research Organisation, Crn Vimiera and Pembroke Roads, Marsfield, NSW 2122, Australia

ARTICLE INFO

Article history:
Received 3 July 2014
Received in revised form 31 March 2015
Accepted 15 April 2015

Keywords:
Automatic text classification
Automatic medical evidence classification
Decision support system
Medical natural language processing
Evidence-based medicine

ABSTRACT

Background: Evidence-based medicine practice requires practitioners to obtain the best available medical evidence, and appraise the quality of the evidence when making clinical decisions. Primarily due to the plethora of electronically available data from the medical literature, the manual appraisal of the quality of evidence is a time-consuming process. We present a fully automatic approach for predicting the quality of medical evidence in order to aid practitioners at point-of-care.
Methods: Our approach extracts relevant information from medical article abstracts and utilises data from a specialised corpus to apply supervised machine learning for the prediction of the quality grades. Following an in-depth analysis of the usefulness of features (*e.g.*, publication types of articles), they are extracted from the text via rule-based approaches and from the meta-data associated with the articles, and then applied in the supervised classification model. We propose the use of a highly scalable and portable approach using a sequence of high precision classifiers, and introduce a simple evaluation metric called average error distance (AED) that simplifies the comparison of systems. We also perform elaborate human evaluations to compare the performance of our system against human judgments.
Results: We test and evaluate our approaches on a publicly available, specialised, annotated corpus containing 1132 evidence-based recommendations. Our rule-based approach performs exceptionally well at the automatic extraction of publication types of articles, with *F*-scores of up to 0.99 for high-quality publication types. For evidence quality classification, our approach obtains an accuracy of 63.84% and an AED of 0.271. The human evaluations show that the performance of our system, in terms of AED and accuracy, is comparable to the performance of humans on the same data.
Conclusions: The experiments suggest that our structured text classification framework achieves evaluation results comparable to those of human performance. Our overall classification approach and evaluation technique are also highly portable and can be used for various evidence grading scales.

1. Introduction

Evidence-based medicine (EBM) is a practice that requires medical practitioners to obtain the best quality clinical evidence from published research when answering clinical queries, in addition to using their own expertise. It has been described as "*the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients*" [1]. To use the best available medical evidence for solving patients' problems, practitioners are required to perform a number of steps including searching for evidence, selecting the best available evidence, extracting relevant information, and appraising the quality of the extracted evidence in the light of the patients' problems. Currently, the process of evidence-based answer generation is a manual process and primarily due to the plethora of electronically available medical documents, practitioners generally face the problem of information overload. Research has shown that practitioners often fail to pursue evidence-based answers to their clinical queries, particularly at point-of-care, due to time constraints [2]. The time associated with seeking and appraising information is largely considered to be the biggest obstacle in EBM practice [3–10]. As such, approaches that can extract relevant information from medical text, and utilise them to automatically perform some of the tasks associated with

[*] Corresponding author at: Department of Biomedical Informatics, Arizona State University, 13212 East Shea Boulevard, Scottsdale, AZ 85259, USA.
Tel.: +1 480 884 0349.
E-mail address: abeed.sarker@asu.edu (A. Sarker).

http://dx.doi.org/10.1016/j.artmed.2015.04.001
0933-3657/© 2015 Elsevier B.V. All rights reserved.

evidence-based decision making, can significantly aid the practice.

The appraisal of the quality of the extracted evidence is a crucial task in the process of evidence-based answer generation, and its purpose is to indicate the reliability of the recommendations that are made based on the available evidence. The quality of the best available evidence may depend on a large number of factors. For example, it may depend on the topic. The reliability of the evidence associated with different topics may vary depending on the amount of research the topics have received. Topics that have received more research attention in the past are likely to contain better quality evidence (*e.g.*, *safe behavioural interventions for obesity*), compared to topics that have received little (*e.g.*, *duration of steroid therapy for contact dermatitis*). Also, sometimes findings from different studies are not consistent, making the evidence unreliable. When making evidence-based recommendations, practitioners have to take these and other factors into account in order to assess the reliability of the extracted evidence. Thus, when extracting evidence from medical publications regarding a topic, practitioners also have to spend significant amounts of time to appraise the quality of the evidence associated with the topic.

In this paper, we describe an approach to automate the process of appraising the quality of the evidence. Our approach attempts to extract relevant information from medical abstract texts and the associated meta-data, and utilise the information to predict the quality of the evidence presented by the data. We apply natural language processing (NLP) techniques to extract features from the texts, and use the features in a supervised machine learning model to perform the quality predictions. Using a corpus that specialises in EBM question answering, we first perform an analysis of the features that are likely to be indicative of the quality of evidence. Following the analysis and the selection of the features, we apply a sequential classification model to automatically predict the quality of evidence on a discrete scale. Our approach achieves an accuracy of 62.84% when evaluated against a gold standard. Our evaluations also show that the difference between the performance of our system and that of human experts on the same data is not statistically significant.

The rest of the paper is organised as follows. We provide background on evidence appraisal including a discussion of the discrete scale that we use, and discuss some related research in Section 2. In Section 3, we discuss the data, our preliminary analysis of features, the fully automatic grade classification model, and our human evaluation experiments. In Section 4, we present the results of all our experiments along with discussions of the results. We conclude the paper in Section 5.

2. Background and related work

Due to the importance of appraising and specifying the quality of evidence in EBM practice, standardised grading scales have been proposed in the literature. Various organisations and publications have their own measure of evidence and, according to a research report produced by the Agency of Healthcare Research and Quality [11], more than 100 evidence grading scales are in use today. The report also proposes that any system for grading the strength of evidence should consider three key elements: quality (the extent to which the identified studies minimise the opportunity for bias), quantity (the number of studies and subjects included in those studies) and consistency (the extent to which findings are similar between different studies on the same topic). Among other requirements, studies have specified the need for a balance between simplicity (such that assessing the quality of evidence is not very time-consuming) and clarity (so that evidence can be easily classified into a specific grade) [12]. Comprehensiveness of grading systems is also seen as an important factor [13] since they need to be applied to studies of screening, diagnosis, prevention, therapy and prognosis. Based on these requirements, we chose the strength of recommendation taxonomy (SORT) [13] as our target grading scale. SORT was designed to provide a uniform recommendation-rating system that could be applied throughout the medicine literature. It is simple and straightforward, and, therefore, easy for practitioners to use during everyday practice. This taxonomy uses only three ratings – A (strong), B (moderate) and C (weak) – to specify the *strength of recommendation* of a body of evidence. Furthermore, the availability of a specialised corpus [14] that uses SORT as the target scale for quality prediction/grading makes this scale an ideal choice for our research. The corpus, described in the next section, enables us to compare the automatically generated evidence grades to grades assigned by human experts, and evaluate the performance of our system.

Research related to ours has focused mostly on text classification in the medical domain and automatic quality assessment of medical publications. Text classification techniques have been applied to clinical text of various granularities (*e.g.*, abstracts, sentences, phrases, and so on), from various types of sources (*e.g.*, scientific articles, clinical notes, electronic health records, clinical free texts, and so on), and with various intents (*e.g.*, quality assessment, content categorisation, polarity classification, entity recognition, and so on) [15–21]. For purposes such as retrieval and post-retrieval re-ranking, approaches based on word co-occurrences [22] and bibliometrics [23] have been proposed for improving the retrieval of medical documents. These approaches, however, do not integrate evidence-based recommendations for appraisal. Tang et al. [24] propose a post-retrieval re-ranking approach that attempts to re-rank results returned by a search engine. Their approach is only tested in a specific sub-domain (*i.e.*, Depression) of the medical domain. Kilicoglu et al. [25] focus on identifying high quality medical articles and build on the work by Aphinyanaphongs et al. [26]. They apply machine learning and obtain 73.7% precision and 61.5% recall. More recently, Kim et al. [27] proposed the use of support vector machine (SVM) classifiers to identify high-quality systematic reviews to help EBM practitioners choose the best quality evidence. A similar classification approach has also been suggested by Adeva et al. [28] to support the creation of systematic reviews. These approaches and related research generally model the problem of quality assessment as a binary classification task, where each article may either be of *good* or *bad* quality. Also, the approaches are suitable for ranking single documents only. Our research has two primary differences with existing research on automatic quality assessment: (i) we use a more standardised and specialised scale, with the intent of automatically recommending evidence-based grades; (ii) our approach is for *bodies* of evidence, which may be single documents or multiple documents on the same topic. In our work, we experiment with some of the features that are suggested to be useful by the SORT guidelines (*e.g.*, publication types of articles), and some features that have been utilised in the past to make quality estimates (*e.g.*, journal names, publication dates) in related literature.

Ebell et al. [13] suggest that the publication types of medical articles are good indicators of their qualities. Literature in the medical domain consists of a large number of publication types such as randomised controlled trials, systematic reviews, cohort studies, case studies and so on.[1] These publication types are of varying qualities (*e.g.*, a randomised controlled trial is often of much higher quality than a case study of a single patient). Greenhalgh [29]

[1] A list of publication types used by the U.S. National Library of Medicine can be found at http://www.nlm.nih.gov/mesh/pubtypes2006.html. This list is not exhaustive [accessed 10.11.14].

mentions some other factors that influence the grade of an evidence, such as the number of subjects included in a study and the mechanism by which subjects are allocated (*e.g.*, randomisation/no randomisation), but the latter is generally indicated by the publication type (*e.g.*, randomised controlled trial) of the article. Lin and Demner-Fushman [30] also acknowledge the importance of publication types in determining the quality of clinical evidence. They use a working definition of the 'strength of evidence' as a sum of the scores given to journal types, publication types and publication years of individual publications. Their scores are used for citation ranking, not evidence grading, and therefore their results cannot be compared to ours. Their research also suggests that the journal names and publication years have an influence on the qualities of individual publications, which in turn may influence the grade of evidence obtained from them.

3. Methods

In this section we describe the design and execution of our experiments. We first briefly describe some preliminary analysis to identify useful features. Following that, we discuss automatic approaches to extract features, and our experiments that utilise these automatically extracted features for strength of recommendation (SOR) classification. Finally, we discuss the design of a human evaluation to compare the performance of our approach with human performance on the same data. The results for all experiments are presented in the next section.

3.1. Experimental data and preliminary analysis

We use a corpus[2] developed in-house [14,31] which is specialised for EBM question answering. Each record in the corpus is a clinical query obtained from the 'Clinical Inquiries' section of the Journal of Family Practice (JFP)[3] (the articles date from 2001 to 2010). Each query is accompanied by one or more evidence-based answers, and each answer is generated from one or more medical publications. Furthermore, each answer contains its SOR, a list of publication references and a brief description of the publications including their publication types. Fig. 1[4] shows a screenshot of the data that is used for our research. The figure shows that the article has three answers associated with the question. Each answer has a SOR specified with a brief justification behind the SOR grade (*i.e.*, the publication types of the articles from which the answers have been derived). Following this, there are some detailed justifications which contain references to the source articles. In our research, the abstracts of these referenced articles constitute the source texts from which the SOR predictions are made.

From the corpus, which contains the JFP data in XML format, we collected all evidence-based answers that had their SORs specified. Our final set consists of 1132 evidence-based answers generated from 2713 medical documents. Of the 1132 answers, 330 are of grade A, 511 of grade B and 291 of grade C. In our preliminary feature analysis, the results of which we have previously published [32], we focused on analysing the effects of the following features in determining the evidence grades:

1. The ***publicationtypes*** of the individual publications associated with the recommendations. In our corpus, the publication types associated with bottom-line recommendations are generally mentioned (as is the case with the articles from the JFP from which the corpus has been collected).[5] For the few cases where the publication types are not mentioned, we manually identified this information from the meta-data in the PubMed[6] XML files containing the abstracts.
2. The ***publicationyears*** of the associated documents. Related literature suggests that recent publications are more relevant/reliable than older publications [30]. We obtained publication years of the documents from the PubMed XML files of the abstracts and used them as numeric features.
3. The ***publicationvenues***. We included this feature to assess if studies published in high-impact journals give better quality evidence compared to studies published in venues of lower impact. Past research suggests that the qualities of individual studies may depend on the publication venues, and, therefore, this information is likely to affect the overall evidence grade [30]. Information about the publication venues are also encoded in the PubMed XML files present in our corpus.
4. The **titles** of the articles. The titles of the articles present useful information such as the topics of the studies, and, often other information such as disorders, interventions and so on. The titles were also obtained from the PubMed XML files.

Note that, for this analysis, we refrained from using raw text from the publication abstracts. This is because using word *n*-grams from the abstracts introduces a large number of features, making the analysis of other features difficult. We utilised lexical information from the abstracts in the work described later in this paper.

Due to the large number of possible publication types that medical articles can have, we grouped together publication types having low frequency and similar quality levels, since it is not possible to accommodate all publication types. This grouping is performed manually. Our final set consisted of 11 groups of known publication types as shown in Fig. 2. Publications for which we could not identify the publication type were labelled as *unknown*. Based on our collected data, we consider 45.1% – the accuracy when all instances are classified as B (the majority class) – as the baseline for our experiments.

This distribution of publication types over SOR grades is shown in Fig. 2. A clear pattern in the distribution of publication types over SORs can be seen. For SOR A, evidence primarily comes from randomised controlled trials, systematic reviews and meta-analyses, and the numbers drop significantly for other publication types. For SOR C evidence, most of the evidence comes from publications presenting expert opinion, case series/reports, and consensus guidelines. The distribution for SOR B has the largest spread with cohort studies having the highest frequency.

To test the extent to which SORs can be predicted from the abovementioned features, we design an experiment using supervised machine learning. We model the grading of evidence as a single-label text classification problem with three classes (A, B, and C), and the features that have already been mentioned. We model the publication type feature as vectors of counts of the 12 publication types shown in Fig. 2. We represent the titles and journal names using uni- and bi-grams; we do not include any larger *n*-grams, since both of these are generally short spans of texts and larger *n*-grams lead to data sparseness. Prior to generating the *n*-grams, we process the titles by removing stop words, lowercasing the words, stemming the remaining words using the Porter stemmer [33] and removing words occurring less than five times across the whole data set. We repeat the experimental procedures mentioned above with various combinations of these feature sets.

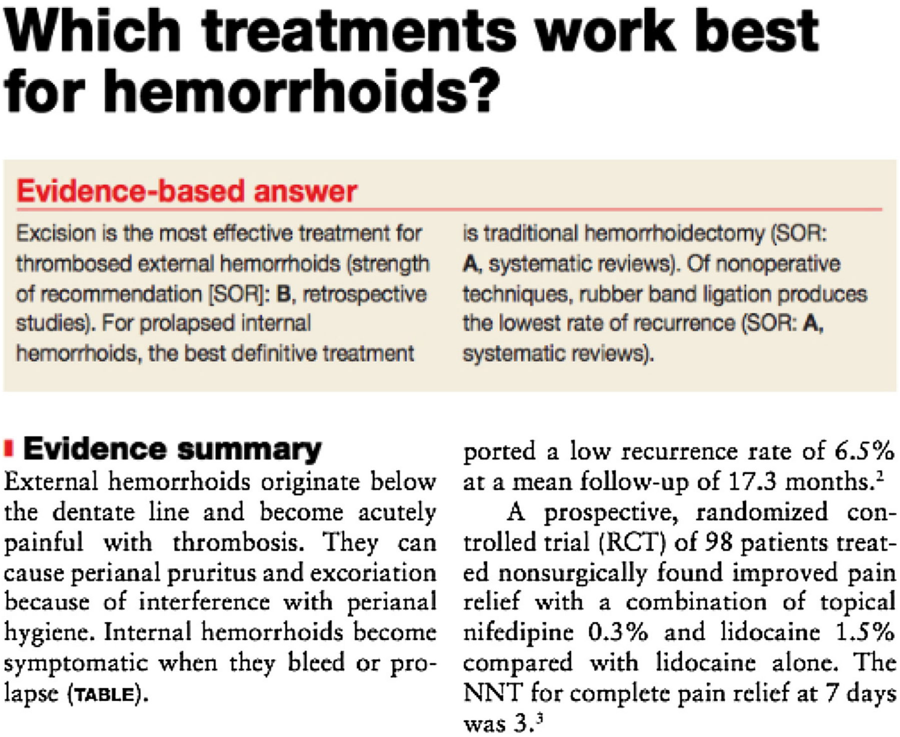
[2] The corpus is available to the research community at: http://sourceforge.net/projects/ebmsumcorpus/ [accessed 10.11.14].

[3] http://www.jfponline.com [accessed 10.11.14].

[4] Reprinted with permission of *The Journal of Family Practice*.

[5] An example of this can be found at: http://www.jfponline.com/pages.asp?aid=8103 [accessed 10.11.14].

[6] http://www.ncbi.nlm.nih.gov/pubmed/ [accessed 10.11.14].



Which treatments work best for hemorrhoids?

Evidence-based answer

Excision is the most effective treatment for thrombosed external hemorrhoids (strength of recommendation [SOR]: **B**, retrospective studies). For prolapsed internal hemorrhoids, the best definitive treatment is traditional hemorrhoidectomy (SOR: **A**, systematic reviews). Of nonoperative techniques, rubber band ligation produces the lowest rate of recurrence (SOR: **A**, systematic reviews).

Evidence summary

External hemorrhoids originate below the dentate line and become acutely painful with thrombosis. They can cause perianal pruritus and excoriation because of interference with perianal hygiene. Internal hemorrhoids become symptomatic when they bleed or prolapse (TABLE).

ported a low recurrence rate of 6.5% at a mean follow-up of 17.3 months.[2]

A prospective, randomized controlled trial (RCT) of 98 patients treated nonsurgically found improved pain relief with a combination of topical nifedipine 0.3% and lidocaine 1.5% compared with lidocaine alone. The NNT for complete pain relief at 7 days was 3.[3]

Fig. 1. A sample article from the *Journal of Family Practice* showing the clinical query, the bottom-line summaries, the associated evidence grade for each answer, and justifications behind the grades.

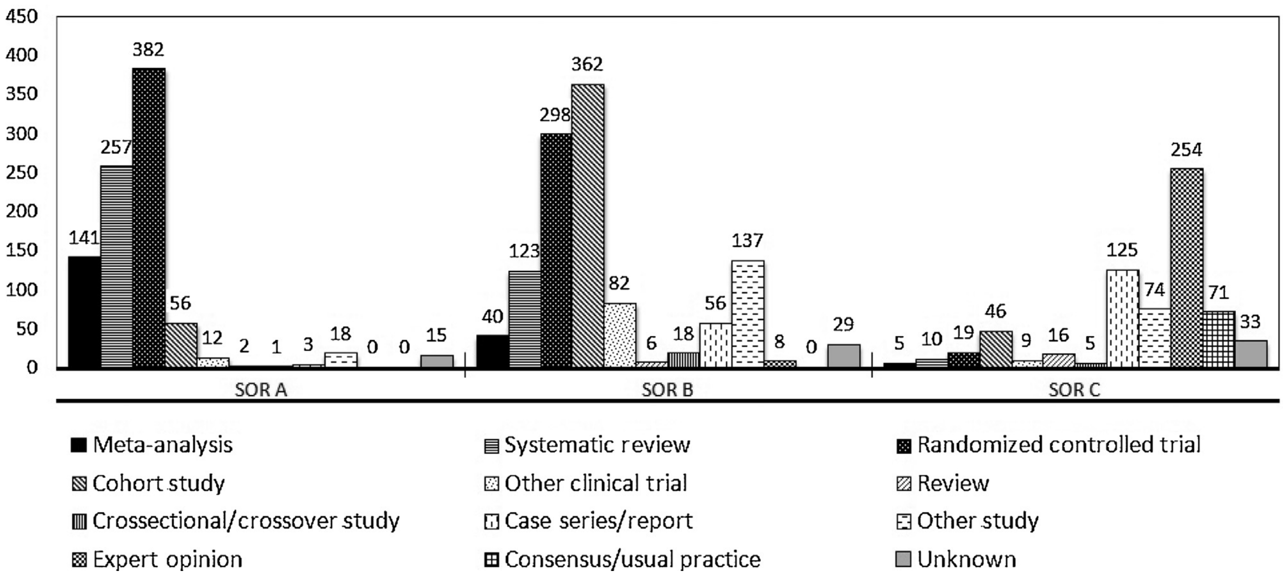


Fig. 2. Distribution of publication types across SORs in our analysis set. *Other study* refers to low frequency studies (*e.g.*, observational study), *Other clinical trial* refers to all clinical trials other than randomised controlled trials, and *Unknown* refers to articles with unidentified publication types.

We use two-thirds of our data for training and the remaining as held-out test data. For both sets, we keep the proportions of instances belonging to the three classes the same as their proportions in the whole data set. We use the software package Weka[7] [34] for our machine learning experiments. We choose five classifiers that produce good results on our training data and have also been shown to produce good results on similar problems in the past. The five chosen classifiers are (the names used in Weka shown in parenthesis): bayes net, support vector machines (SMO), *k*-nearest neighbour (IBk), multinomial logistic regression (Logistic) [35], and C4.5 decision tree (J48) [36]. We perform simple parameter tuning for some of the chosen classifiers and choose parameter values that produce the best results for stratified 10-fold cross validations on the training set. For the bayes net classifier, we use the default implementation in Weka with the K2 search algorithm [37] for local score metrics and the simple estimator for estimating conditional probability tables. For the SMO algorithm, we use John Platt's [38] sequential minimal optimisation algorithm and solve

[7] http://www.cs.waikato.ac.nz/ml/weka/ [accessed 10.11.14].

Table 1
Accuracies, 95% confidence intervals, and best performing classifiers for various feature sets.

Features	Accuracy (%)	95% CI	Classifier
Journal, pub. year, title and pub. type	63.636	58.5–68.5	C4.5
Pub. type and pub. year	66.578	61.6–71.3	C4.5
Pub. type and title	67.380	62.4–72.1	C4.5
Pub. type and journal	63.904	58.8–68.8	C4.5
Journal, pub. year and title	50.802	45.6–56.0	SMO
Journal and pub. year	46.257	41.1–51.5	SMO
Title only	51.070	45.9–56.2	SMO
Pub. year only	47.594	42.4–52.8	Bayes net
Journal only	47.326	42.2–52.5	Bayes net
Pub. type only	68.717	63.8–73.4	IBk

our multi-class problem using pairwise (1-*vs*-1) classification. We use an RBF kernel for the SMO classifier, normalising all attributes and using a grid search to find good values for the parameters γ and C. For the C parameter, we searched through all powers of 2 from 2^{-5} to 2^{15}; for the γ parameter, we searched through all powers of 2 from 2^{-15} to 2^{3}. To find the best value of k for the k-nearest neighbour algorithm, we search through all odd values of K from 1 to 101. For the C4.5 decision tree classifier, we search between 2^{-5} and 2^{-1} to find the best value for the confidence factor parameter.

Using only manually specified publication types as a feature set, we obtain classification accuracies of approximately 66–69% (over 20% improvement over the majority-class baseline) with various classifiers on our held-out test set. We found no statistically significant difference between the accuracies of the different classifiers, meaning that the most important role to determine classification accuracy is played by the features used, not the type of classifier. Adding features such as journal names, publication years and article titles to the publication types does not significantly influence the SORs. Table 1 shows the highest accuracies obtained using various combinations of feature sets, the 95% confidence intervals,[8] and the classifiers producing these results. From the table it is evident that the absence of publication types as a feature set causes significant drops in accuracy. Although incorporation of article titles as a feature set produces marginally better accuracies compared to our baseline, no significant improvements are achieved when this feature set is combined with publication types. The other feature sets, alone or in combination with each other, do not give a statistically significant improvement over the baseline.

We also experimented to verify the usefulness of several other features for automatic SOR classification, but found none that could influence classification accuracies. For example, we attempted to ascertain if there is a correlation between the number of publications associated with an evidence and the quality of the evidence, but there was none. Based on our manual inspection of the data, it appeared that some topics may be associated with high quality evidence, perhaps due to the significant amount of research performed on those topics. However, due to the large number and diversity of topics present in the corpus, incorporating topic information (*e.g.*, from the question) and using them as features leads to data sparseness. In conclusion, only the publication types showed a strong association with the SOR grades, and the article titles produced small improvements in classification accuracies.

3.2. Identifying publication types of medical articles

Because of the importance of publication types in SOR classification, as revealed by our preliminary analyses, we first focus on the automatic identification of publication types. As mentioned earlier, the PubMed XML files of the abstracts contain associated meta-data, including the publication types of the articles. However, not all publication types that we require are present in the PubMed articles. For example, PubMed does not have a specialised tag for systematic reviews. The *Review* tag is used to represent both systematic reviews and non-systematic reviews. In many cases, simply the default *Journal Article* tag is used for the articles. There is also no specialised tag for cohort studies, although a large number of articles in our data set belong to that broad category. Because of these reasons, we cannot rely fully on the meta-data associated with the articles to identify the publication types.

We apply a simple rule-based approach to automatically identify several important publication types. We have reported preliminary results of our regular expression-based approach in the past [39], and here we provide a comprehensive description of the task. We combine the publication types identified by this approach along with the publication type information provided as meta-data. In our approach to identify the publication types, we only use texts from the article titles and abstracts. Abstracts, and often titles, of medical articles contain information about the types of studies, and therefore, provide evidence of their publication types. Fig. 3 presents three examples of evidence of randomised controlled trials. The first example shows evidence from the title. The second and third are examples of sentences from article abstracts that contain evidence.

Our approach relies on regular expressions to identify relevant patterns (evidence) from titles and abstracts. In addition to developing regular expressions for publication types that do not have specific tags in PubMed, we implemented some expressions for publication types that have specific tags. This is for two reasons:

i often multiple publication types are assigned to a single article, and our expressions are aimed at detecting the most relevant tag for an article; and
ii in some cases, the tags assigned on PubMed are not consistent, and applying our regular expression-based classifier, the publication type can be detected with more accuracy for certain publication types.

An example of (i) may happen when an article is tagged as a *randomised controlled trial*, *controlled clinical trial*, and *clinical trial* in the PubMed meta-data. Using our rule-based approach, if the article is found to be a randomised controlled trial, only that tag is kept and the others are discarded. Details of this are provided later in this section when we discuss our features for the machine learning classifiers in detail. As an example of (ii), we found a number of cases where an article was tagged as *multi-centre study* only, while it was actually a multi-centre randomised controlled trial. The intent of the rule-based approach, in this case, is to tag the article as a randomised controlled trial since that is the most relevant for our work.

We developed regular expressions to classify articles by manually studying the titles and abstracts of articles belonging to several publication types. We collected our development set from a mixture of sources. For articles which have associated *PublicationType* tags in PubMed (*e.g.*, randomised controlled trials and meta-analyses), we retrieved about two hundred of each type. We studied each article individually, identified the evidence of publication type and developed patterns to pick up the evidence. During the development of the rules, we used an incremental approach similar to the ripple down rules [40] philosophy – after adding a new regular expression we tested its effect on our development set and added more expressions based on the articles that are not correctly identified. We utilised this strategy for meta-analyses, randomised controlled trials, consensus development conferences,

[8] Calculated using the package R's binom.test function (http://www.r-project.org) [accessed 10.11.14].

Evidence of publication type in title:

A ***randomised controlled trial*** *of self-help interventions in patients with a primary care diagnosis of irritable bowel syndrome.*

Evidence of publication type in abstract:

Prospective randomised controlled trial *of low risk women admitted in spontaneous labour, with intact membranes.*

In this study, 200 participants who met the diagnostic criteria for cervicogenic headache were ***randomised into four groups****: manipulative therapy group, exercise therapy group, combined therapy group, and a control group.*

Fig. 3. Examples of evidence of publication type in title and abstract texts. The first example shows how the title can provide evidence of the publication type. The second and the third examples show how abstract sentences can provide evidence about the publication type.

```
Evidence of randomisation for randomised controlled trials:
random.*alloc
desig:.*random
random.*animal
patien.*random
subjec.*random
randomi[sz].*group
random.*doub.*blind
random.*open[\W]*label
randomi[sz]e.*trial

Evidence of no or unacceptable randomisation:
coin\W*flip
non\W*random
odd\W*even
uncontrol\W*stud
```

Fig. 4. Sample patterns used for detecting randomised controlled trials. Sample patterns used for detecting unacceptable randomisation techniques are also shown.

practice guidelines, and reviews. For example, in the case of randomised controlled trials, we primarily developed expressions to detect evidence of randomisation in the abstracts. Once evidence of randomisation is found, the expressions attempt (from false positives) to detect evidence(s) of unacceptable randomisation.[9] Some of the expressions we use to identify randomised controlled trials and unacceptable randomisation techniques are shown in Fig. 4 (the list is not exhaustive).

For articles without an associated *PublicationType* tag in PubMed (*e.g.*, systematic reviews), obtaining a large development set was considerably more difficult. We therefore used a mixture of secondary sources of evidence such as the Journal of Family Practice and the Cochrane Library for obtaining approximately fifty of each and developed our expressions from that set. In these sources, the publication types are manually specified. In our corpus, the publication types of the cited articles are often given, and we used those annotations as our gold standard. Furthermore, we studied search techniques suggested by PubMed[10] for the efficient retrieval of systematic reviews and developed expressions based on their suggestions. We also developed expressions based on search keywords and techniques suggested in the literature for obtaining articles of specific publication types [41,42]. Developing rules was easier for publication types of higher qualities (*e.g.*, systematic reviews). This is primarily because articles belonging to these publication types often have standard discourse structures, and also, almost invariably, clearly state the type of publication in the abstract or the title. The same is not true for publication types of lower qualities.

We applied this approach for systematic reviews and cohort studies. We also attempted to apply this technique for case series, case control studies, and some other publication types. However,

[9] Details about unacceptable randomisation techniques for randomised controlled trials and other publication types can be found at http://www.nlm.nih.gov/mesh/pubtypes2004.html [accessed 10.11.14].

[10] The techniques can be found at http://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html [accessed 10.11.14].

due to the inconsistent nature of the abstracts of articles belonging to these publication types, and the variety of ways in which the abstracts are written, the accuracies achieved by our preliminary experiments are low. Therefore, for such lower quality publication types, we primarily rely on the *PublicationType* tags provided by PubMed. Fig. 5 shows some of the regular expressions used for detecting systematic reviews, meta-analyses, and some other publication types (the list is not exhaustive).

We apply a decision list to identify the publication types of articles. Each article is initially assigned an empty tag and passed through a sequence of tests, each responsible for checking for patterns indicating a specific publication type. At any stage of the sequence, both the PubMed *PublicationType* tag and the regular expressions corresponding to a specific publication type are utilised to check if an article belongs to that category. For the regular expression matching, the title of the article is first checked, and, if no evidence is found, the abstract is checked. If sufficient evidence of a particular publication type is found (with no further evidence of negation), the article is tagged and removed.

The sequence in which the operations are applied is very important as the number of false positives may increase significantly if the sequence is changed. For example, if systematic reviews and meta-analyses are not removed before searching for randomised controlled trials, many of the former are falsely tagged as the latter. This is because abstracts of systematic reviews and meta-analyses are often produced by collecting information from multiple randomised controlled trials, and they usually mention the number and types of studies being reviewed/analysed (*e.g.*, *We conducted a systematic review of five double-blind, randomised controlled trials to investigate, etc.*). Thus, both these publication types generally mention other publication types and must be detected early on in the sequence. The following list elaborates the actions performed at each stage of the sequence:

```
Input: $a \leftarrow$ untagged article
begin
    if isMetaAnalysis($a$) then
        tag $a$ as meta-analysis; remove $a$;
    else if isSysRev($a$) then
        /* regular expressions only                    */
        tag $a$ as systematic review; remove $a$;
    else if isReviewOrCDC($a$) then
        if isReview($a$) then
            tag $a$ as non-systematic review; remove $a$;
        else if isCDC($a$) then
            tag $a$ as consensus development conference; remove $a$;
    else if isGuideline($a$) then
        tag $a$ as practice guideline; remove $a$;
    else if isRCT($a$) then
        tag $a$ as randomised controlled trial; remove $a$;
    else if isCT($a$) then
        /* meta-data only                              */
        tag $a$ as clinical trial; remove $a$;
    else if isCohort($a$) then
        /* regular expressions only                    */
        tag $a$ as cohort study; remove $a$;
    else if isOther($a$) then
        /* this step utilises meta-data                */
        tag $a$ as other publication type; remove $a$;
end
Output: $a$ with publication type tag
```

```
Evidence of systematic reviews:
sytemat.*rev
cochr.*medlin
search.*cochr
search.*medlin.*datab

Evidence of cohort studies:
retrospect.*stud
foll.*prospect
retrospect.analys
patien.*foll.*year

Evidence of practice guidelines:
guidel.*diag.*treat
clinic.*guidel

Evidence of consensus development conferences:
consens.*confer
consens.*statem

Evidence of review (non-systematic):
postmar.*survei.*surv
retrospec.*char.*rev
```

Fig. 5. Sample patterns used for detecting specific systematic reviews, cohort studies, practice guidelines, consensus development conferences, and non-systematic reviews.

In all cases, the meta-data is checked first (if available), then the title and finally the abstract text. While checking the abstract of an article for evidence, each sentence is searched separately. We have attempted other approaches such as searching the whole abstract and using a sliding window. However, we have found sentence level searching to produce the best results primarily because evidence of publication or study type is usually stated or described in a single sentence of an article abstract. Once a pattern match occurs, the entire abstract is searched again to identify patterns that negate the evidence in specific cases (such as unacceptable randomisation techniques in the case of randomised controlled trials), and the article is only tagged if no evidence of negation is found. The results of the evaluation process are provided in the next section.

3.3. Automatic SOR classification

3.3.1. Features and methods for SOR classification

For our fully automatic classification approach, we use the data from the 2011 ALTA shared task [43]. The task was based on the problem of automatic grading of evidence, and the data set was prepared from our corpus. The data for the shared task consisted of a set of 'evidences' with the SORT grade for each. Each evidence was represented as a list of publications (PubMed IDs) from which the evidence had been generated. Information for each publication was provided in the form of an XML file per publication obtained from PubMed. Two sets of such data were provided initially for training (677 evidences) and development time testing (178 evidences), and an additional set was used for testing the final system (183 evidences). Bottom-line summaries with no associated abstracts and abstracts containing no text were not included in this data set.

```
75474 B 18492531
75475 B 12597676
75476 C 9394980
75477 B 16536797 16308411
75478 B 8582464
75479 C 10960901
15561 B 3822681 10940092 10937475 12390647 10971664 11669346
15562 B 10937475 2191938 2861907 3282670
74132 C 9215014
74133 A 11882771 9215014 9606614
74134 B 9606614
74135 B 17065896 14970960 12373695 11166969 16213659
17512 A 3147087 2040866
17513 A 2004476 7641412 8481068
```

Fig. 6. Sample data from the 2011 ALTA shared task.

Fig. 6 illustrates how the data in the shared task was provided. In the figure, the first column is the instance ID, the second column is the grade, and the following columns represent the PubMed IDs of the abstracts associated with each instance.

Based on the findings of our preliminary experiments, we use publication types and article titles as feature sets. In addition, we introduce word *n*-grams from the article abstracts as a feature set.[11] We now provide a description of these feature sets.

***N*-grams.** The most important information contained in the articles lies in the text of the abstracts. These include types of studies, sizes of studies, background information, results and outcomes. To attempt to capture this information, we generate *n*-grams ($n = 1$, 2, 3 and 4) for each of the abstracts in the training set. The abstracts contain larger volumes of texts compared to the titles, therefore, up to 4-grams is possible without introducing data sparseness issues. Prior to generating the *n*-grams we perform some preprocessing of the text. It is common for medical concepts to have different lexical representations. For example: *hbp*, *hypertension*, and *high blood pressure* represent the same medical concepts. The concepts can further be generalised into broad categories representing classes of these concepts. In our approach, we replace specific medical concepts in the texts with generic *'sem_type'* tags. We use MetaMap[12] [44] to identify domain specific concepts as defined in the unified medical language system (UMLS).[13] The UMLS provides a vast vocabulary of the medical concepts and the broad semantic groups into which the concepts can be classified. For example, all disease names fall under the semantic category *disease or syndrome (dsyn)*. Replacing each occurrence of a disease, syndrome, or any medical problem mention with the generic tag ensures that the mention does not have an influence on the classifiers used and reduces overfitting. To represent all medical problems, we use a set of semantic types previously used for similar text classification tasks [45].[14] We also preprocess the *n*-grams by stemming using the Porter stemmer [33], lowercasing, and removing stop words.

Publication types. We employ the approach described in the previous section to automatically identify the publication types of the articles. Due to the difficulty of accurately identifying all the 11 types of publications mentioned in Fig. 2, we further condense the number of publication types and only use the ones mentioned in the previous subsection for automatic detection. Abstracts that are

[11] We have experimented with other features, but this combination produced the best results.

[12] http://metamap.nlm.nih.gov/ [accessed 10.11.14].

[13] http://www.nlm.nih.gov/research/umls/ [accessed 10.11.14].

[14] *pathological function*, *disease or syndrome*, *mental or behavioral dysfunction*, *cell or molecular dysfunction*, *virus*, *neoplastic process*, *anatomic abnormality*, *acquired abnormality*, *congenital abnormality*, and *injury or poisoning*.

not captured by our automatic approach are given the tag specified in the PubMed meta-data accompanying the abstract text. For articles with multiple publication types, we only keep the tag that represents the highest quality. For example, if an article is tagged as a randomised controlled trial, a clinical trial and a journal article, we only keep the randomised controlled trial tag, since it has the highest quality among the three types. This approach produces a total of 23 publication types, including the automatically detected publication types. Articles that can neither be classified by our rule-based approach nor by the associated meta-data are given the default *Journal Article* tag. For each instance of SOR in our data set, this feature set is added as a vector of the counts of each publication type.

Titles. We generate word uni- and bi-grams from the titles, preprocessing them the same way as in the preliminary analysis experiments.

3.3.2. *Classification approach*

We apply the same classifiers that we used for our preliminary research, and use the software package Weka. Namely, we experiment with the following classifiers: bayes net, support vector machines (SMO), *k*-nearest neighbour (IBk), multinomial logistic regression (Logistic) [35], and C4.5 decision tree (J48) [36]. We also use a naïve bayes classifier for comparison. The results are presented in the next section.

In addition to the accuracy values, we add another measure, which we call the *average error distance* (AED). The intent of this measure is to estimate the extent to which the predictions made by our system differs from the actual grades. For an instance, if the actual grade is A and our system predicts B, then the *Error Distance* (ED) is 1. Similarly, if, for the same instance, the prediction by our system is C, the ED is 2. For correct predictions the ED is 0. The formula to compute AED is as follows:

$$AED = \frac{\sum_{g \in G} ED(g_p, g)}{(2 \times (N_a + N_c)) + N_b} \qquad (1)$$

where N_a, N_b and N_c represent the number of instances with actual grades A, B and C, g_p is the predicted grade, g is the actual grade, and the function $ED(g_p, g)$ gives the ED for that instance. Note that in the equation, N_a and N_c are multiplied by 2 because the maximum ED for these two classes is 2, whereas, for N_b, the maximum ED is 1.The formula for AED ensures that, for a given data set, the worst performing system will obtain an AED of 1, while the best performing system will obtain an AED of 0. Thus, the lower the AED for a system, the more likely it is to make accurate predictions. In other words, smaller AEDs mean that the classifier predictions are *closer* to the actual predictions. The AED metric makes it easier to compare different systems on the same data set. The performance of two systems having comparable or equal accuracies may be very different in practice. For example, a system which frequently classifies C grade evidence as B is better in practice than a system that classifies the same instances as A. In other words, the errors made by the second system are *bigger*, and the grades predicted are further from the actual grades. Although the accuracies of these two systems may be equal, their AED values will indicate their relative performances.

3.3.3. *Combining classifiers*

In an attempt to further improve the performance of our system over the performances of the single classifiers, we apply a sequence of classifiers, each of which is provided with a specific feature set instead of all the feature sets at the same time. The intent of each classifier in the sequence is to attempt to identify instances belonging to the A and C classes with relatively high precision, at the expense of recall. Thus, each classifier in the sequence classifies only a small number of instances as A and C classes. When a number of such classifiers are utilised, the number of correctly classified A and C grade evidences increases at each step, with a lower number of false positives for both these classes compared to the number of false positives when all feature sets are combined in a single classifier. This approach is similar to the idea of boosting in machine learning [46]. Based on the results of the single classifier experiments, which are presented in the next section (Table 3), we use the SMO classifier for these experiments. The sequence in which the classifiers are applied and specific details about each of them are as follows:

Step 1: Classify all evidences as grade B (majority class).
Step 2: SMO with *n*-grams (n = 1, 2, 3, 4 and semantic types replaced) as features. Parameters: c = 2.0 and γ = 0.0. Attribute selection: using the information gain measure to select the top 400 *n*-grams.
Step 3: SMO with publication types as features. For each instance, the frequency of each publication type is used. Parameters: c = 1.0 and γ = 0.0.
Step 4: SMO with titles as features. Parameters: c = 32.0 and γ = 0.002.

The parameters for the SMO are tuned using the training set for training and the development time test set for evaluation. Each of the above classifiers and their parameters are chosen based on their precision in classifying A and C grade evidences. Thus, in our algorithm, each classifier classifies most instances as B but identifies some A and C class instances with high precision. At each step, instances classified as A or C are removed from the set, and the new grades are assigned to these instances. Using this approach, the classification accuracy increases with each step of the algorithm as more instances are classified as A and C.[15] Fig. 7 graphically summarises our fully automating SOR classification model.

3.4. *Human evaluation*

To fully understand the applicability of our system in real life grading of evidence, it is pertinent to compare its performance against the performance of human experts on the same data. This is particularly important for the following two reasons. Firstly, the articles in the *Clinical Inquiries* section of the JFP are authored by different domain experts who follow the guidelines of EBM to answer the clinical queries that are the topics of the JFP articles. For each article, the authors retrieve the relevant literature available on the topic, read and analyse them to identify the key recommendations, synthesise information from multiple relevant documents, appraise the quality of the information they have gathered, and then choose a grade based on the guidelines of the SORT. To specify the final grade of evidence, they combine their domain knowledge and the information available to them, and make the judgments based on them. The authors are required to compare the information available to them with the guidelines available and make the judgments. In many cases, as expected, the information available to the authors cannot be directly mapped on to the rules of the guidelines. This may happen, for example, when there are multiple articles available on the same topic, with contrasting information in some of them. The authors may choose to refer to the *best* papers only when making the final recommendations, and ignore the lower quality articles presenting contrasting information. The final grade assigned, therefore, will depend only on the papers on

[15] For empirical reasons: we experimented by changing the sequence of classifiers, but found no significant differences in performances.

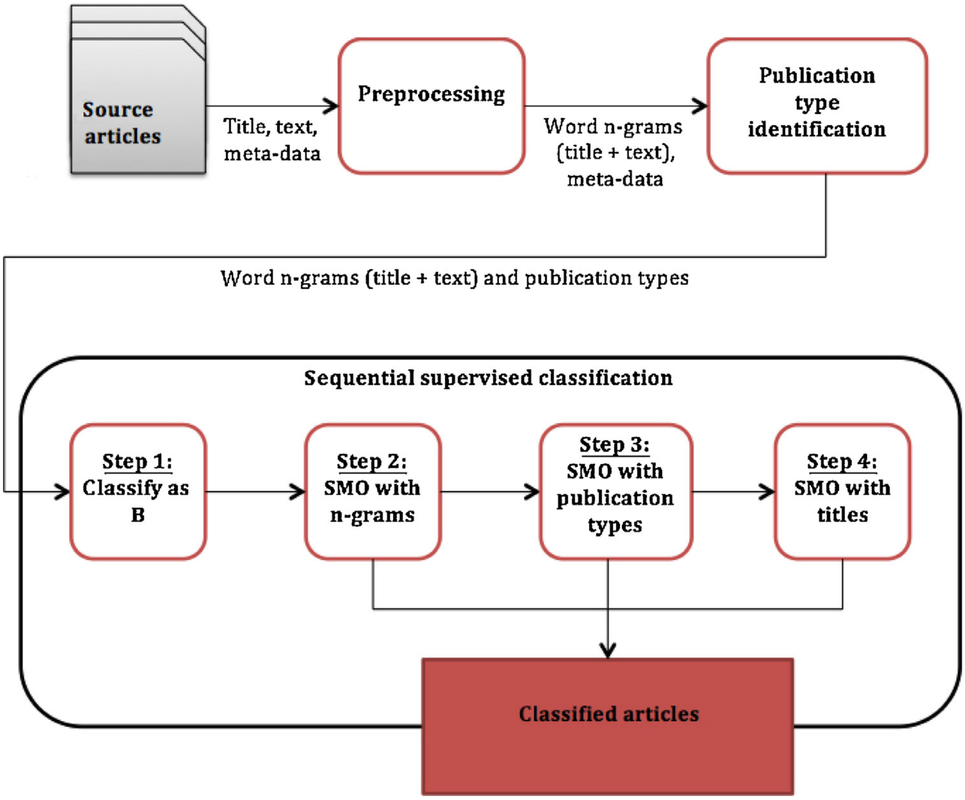

Source articles
Title, text, meta-data
Preprocessing
Word n-grams (title + text), meta-data
Publication type identification
Word n-grams (title + text) and publication types
Sequential supervised classification
Step 1: Classify as B
Step 2: SMO with n-grams
Step 3: SMO with publication types
Step 4: SMO with titles
Classified articles

Fig. 7. The complete SOR classification model, including the sequential, high precision supervised classification.

which the authors chose to rely on. However, there is a strong possibility that if other experts were given the same set of documents to make a decision, they would apply their own methodology to derive the final grade for the evidence. As such, there is also a strong possibility that different experts, when faced with the same clinical query and the same set of supporting information, will choose different grades to specify the qualities of the evidences, particularly when there are time-related constraints. It is crucial to compare the agreement levels among different experts in order to understand how good or bad the performance of our system is.

Secondly, the JFP contributors have access to more information than our system. The experts can utilise information from the full articles, while our system has to rely on information present in the abstracts only. Furthermore, in a number of cases (approximately 10%), the abstracts do not contain any text, making it harder for the system to achieve its goal. Therefore, it is likely that the system's predictions are affected by the lack of information that is provided to it.

We wanted to investigate the following two issues:

i The extent to which human experts agree with the gold standard annotations in our corpus, given the same data that our system receives.
ii The extent to which human experts agree amongst themselves regarding the evidence grades, given the same data that our system receives.

To commence this experiment, we first selected a random set of 100 instances from the text classification task described in the previous section. Each instance consisted of a clinical query, a set of article abstracts that were referenced to generate the bottom-line summary associated with the evidence, and a grade indicating the quality of evidence. Among the 100 instances, 38 were of grade A, 36 were of grade B, and 26 were of grade C.[16]

We employed four human experts to perform the grading task on this data set, using their own expertise. The human experts chosen were from different backgrounds, but prior to introducing them to the task, they were tested on their knowledge of EBM practice. Three of the human experts were practising chiropractors who graduated from Macquarie University, and the fourth expert was a final year medical science student from the Australian National University. For the task, we implemented a web-based tool that enabled the experts to read the queries and the associated abstracts, and choose a grade for each evidence based on the given information. Each expert received one hour of training prior to the grading task. During the training process, the experts were introduced to the SORT guidelines, and they performed grading on a small set of examples that was separate from the 100 instances mentioned earlier. Following the grading of this sample set, the JFP grades were revealed to the experts, and the reasoning behind the grades were also explained. Using the web-based tool, the experts were able to access the instances at any time; the tool also allowed them to revise the grades they assigned, and leave comments justifying their decisions.

4. Evaluations, results and discussions

We now present the results of the various experiments described in the previous section. We first discuss the results of our rule-based approach for detecting the publication types of medical articles. Following that, we present the results of our classification approaches, and finally, we discuss the results of our human

[16] The proportions for each of these three grade categories are not the same as in the full data set purely as a result of the random selection process.

Table 2
Performance of the rule-based approach for the automatic classification of publication types.

Publication type	Recall	Precision	*F*-score
Meta-analysis and systematic review	0.99	1.00	0.995
Randomised controlled trial	0.96	0.99	0.975
Cohort	0.81	0.78	0.795
Consensus development conference	0.76	0.92	0.832
Practice guideline	0.90	0.86	0.883
Other clinical trials	0.84	0.79	0.814
Other	0.78	0.65	0.711

Table 4
Average error distances for the best performing classifier (SMO) for individual features and all features combined.

Features	Average error distance
Abstract text *n*-grams	0.332
Title text *n*-grams	0.339
Publication types	0.300
All	0.289

evaluations. For each set of results presented, we also provide discussions of the key findings and error analyses.

4.1. Publication type detection results

To evaluate the performance of our rule-based approach, we required a set of test articles that were different from the development set and at the same time reliably annotated. To achieve this, we use the articles in our corpus whose publication types are explicitly mentioned by the JFP authors. Importantly, the chosen articles are not actually written by JFP authors, but are cited by them within the JFP articles and the publication types are also identified by these authors. These JFP classifications are considered to be the gold standard in our evaluations. Relying on JFP for the test data also allows us to include articles from a wide range of medical topics, thus ensuring that our approach is not topic dependent. To further prevent bias, all articles identified are added to the test set regardless of their structure/content, and the abstracts of the articles are not reviewed during the annotation process.

We use a total of 518 articles for evaluation, which includes 111 systematic reviews and meta-analyses, 100 randomised controlled trials, 78 cohort studies, 17 consensus development conferences, 31 practice guidelines, 92 non-randomised clinical trials, and 89 other studies (other). Table 2 presents the performance of our rule-based approach. Note that in the table we do not distinguish between systematic reviews and meta-analyses because these two publication types are essentially the same (*i.e.*, meta-analyses are types of systematic reviews). For systematic reviews and meta-analyses, our approach produces perfect precision but fails to identify one systematic review. Our approach tags a total of 97 articles as randomised controlled trials, of which 96 are correctly identified. In the case of randomised controlled trials, the falsely tagged article is a review (non-systematic) which mentions ‘*one randomised, placebo-controlled study*’ and is therefore picked up by our rules. As for the four randomised controlled trials that are missed, none of their abstracts contain any evidence of randomisation although for one of the randomised controlled trials, there is clear evidence of randomisation in the full article text (identified by manual inspection of the full text). In the case of systematic reviews and meta-analyses, the unpicked article is a systematic review in which the abstract does not contain any detail of the study type. It can be seen that the *F*-scores for the lower quality publication types are lower than those for the higher quality publication types, indicating that automatic classification gets increasingly difficult as the qualities of the articles decline.

The results indicate that a rule-based approach such as ours is very effective in classifying high quality publication types, such as systematic reviews, meta-analyses, and randomised controlled trials. The high *F*-scores can be attributed to the fact that articles belonging to these three publication types are very structured (since there are very specific guidelines that must be followed when writing these articles), therefore, their titles and abstracts almost invariably contain sufficient evidence of the type of publication, which can be automatically identified. In contrast, lower quality publication types such as non-randomised clinical trials and cohort studies vary significantly in terms of abstract structure and content. Thus, the rule-based approach is less accurate for these publication types. However, for most of these publication types (other than cohort studies), the PubMed *PublicationType* tag is generally correctly specified. Thus, with these two information types combined, it is possible to detect the abovementioned publication types fairly reliably.

4.2. Automatic SOR classification results

For these experiments, both the training set and the development test (855 instances) from the ALTA shared task are used for training, and the test set (183 instances) used for evaluation. Table 3 presents the performances of the different classifiers for each feature set, and all the feature sets combined, using automatically extracted features. Consistent with our preliminary experiments, publication types prove to be the best feature set, followed by *n*-grams. The best performing classifier is SMO, which performs better than the other classifiers when the feature sets are combined. Table 4 presents the AEDs for the best performing classifier for the different feature sets, showing that the AED is minimised when all the features are combined.

Automatically extracted features do not provide as good a performance as manually extracted features. Thus, despite the use of more features, we are not able to obtain results as good as those from our preliminary experiments. When performing analysis on the training set to optimise various classifier parameters, we notice that a major problem in this classification process is increasing the recall and precision for the A and C classes. Attempting to improve recall significantly decreases precision and vice versa. For the best feature combination and classifier, our system obtains an average precision and recall of 0.56 and 0.51 respectively. The B class, being the majority class, has a recall of 0.89, while the A and C classes have recalls of 0.39 and 0.24 respectively. Optimising the

Table 3
Individual classifier accuracies and 95% confidence intervals for six classifiers using all three feature sets. All features are automatically extracted.

Classifier	Abstract (%)	Title (%)	Publication type (%)	All (%)
Naïve bayes	48.3 (41–56)	45.2 (38–53)	54.9 (47–62)	53.5 (46–61)
Bayes net	46.7 (39–54)	48.6 (41–56)	55.7 (48–63)	56.1 (49–64)
K-nearest neighbour	39.3 (32–47)	47.1 (40–54)	54.6 (47–62)	51.4 (48–63)
Logistic regression	44.2 (37–52)	48.1 (41–56)	55.9 (48–63)	56.6 (49–64)
C4.5	41.0 (34–48)	47.5 (40–55)	58.5 (51–66)	57.4 (50–65)
SMO	49.7 (42–57)	52.5 (45–60)	57.4 (50–65)	60.1 (53–67)

Table 5
Confusion matrix showing number of correctly and incorrectly classified instances when the sequential classification approach is used. The rows show the actual classes and the columns represent the system classifications.

	A	B	C
A	25	29	2
B	6	79	4
C	5	22	11

Table 6
Classification *F*-scores for each of the three classes at each step of the sequential classifier.

Class	Step 1	Step 2	Step 3	Step 4
A	NA	0.278	0.500	0.543
B	0.486	0.666	0.706	0.721
C	NA	0.094	0.415	0.400

classifier parameters to improve the recall values of the A and C classes also causes significant drops in the recall for class B. Also, a number of A and C grade evidences are mis-classified as each other, resulting in high AED values overall. Interestingly, our post-classification error analysis showed that in some cases an instance that is correctly classified when a single feature set is used may get classified incorrectly when all the feature sets are combined, although the overall accuracy tends to increase. Due to this, we apply our sequential classification approach, which attempts to minimise mis-classifications at each step.

4.3. Combined classification results

For the final evaluation, we train our classifiers using the training set and the development test set, and evaluate the performance using test set instances. Among the 183 instances of the test set, our classifiers classify 36 as grade A, 130 as grade B, and 17 as grade C. This achieves an overall accuracy of 62.84%, meaning that 115 instances out of the 183 were correctly classified. This is significantly better than the baseline of classifying all instances as grade B, which has an accuracy of 48.63% (CI: 41.50–55.83). The AED value for this is 0.271, which is smaller than those of the previous experiments. Table 5 presents the confusion matrix for this classification strategy, and Table 6 presents the *F*-score for each class at each stage of the sequence. From the table, it can be seen that the *F*-scores for all three classes tend to increase with each step of the sequence. The average *F*-score for the three classes is 0.555, compared to the average *F*-score of 0.506 when all feature sets are used in a single classifier. Using the features separately in a sequential, high-precision classifier model ensures that once A and C grade evidences are classified, with high precision at a specific step, they cannot be mis-classified at the later stages when other features are used. Furthermore, since all instances are classified as B at the beginning, large ED mis-classifications (*i.e.*, A to C or C to A) are reduced. In addition to increasing accuracy and decreasing AED, an advantage of this approach is that more high precision classifiers can be plugged into this pipeline. This can further increase accuracy while also ensuring that very few A grade evidences are classified as C and vice versa.

4.4. Human evaluations

We compute a number of statistics which enable us to better understand the performance of our system and the validity of our supervised classification model. Fig. 8 shows the grade distributions for the gold standard, our system, and the four experts. The only significant distinctions between the distributions that can be noticed from the figure is the high number of B grades assigned by our system and the low number of C grades. This is due to the design of our system, as explained in the previous section.

Table 7
Accuracy values for the four experts, along with 95% confidence intervals, and our system when compared to the gold standard annotations.

	System	Expert 1	Expert 2	Expert 3	Expert 4
Accuracy	0.61	0.58	0.69	0.71*	0.62
95% CI	0.51–0.70	0.48–0.68	0.59–0.78	0.61–0.80	0.52–0.72

* Indicates statistical significance.

By considering the grades in the gold standard as the correct grades, we compute the *accuracies* for the four experts and our system. Table 7 shows the *accuracies* obtained by the four expert graders and our system on this small data set of 100 instances. Three of the four experts obtain better *accuracies* than our system, and one expert obtains a lower *accuracy*; only one of the expert's *accuracy* is statistically significantly better than that of our system. The results are encouraging for our system because they suggest that the performance of our approach is at least close to the performance of human experts. Also, given the same data, our system's grades and the experts' grades have similar differences to the gold standard grades in terms of accuracy. However, the results do not explain whether the reason behind this is the lack of available information, or if there is a good amount of disagreement among the different human authors on the same information. Therefore, we investigate the agreements among the different experts.

We use the Cohen's Kappa [47] measure (shown below) to compute inter-expert agreements.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{2}$$

where $Pr(a)$ is the relative observed agreement among two experts, and $Pr(e)$ is the hypothetical probability of chance agreement. We apply this statistic in two different ways to obtain an understanding of agreement between experts. In the first approach, we compute the pairwise agreements between the expert grades and the gold standard grades. This gives us a better understanding of the extent to which the experts employed by us agree with the JFP contributors. In the second approach, we compute the pairwise agreements between the experts only, and this gives us an idea of the extent to which the experts agree amongst themselves, based on the data available to them.

Table 8 presents the pairwise agreements between each expert and the gold standard grades along with the mean and standard deviation. From the table, it can be seen that the agreement values for experts 1 and 4 are much lower than the agreement values for the two other experts. The mean agreement is 0.47, which can be regarded as moderate agreement [48]. The standard deviation in agreement, at 0.09, is also relatively high. These agreement values indicate that experts only have moderate agreement amongst themselves regarding the grades. Using the same measure, our system's grades have an agreement of 0.40 with the gold standard grades. These agreement values support the idea that there is not complete agreement among the experts when it comes to assigning

Table 8
Pairwise agreements between each expert and the gold standard grades, along with the mean and standard deviations.

	Agreement (κ)
Expert 1	0.37
Expert 2	0.54
Expert 3	0.56
Expert 4	0.42
Mean agreement	0.47
Standard deviation	0.09

Table 9
Pairwise agreements between the experts.

	Agreement (κ)
Expert 1, Expert 2	0.50
Expert 1, Expert 3	0.42
Expert 1, Expert 4	0.46
Expert 2, Expert 3	0.54
Expert 2, Expert 4	0.61
Expert 3, Expert 4	0.60
Mean agreement	0.52
Standard deviation	0.08

Table 10
Average error distances for the experts' grades and our system's grades.

	Average error distance
Expert 1	0.28
Expert 2	0.24
Expert 3	0.21
Expert 4	0.26
System	0.26

evidence grades. This also suggests that our system's performance is promising, given the information with which it is provided.

Table 9 presents the pairwise agreements between the experts, along with the mean and the standard deviation. The mean agreement, at 0.52, is slightly higher than the mean agreement with the gold standard, although this increase is not statistically significant. From the table, it can be seen that Expert 1 has particularly low agreement with the other experts, and if the grades assigned by this expert are ignored, the mean agreement rises to 0.58. However, the level of agreement can still be considered to be moderate [48]. This further supports the possibility that, despite the presence of a clear and simple guideline for the SORT, experts still vary in their assignment of grades.

To further compare our system against the expert grades, we compare the AEDs of our system on this data with the AEDs of the experts' grades. Table 10 presents the AEDs, which are computed relative to the gold standard grades. It can be observed from the table that the AED values resemble the accuracy values, with the system's AED falling between the AEDs of the four experts' grades, and Expert 3's grades having the best (lowest) AED value. The comparable AED value of our system to the expert annotations further shows the promise of our classification model and approach.

Table 11
Pairwise agreements between the experts and our system.

	Agreement (κ)
Expert 1, System	0.35
Expert 2, System	0.29
Expert 3, System	0.35
Expert 4, System	0.35
Mean agreement	0.34
Standard deviation	0.03

As the final comparison between our system's grades and the experts' grades, we compute the pairwise agreements between our system and the grades assigned by the experts. Table 11 shows the agreement values, along with the mean and standard deviation. The mean agreement is lower than the mean agreement between the experts, and this can be considered to be *fair* agreement [48].

4.5. Human evaluation discussions

We have obtained some interesting results from the human evaluations described in this section. We can summarise the key findings of this evaluation as follows:

1. From the agreement measurements we see that the mean κ for the agreements among the experts employed by us is marginally higher than the mean κ for the expert – gold standard agreements;
2. The *accuracies* and AEDs of our system and the experts are comparable in general; and
3. There is significantly lower agreement between the experts and our system compared to the agreement between the experts only.

The most likely reason behind finding (1) is the fact that the JFP contributors have access to the full articles, whereas the experts employed by us had access to the abstracts only. When generating evidence grades, it may be beneficial to incorporate information from full documents, rather than just the abstracts. However, the performance of our supervised classification model and system

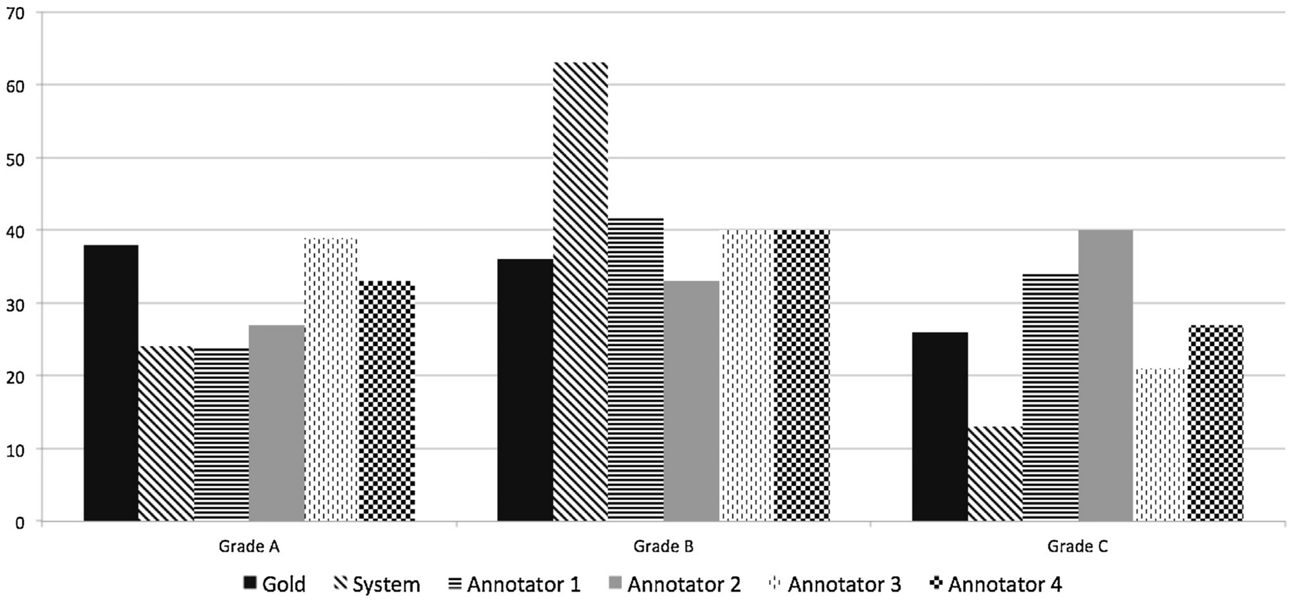


Fig. 8. Evidence grade distributions for the gold standard, our system, and four experts.

using features derived from full articles can not be determined at this point due to the unavailability of that data. Furthermore, the JFP contributors are likely to review more articles than those cited. Thus, if we provide our system with all the articles retrieved, its performance is likely to change.

Finding (2) shows the promise of our supervised classification model and the approach we use for the classification task. Since our system achieves accuracies and AED values comparable to the human experts, we can conclude that it is not possible to significantly improve the accuracy of our system with the data that is currently available. Comparable AED values also indicate that the magnitude of error made by our system is similar to the amount of error human experts would make, given the same data.

Finally, finding (3) shows that, despite the good relative performance of our system, there is still room for modifications that can be made to our system, and which can perhaps provide marginal improvements. In other words, there is still a significant difference in the way our system predicts a grade and how experts derive their decisions. There is only little agreement among the experts and the system, showing that future research could focus on updating the automatic grading process so that it can resemble the human grading process more accurately.

5. Conclusions

In this paper, we addressed the problem of automatic grading of evidence on a chosen discrete scale. We first discussed the grading scale (SORT), the various grading criteria, and some research related to ours. Following that, we described our analysis, which was carried out with the intent of identifying useful factors that influence evidence grades, and the suitability of a supervised machine learning model for this task. Our experiments produced significantly better results than the baseline, and suggested that supervised machine learning has the potential for being applied to this task. We also made some key discoveries regarding the importance of various factors in determining evidence grades. Specifically, we discovered that the publication types of individual articles are useful predictors of evidence grades, and the titles of articles are also useful. However, publication dates (years) and publication venues (*i.e.*, journal names) of individual articles are not useful predictors of evidence grades according to our analysis. Due to the importance of the information regarding the publication types of individual articles in the grade classification process, we attempted to devise an automatic approach for identifying the publication types of medical articles. We showed that a rule-based approach can efficiently identify the publication types of high quality articles such as systematic reviews and randomised controlled trials by utilising information from the article titles, abstracts, and the associated meta-data. Automatic identification of lower quality publication types such as case studies is more challenging, since the article titles and abstracts often do not contain the necessary information.

We applied supervised machine learning with automatically extracted features to perform the grading task. In our model, we applied a sequence of classifiers that attempted to separate A and C grade evidences from B grade ones. We obtained an accuracy of 62.84% using this approach, which was a significant improvement over the baseline. We introduced an evaluation metric, AED, which attempts to estimate the *closeness* of a system's grades to actual grades, and we showed that our sequential classification model achieves improved AEDs compared to the baseline.

To conclude our research on this topic, we conducted a human evaluation and compared the performance of our system with human experts. Our experiments revealed that when human experts are given the same data as our machine learning algorithm, they only have *moderate* agreement regarding the grades. The experiments also revealed that although the performance of the experts is comparable to our system when compared against the gold standard, there are still significant disagreements between the expert assigned grades and the grades assigned by our system. Based on our findings, we can conclude that supervised classification is a promising approach for automatic grade recommendations. Considering the relatively low level of agreement between human-generated and automatically-generated grades, there is still room for modifications/adjustments to the system to increase its agreement with human experts. Importantly, our evaluations suggest that it may not be possible for an evidence grading system to significantly improve its performance using the data that is currently available.

Future research can benefit from the use of more annotated data. The amount of data used in this supervised classification model is relatively small, and this affects the performance of the classifier particularly for the smaller classes such as C. Furthermore, in our grading model we have only incorporated features from individual documents, and we have not utilised multi-document features such as consistency. Automatic extraction of such multi-document features is challenging, and current research in this area is limited. We have explored document level polarity classification for this domain [19], but the accuracies of such techniques are still not sufficiently high to be applied as an intermediate step in the evidence grading task. Future improvements to such techniques may deem them suitable for use in the automatic grading task. Based on the findings of our human evaluations, however, it appears that the amount of improvement that can be achieved is limited. In the future, we would like to extend our human evaluation by involving experienced medical practitioners rather than trained students. Finally, we tested our feature sets and the sequential classification approach using the SORT grading scale, which has three classes. However, the same set of feature sets (*e.g.*, publication types and *n*-grams) and approach (*e.g.*, a sequence of high-precision classifiers) may be applied for evidence grading using other similar scales with more or fewer grades.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This research was jointly funded by Macquarie University and CSIRO. Thanks to Dr. Shannon Casinto, Dr. Krystle Tran, Dr. Claire McVittie, and Hridita Nasim for their efforts in the evaluations.

References

[1] Sackett DL, Rosenberg WMC, Gray JAM, Haynes BR, Richardson WS. Evidence based medicine: what it is and what it isn't. Br Med J 1996;312(7023):71–2.

[2] Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, et al. Analysis of questions asked by family doctors regarding patient care. Br Med J 1999;319(7206):358–61.

[3] Verhoeven AAH, Boerma EJ, de Jong BM. Which literature retrieval method is most effective for GPs? Fam Pract 2000;17(1):30–5.

[4] Smith R. What clinical information do doctors need. Br Med J 1996;313:1062–8.

[5] Westberg EE, Miller RA. The basis for using internet to support the information needs of primary care. J Am Med Inform Assoc 1999;6:6–25.

[6] Dorsch JL. Information needs of rural health professionals: a review of the literature. Bull Med Libr Assoc 2000;88(4):346–54.

[7] McColl A, Smith H, White P, Field J. General practitioner's perceptions of the route to evidence based medicine: a questionnaire survey. Br Med J 1998;316:361–5.

[8] Wilson SM. Impact of the Internet on primary staff care in Glasgow. J Med Internet Res 1999;1(2):E7.

[9] Ely JW, Osheroff JA, Ebell MH, Chambliss ML, Vinson DC, Stevermer JJ, et al. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. Br Med J 2002;324(7339):710.

[10] Coumou H, Meijman F. How do primary care physicians seek answers to clinical questions? A literature review. J Med Libr Assoc 2006;94(1):55–60.

[11] West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence; 2002 http://www.ncbi.nlm.nih.gov/books/NBK33881/ [accessed 10.11.14].

[12] Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. Br Med J 2004;328(7454):1490–7.

[13] Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman J, et al. Strength of Recommendation Taxonomy (SORT): patient-centred approach to grading evidence in the medical literature. Am Fam Physician 2004;69(3):548–56.

[14] Mollá D. A corpus for evidence based medicine summarisation. In: Indurkhya N, Zwarts S, editors. Proceedings of the Australasian language technology association workshop, vol. 8. Melbourne, Australia: ALTA, University of Melbourne; 2010.

[15] Elkin P, Brown S, Bauer B, Husser C, Carruth W, Bergstrom L, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inf Dec Mak 2005;5(13), 13:1-13:7.

[16] Ruiz-Rico F, Vicedo JL, Rubio-Sánchez M-C. MEDLINE abstracts classification based on noun phrases extraction. In: Fred A, Filipe J, Gamboa H, editors. Proceedings of biomedical engineering systems and technologies, international joint conference (BIOSTEC), communications in computer and information science. Berlin/Heidelberg: Springer; 2009. p. 507–19.

[17] Sasaki Y, Rea B, Ananiadou S. Clinical text classification under the open and closed topic assumptions. Int J Data Mining Bioinform 2009;3(3):229–313.

[18] Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence based medicine. BMC Bioinform 2011;12(Suppl. 2). S5:1–S5:10.

[19] Sarker A, Mollá D, Paris C. Outcome polarity identification of medical papers. In: Mollá D, Martinez D, editors. Proceedings of the Australasian language technology association workshop. Canberra, Australia: ALTA, Australian National University; 2011. p. 105–14.

[20] Sarker A, Mollá D, Paris C. An approach for automatic multi-label classification of medical sentences. In: Suominen H, editor. Proceedings of the fourth international workshop on health document text mining and information analysis, vol. 4. Sydney, NSW, Australia: NICTA; 2013.

[21] Skeppstedt M, Kvist M, Nilson GH, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. J Biomed Inform 2014;49:148–58.

[22] Goetz T, von der Lieth C-W. Pubfinder: a tool for improving retrieval rate of relevant PubMed abstracts. Nucleic Acids Res 2005;33:W774–8.

[23] Plikus MV, Zhang Z, Chuong C-M. PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. BMC Bioinform 2006;7(1):424–39.

[24] Tang T, Hawking D, Sankaranarayana R, Griffiths K, Craswell N. Quality-oriented search for depression portals. In: Boughanem M, Berrut C, Mothe J, Soule-Dupuy C, editors. Advances in information retrieval, vol. 5478 of Lecture Notes in Computer Science. Berlin/Heidelberg: Springer; 2009. p. 637–44 [Chapter 60].

[25] Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes BR. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc 2009;16(1):25–31.

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

[27] Kim S, Choi J. An SVM-based high-quality article classifier for systematic reviews. J Biomed Inform 2014;47:153–9.

[28] Adeva JJG, Atxa JMP, Carrillo MU, Zengotitabengoa A. Automatic text classification to support systematic reviews in medicine. Exp Syst Appl 2014;41(4):1498–508.

[29] Greenhalgh T. How to read a paper: the basics of evidence-based medicine. 3rd ed. Chichester, West Sussex, United Kingdom: Wiley; 2006.

[30] Demner-Fushman D, Lin JJ. Answering clinical questions with knowledge-based and statistical techniques. Comput Linguist 2007;33(1):63–103.

[31] Mollá D, Santiago-Martinez ME. Development of a corpus for evidence based medicine summarisation. In: Mollá D, Martinez D, editors. Proceedings of the Australasian language technology association workshop, vol. 9. Canberra, Australia: ALTA, Australian National University; 2011. p. 86–94.

[32] Sarker A, Mollá-Aliod D, Paris C. Towards automatic grading of evidence. In: Nytrø O, Slaughter L, Moen H, editors. Proceedings of LOUHI: third international workshop on health document text mining and information analysis, vol. 3. Bled, Slovenia: CEUR Workshop Proceedings; 2011. p. 51–8.

[33] Porter MF. An algorithm for suffix stripping. Program 1980;14(3):130–7.

[34] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor 2009;1(11):10–8.

[35] Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. Appl Stat 1992;41(1):191–201.

[36] Quinlan JR. C4.5: programs for machine learning. Burlington, MA: Morgan Kaufmann Publishers Inc; 1993.

[37] Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992;9:309–47.

[38] Platt JC. Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel methods: support vector learning. Cambridge, MA: MIT Press; 1998. p. 185–208.

[39] Sarker A, Mollá-Aliod D. A rule-based approach for automatic identification of publication types of medical papers. In: Scholer F, Trotman A, Turpin A, editors. Proceedings of the ADCS annual symposium, vol. 15. Melbourne, Australia: ADCS, RMIT University; 2010. p. 84–8.

[40] Compton P, Jansen R. Knowledge in context: a strategy for expert system maintenance. In: Barter CJ, Brooks MJ, editors. AI'88, Vol. 406 of Lecture Notes in Computer Science. Berlin/Heidelberg: Springer; 1990. p. 292–306.

[41] Hunt DL, McKibbon KA. Locating and appraising systematic reviews. Ann Intern Med 1997;126(7):532–8.

[42] Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. Br Med J 2005;330(7482):68–73.

[43] Mollá D, Sarker A. Automatic grading of evidence: the 2011 ALTA shared task. In: Mollá D, Martinez D, editors. Proceedings of the Australasian language technology association workshop. Canberra, Australia: ALTA, Australian National University; 2011. p. 4–8.

[44] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Bakken S, editor. Proceedings of the American medical informatics association annual symposium. Washington, DC: AMIA; 2001. p. 17–21.

[45] Uzuner O, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. J Am Med Inform Assoc 2009;16:109–15.

[46] Schapire RE. The strength of weak learnability. Mach Learn 1990;5:197–227.

[47] Carletta J. Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 1996;22(2):249–54.

[48] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74.