Health Informatics Journal

Original Article

Health Informatics Journal
1–17
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1460458217719560
journals.sagepub.com/home/jhi

Context-aware grading of quality evidences for evidence-based decision-making

Muhammad Afzal and Maqbool Hussain
Sejong University, South Korea; Kyung Hee University, South Korea

Robert Brian Haynes
McMaster University, Canada

Sungyoung Lee
Kyung Hee University, South Korea

Abstract

Processing huge repository of medical literature for extracting relevant and high-quality evidences demands efficient evidence support methods. We aim at developing methods to automate the process of finding quality evidences from a plethora of literature documents and grade them according to the context (local condition). We propose a two-level methodology for quality recognition and grading of evidences. First, quality is recognized using quality recognition model; second, context-aware grading of evidences is accomplished. Using 10-fold cross-validation, the proposed quality recognition model achieved an accuracy of 92.14 percent and improved the baseline system accuracy by about 24 percent. The proposed context-aware grading method graded 808 out of 1354 test evidences as highly beneficial for treatment purpose. This infers that around 60 percent evidences shall be given more importance as compared to the other 40 percent evidences. The inclusion of context in recommendation of evidence makes the process of evidence-based decision-making "situation-aware."

Keywords

context-aware evidence grading, evidence-based medicine, evidence-based practice, evidence informed decision, quality recognition

Introduction

There is an exponential growth in the medical literature, and medical practitioners are finding it difficult to obtain the most relevant information in their limited time span. Young physicians,

Corresponding author:
Sungyoung Lee, Department of Computer Science and Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si 446-701, Gyeonggi-do, Korea.
Email: sylee@oslab.khu.ac.kr

particularly, are open to innovation, but they seek to minimize the costs either to themselves or to their patients.[1] Without automation, processing of huge amount of literature is a costly and a challenging task for even the existing systems, leave aside manual efforts of the medical practitioners. Practicing evidence-based medicine[2–4] requires the medical practitioners to extract high-quality evidence from published research in addition to their own knowledge and experience.[2] Finding high-quality evidence is essential for successful practice,[3] but medical practitioners face many barriers in using evidence-based answers at point-of-care.[4] If done at all, most of the time, seeking best evidence is done manually.[5] It requires a lot of manual computation time in order to reach to the desired quality appraised evidences. The importance of recognizing and appraising the evidences can be realized from the fact that more than 100 grading scales are in use today as reported in Agency of Healthcare Research and Quality research report.[6] Regardless of a grading scale, the strength of computing evidences should consider three key elements: quality, quantity, and consistency. A few of them such as Grading of Recommendations Assessment, Development, and Evaluation (GRADE)[7,8] and Strength Of Recommendation Taxonomy (SORT)[9] focus on developing guidelines for quality of evidences and strength of recommendations. GRADE provides the definitions for grading the quality of the evidence on four levels: high, moderate, low, and very low. SORT, on the other hand, provides a taxonomy to determine the strength of the recommendation of a body of evidence based on three ratings: A (strong), B (moderate), and C (weak).

Currently, some approaches[10–12] focus on query building to find information resources but lack automatic appraisal of evidence quality. Using Boolean approaches with search filters, “hedges” can improve the retrieval of clinically relevant and scientifically sound studies from MEDLINE and similar databases,[13,14] but the statistical approaches[3,5] presented a proof of better accuracy in recognizing quality articles as compared to Boolean approaches. Very recently, Sarker et al.[15] presented an approach of evidence quality prediction through supervised classification model. The approach uses the SORT[9] to grade the evidences. A number of other approaches[16–18] are proposed in the area of text classification. Ruiz-Rico et al.[16] combine the existing techniques innovatively for the classification of MEDLINE abstracts based on a noun phrase extraction. Kim and Choi[18] provide automatic classification of key sentences to support evidence-based medicine. A support vector machine (SVM)-based approach is presented for systematic review of related high-quality article classification.[18] Domain-specific post-retrieval re-ranking approach[19] is proposed in the domain of depression that attempts to re-rank the articles returned by the search engine.

The investigation leads us to the conclusion of utilizing quality-based context-aware graded evidences in the evidence-based decision-making process. The strengths of the existing work motivated us to explore improvements in the area of automatic quality processing and grade computation. Based on this motivation, we formulated a consistent two-level methodology: quality evidence recognition (level 1) and context-aware grading (level 2). The methodology is capable of identifying high-quality evidence for recommendation to the medical practitioners or caregivers, or even patients according to their contexts.

We differentiate our proposed approach with the existing ones in the following areas: (a) automatic extraction of metadata features and standardization for improved quality prediction through a supervised classification model called quality recognition model (QRM) and (b) context elements identification, mappings, and aggregation to grade evidences through our proposed method called context-aware grading (CAG). In our approach, we exploited the strengths of existing approaches[3] using the data features (title, abstract) and metadata features (MeSH terms and publication type) for developing the QRM model. We adhere to the suggestions of GRADE[8] and SORT[9] grading scales in CAG-based grading with the exception of involving decision-making context influenced from the conceptual framework for context-based evidence-based decision-making.[20,21]

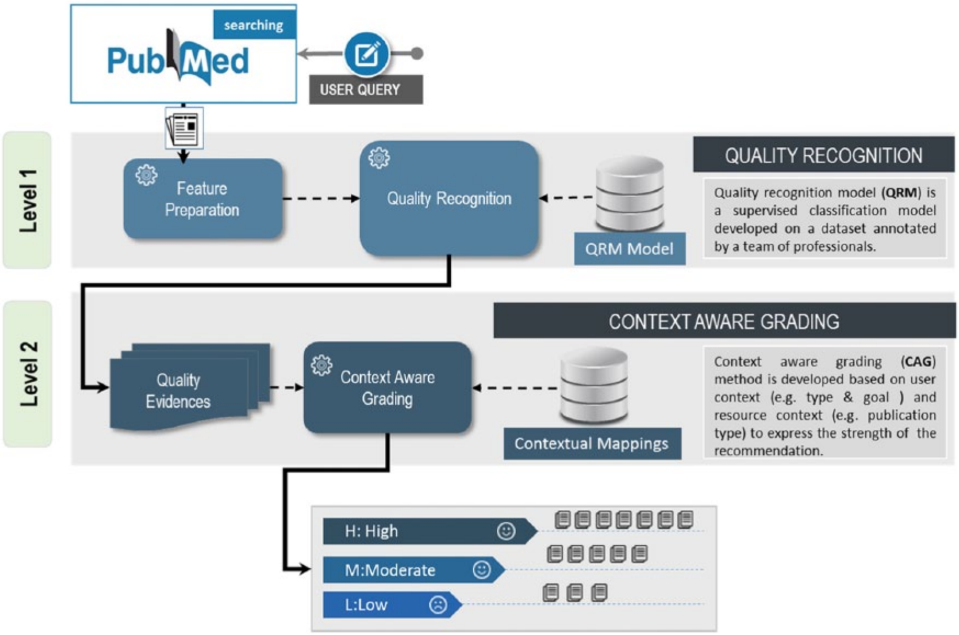
searching
PubMed
USER QUERY
Level 1
Feature Preparation
Quality Recognition
QRM Model
QUALITY RECOGNITION
Quality recognition model (QRM) is a supervised classification model developed on a dataset annotated by a team of professionals.
Level 2
Quality Evidences
Context Aware Grading
Contextual Mappings
CONTEXT AWARE GRADING
Context aware grading (CAG) method is developed based on user context (e.g. type & goal) and resource context (e.g. publication type) to express the strength of the recommendation.
H: High
M:Moderate
L:Low

Figure 1. Two-level evidence evaluation: quality recognition and context-aware grading.

Objective

Physicians, whether serving individual patients or populations, always have sought to base their decisions and actions on the best possible evidence.[22] Based on evidence-adaptive clinical decision support systems,[23] the researchers and developers need to customize the literature-based evidence for local conditions. Adhering to these needs and recommendations, the task of finding best possible evidence from the literature, customized to the local conditions, becomes a priority. We aim at developing methods to automate the process of finding quality (best possible) evidences from a plethora of literature documents and grade them according the context (local condition). Previously, we developed automated methods for knowledge-based query construction, assisting the medical practitioners in query preparation.[24,25] This work is a step forward to focus on evidence quality evaluation and CAG.

Materials and methods

We propose a hierarchical strategy for the evaluation of quality evidence at two different levels as depicted in Figure 1. At first level, the quality of evidences is recognized on the basis of methodological rigorousness through the QRM classification model. If an article passes the criteria of being methodological rigorous, the article is recognized as a quality evidence. At second level, the recognized quality evidences are graded on the basis of user and resource contextual information using CAG method.

Level 1: quality recognition

Prior describing the method of quality evidence recognition, it is necessary to agree upon quality parameters. Quality of an evidence and what makes an evidence a quality evidence for a user are two different considerations. The definitions of a quality evidence are available in the literature for clinical care. SORT[9] includes ratings of A, B, or C for the strength of recommendation for a body of evidence. The analogy of a best evidence aligned with category "A" of SORT grading which is defined as "Recommendation based on consistent and good quality patient oriented evidence."[9] Good-quality patient-oriented evidence has different meanings with respect to different purposes such as diagnosis, treatment, and prognosis. For treatment purposes, the meaning of good quality evidence is provided in Definition 1.

Definition 1. "Systematic Review or meta-analysis of randomized controlled trials (RCTs) with consistent findings or high-quality individual RCT.[9]"

In a study protocol,[14] an article is considered as high quality if it passes the "methodological rigorous" criteria. Methodological rigorous article for different purposes has different meanings. For treatment purpose, a methodological rigor article is defined as in Definition 2.

Definition 2. "Random allocation of participants to comparison groups, outcome assessment of at least 80% of those entering the investigation accounted for in 1 major analysis at any given follow up assessment, and analysis consistent with study design."[14]

For this study, Definition 2 is considered for quality evaluation of the evidences. For quality evaluations, we develop a supervised classification model called QRM. We follow the steps of data collection, feature selection, corpus preparation, algorithm selection, and parameter tuning for QRM development.

Data collection. We use the data that were manually created by a team of specialized experts for the purpose of clinical query filters in PubMed.[14] The data collection consists of 50,594 MEDLINE documents, of which 49,028 documents are unique. The collection is classified across four dimensions: format (O = original study, R = review, GM = general and miscellaneous articles, and CR = case report), human healthcare interest (yes/no), scientific rigor (yes/no), and purpose (diagnosis, etiology, prognosis, treatment, economic studies, reviews, and clinical predication guides). Among 50,594 documents, 3363 are labeled as being scientifically rigorous.

Feature selection. Feature selection plays an important role in predicting performance. From the existing studies, we come across features including data features (title, abstract) and metadata features (MeSH terms, publication type, publication year, publication venue, and publication authors). In some studies, concepts used are semantic prediction, UMLS concepts, and UMLS relation in predictions.[3] The data features that are used in earlier studies[3,15] have proved their importance. Publication type (one of the metadata features) is the most important feature reported from the same studies. MeSH terms is also reported in Kilicoglu et al.[3] as one of the important contributors. Other metadata features including publication year and publication venue are reported as less significant features to affect the classification accuracy. In our experiments, we also found that publication year, venue, and author are the least significant in metadata feature list as compared to other metadata. Finally, we select four features; title, abstract, MeSH, and publication type.

Corpus preparation. Getting the data for the selected features, we implemented eUtils service API[26] to retrieve the documents from the PubMed database. The documents are processed to get individual features and store to MS SQL Server database for experimentation.

Word vector preparation. The selected features are composed of "bag-of-words" which need to be cleaned prior to use for learning methods. We apply the text processing method "process

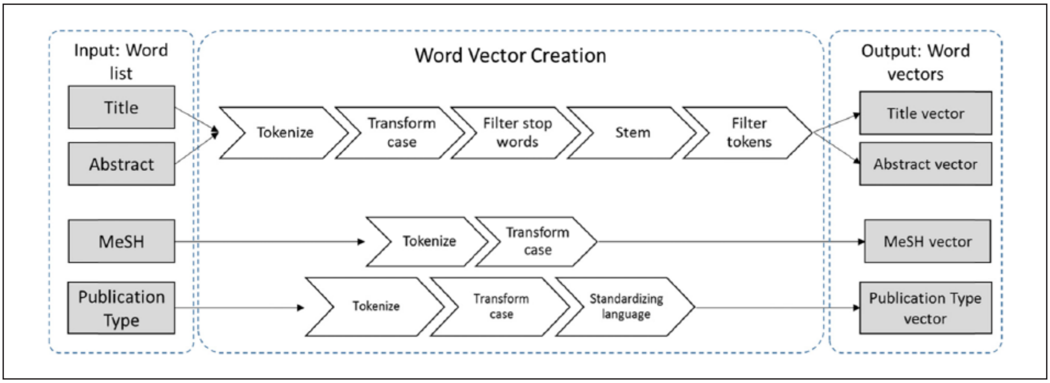

Figure 2. Process of word vectors (title, abstract, MeSH, Publication type) creation.

documents from data" using TD-IDF[27] in RapidMiner tool.[28] To remove the least and too frequent words from the list, the prune method is set to absolute with below absolute = 2 and above absolute = 100. The preprocessing steps are described in Figure 2 to get the final word vector. Title and abstract are first tokenized, transform the case, remove English stop words, stemming the words using the Porter[29] stemmer, and filtering the tokens by length having minimum characters of 2 and maximum of 999 characters. Unlike data features (title, abstract), vectors of the metadata features (MeSH, publication type) are created through tokenization and case transformation as there was no need to remove stop words and stemming. The complete workflow designed in RapidMinor tool is made available on a public domain (https://www.myexperiment.org/workflows/4958.html?version=1) for the general public to reuse for their own experimentation.

Standardizing language of publication type. The publication types text retrieved through eUtils API[26] are not consistent with the vocabulary of publication types provided by PubMed. Publication types found in PubMed are reported in count as 73,[30] which is quite less than the count 248 returned for the documents in our selected data set. Algorithm 1 mapped the inconsistent publication types to standard publication types taking the list of articles as input. The publication type of each article is a string which may contain one or more than one publication types. Using getPType() function, the string is parsed into a list of atomic publication type. For each atomic publication type, rank is determined with getRank() function. The getRank() function finds the rank of each publication type in R mapping table. Ranks of each publication type are dependent on the goal of the study such as diagnosis, treatment, and others. The ranks for publication types based on their importance and effectiveness are derived from the literature evidences[8,9,14,31,32] as shown in Table 1. The rank value 1 shows the highest rank of publication types of the treatment goal with respect to their importance. For instance, meta-analysis of RCTs is considered the most important publication type for treatment, so it is ranked on top by assigning value 1. Table 1 is not an exhaustive representation to have a rank entry for each possible publication type rather it holds the most prominent and influential publication types for the treatment goal.

Parameter setting. Rigorous recognition on the articles is a binary classification problem. We surveyed multiple methods from different sources and selected some that work well with text categorization tasks.[33,34] For the chosen methods, Naïve Bayes (NB) kernel,[35] k-nearest neighbor (kNN),[36] SVM linear,[37] and decision tree (DT),[38] we tested the performance at different parameter settings. NB is experimented with kernel values 5, 10, and 15 with a minimum

Algorithm 1. Standardizing language of publication types.

```
Begin
    inputs: $A - \{a_1, a_2, \ldots, a_n\}$ ;//the list of articles
    output: $A' - \{a_1, a_2, \ldots, a_n\}$ ;//the list of articles with standardized publication type
    Let;
            pt represents publication type;
            rank represents the rank of pt;
            tempRank = 0; //holds the previous rank temporarily for comparison
            spt represents the standardized publication type;
    for each a in A
        do
            pt ← a.getPublicationType();
            rank ← getRank(pt, R); //where R is the rank table for publication types.
            if (rank > tempRank)
                tempRank ← rank;
                spt ← pt;
            endif
        while (a.getPublicationType exists)
        a.PublicationType ← spt;
        A'.add(a);
    endfor
    return A';
End
```

Table 1. Rank values of publication types (1 shows the highest and 4 is the lowest).

Publication type	Rank
Meta-analysis of RCTs	1
Systematic Review of RCTs	2
RCT	3
Meta-analysis of CTs	4
Systematic review of CTs	5
CT	6
Cohort study	7
Case-control study/report	7
Guidelines	8
Opinion	9
Observational study	10
Any other publication type	11

RCT: randomized controlled trial; CT: control trial.

Table 2. SVM complexity cost (*C*) parameter setting and the corresponding results.

Method	Parameter	Value	Result (accuracy)
SVM	Complex cost parameter *C*	−0.2	80.15
		−0.1	80.15
		0.0	80.15
		0.1	75.38
		0.2	75.38

SVM: support vector machine.

Table 3. Performance of machine learning algorithms in terms of accumulative sum score of F-measure, accuracy, and AUC using data and metadata features with standard publication type on training and development test data.

Algorithm/ criteria	Training			Testing			Sum score
	F-measure	Accuracy	AUC	F-measure	Accuracy	AUC	
SVM	0.849	0.771	0.807	0.870	0.785	0.735	**4.818**
DT	0.914	0.883	0.969	0.289	0.316	0.762	4.134
NB	0.835	0.764	0.752	0.721	0.602	0.548	4.223
kNN	0.812	0.707	0.782	0.847	0.752	0.777	4.678

The bold value shows that SVM secured the highest accumulative sum score thus selected for the experiment;
AUC: area under curve; SVM: support vector machine; kNN: k-nearest neighbor; DT: decision tree; NB: Naïve Bayes.

bandwidth of 0.1, and it was found that kernel value = 10 showed slightly better performance. Finding the best value of *k* for kNN, we experimented *k* values in the range of 1–20 for odd values and found $k = 5$ with measure type = NumericalMeasure and numerical measure = CosineSimilarity as better setting. DT performed better on RapidMiner default settings with confidence value of 0.25 for the pessimistic error calculation of pruning. SVM with different parameter settings is tested to find the best value of complex cost parameter *C*. Values less than 0.0 showed similar results to $C = 0.0$. Similarly, values greater than 0.1 produces similar results to $C = 0.1$. The kernel cache value is set to 200 and maximum iterations are set to 100,000. Finally, we were left with $C = 0.0$ and $C = 0.1$ to choose from; however, $C = 0.0$ for our experiment produced better results as compared to $C = 0.1$. The performance for different parameter settings of SVM is shown in Table 2.

Method selection. We choose a subset that is treatment-related documents of our selected data set for the experiment to find quality evidence. The subset includes 6882 documents out of which 4999 are labeled as "non-rigor" and 1883 are labeled as "rigor." We determine the performance of chosen methods on F-measure, accuracy, and area under curve (AUC) criteria (Table 3). F-measure and accuracy are included to judge how accurately the rigorousness of an article is predicted and AUC criterion is included to judge how consistently they are predicted. In the literature, it is reported that AUC is statistically consistent and more discriminant than accuracy.[39,40] SVM classifier performs the best in accuracy than DT and kNN; however, it is lower than NB. AUC of SVM was lower than DT; however, it was higher than NB and kNN. Overall, SVM showed better overall ranking score than all other competing algorithms and kNN showed poor performance as compared to others. Because of the higher performance, SVM is chosen for the development of QRM.

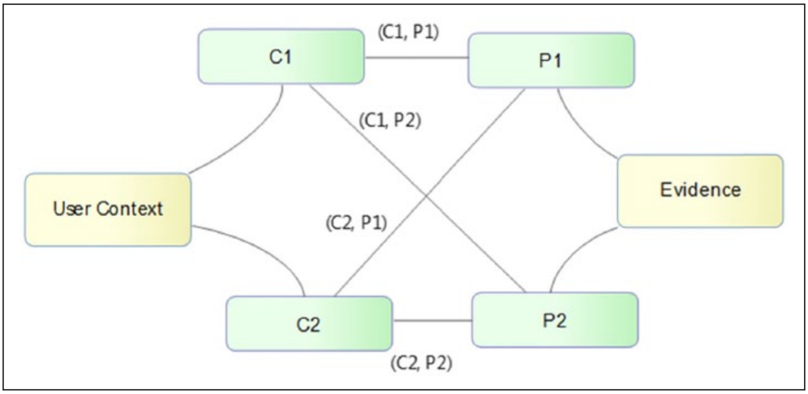

Figure 3. User context mapping with evidence properties.

Level 2: context-aware evidence grading

Evidence recognition on the basis of user query and statistical methods may not fully determine the user preferred evidences. The statistical approach described in Level 1: quality recognition recognizes the evidence quality on the basis of methodological rigorousness, which is a necessary step; however, it is not sufficient to reflect the user perspective. In order to reflect the user perspective, we conceive the user context in relation to a resource (evidence) context. Context has a vast meaning, it exhibits its characteristics according to the goal and application domain. Vebert et al.[41] present a context framework that identities relevant context dimensions for technology enhanced learning applications. We derive the classification of context information that is relevant to evidence-based clinical applications. In evidence-based clinical applications, user's main objective is to interact with online resources for finding support in evidence-based decision-making. We derive the contextual elements from the context framework in Dobrow et al.,[20] Verbert et al.,[41] and Rycroft-Malone[42] that is relevant to the objective of evidence-based clinical applications. User context has multiple elements such as basic information which shows user educational level, background is the experience of the user, goal shows short-term learning or long-term learning, interest represents the preferences, and learning style is the pattern of user learning such as textual and visual. An evidence possess multiple properties such as the publication type, publication avenue (journal, book, etc.), and year of publication. For grading an evidence, we design a method as shown in Figure 3 and describe in Algorithm 2, which evaluates an evidence on the basis of different user context elements.

First, the properties associated with the evidences are extracted and each property is evaluated with each of the elements of different contexts. For instance, an evidence E has properties P_1 and P_2 and user U who is interested in E possesses the contexts C_1 and C_2. The algorithm first evaluates the property P_1 of E according to C_1 and C_2 by putting the grading value from expert-based contextual mappings. The process is repeated for property P_2 in the similar way as that of P_1. If there are more contexts or properties, this process will occur for all of them. In Figure 3, user contexts C_1 and C_2 are mapped to the two properties P_1 and P_2 of an evidence. The mappings of context to evidence are made based on two type of analysis: literature-based and expert-based. We investigate the well-known study protocols and grading systems[8,9,14] and two senior physicians to grade evidence with different contexts. The grade values are chosen as L=low, M=Medium, H=High, and U=Unknown, for each user context against a property of an evidence. The grade values for evidences are stored in the form of matrix where rows represent the user context elements and columns represent the properties of evidence as shown in Table 4.

Algorithm 2. Grading evidences based on user context.

```
Begin
    input: $E - \{e_1, e_2, \ldots, e_n\}$ ;//the list of rigor evidences
    output: $GE - \{\{e_1, g_1\}, \{e_2, g_2\}, \ldots, \{e_n, g_n\}\}$; //where g represents the grades h, m, l, u.
    Let;
        $C - \{c_1, c_2, \supset, c_n\}$; //current context
        $P - \{p_1, p_2, \ldots, p_n\}$; //properties of E
        $G - \{g_1, g_2, \supset, g_n\}$; //properties of E
    for each e in E
        for each p in P
            for each c in C
                grade ← computeGrade(p, c);
                G.add(grade);
            endfor
        endfor
        finalGrade ← getHighestGrade(G);
        GE.add (e, finalGrade);
    endfor
    return GE;
End
```

Table 4. Grade value population for an evidence with respect to contexts.

Context\evidence	P_1	P_2	…	P_n
C_1	(*H* or *M* or *L* or *U*)	(*H* or *M* or *L* or *U*)	…	(*H* or *M* or *L* or *U*)
C_2	(*H* or *M* or *L* or *U*)	(*H* or *M* or *L* or *U*)	…	(*H* or *M* or *L* or *U*)
…	…	…	…	…
C_n	(*H* or *M* or *L* or *U*)	(*H* or *M* or *L* or *U*)	…	(*H* or *M* or *L* or *U*)

Context aggregation. Based on the grade values, the aggregate contextual grade values are inferred from each column of Table 5. The aggregate contextual grade values accumulatively make the aggregate contextual vector. Table 5 shows the aggregate contextual grade vector (ACGV) consisting of aggregate contextual grade values. The aggregate contextual grade values are inferred using a simple rule of picking the highest rank value among *H, M, L*, and *U* in the respective column. Highest to lowest definition is provided in equation (1). For instance, *L* is selected as the aggregate value because $L > U$.

$$H > M > L > U \tag{1}$$

Final grade value (FGV) is inferred from the values of ACGV on the same rule as in equation (1). For the user explanation, the FGV value is interpreted according to equation (2)

Table 5. Aggregate contextual grade values and vector.

Context\evidence	P_1	P_2	…	P_n	
C_1	(*H* or *M* or *L* or *U*)	(*H* or *M* or *L* or *U*)	…	(*H* or *M* or *L* or *U*)	
C_2	(*H* or *M* or *L* or *U*)	(*H* or *M* or *L* or *U*)	…	(*H* or *M* or *L* or *U*)	
…	…	…	…	…	
C_n	(*H* or *M* or *L* or *U*)	(*H* or *M* or *L* or *U*)	…	(*H* or *M* or *L* or *U*)	
Aggregate contextual grade values	(*H* or *M* or *L* or *U*)	(*H* or *M* or *L* or *U*)	…	(*H* or *M* or *L* or *U*)	(*H* or *M* or *L* or *U*)

Aggregate Contextual Grade Vector

Final Grade Value

$$F(\mathrm{FGV}) = \begin{cases} \text{if } H \rightarrow \text{highly beneficial} \\ \text{if } M \rightarrow \text{moderate beneficial} \\ \text{if } L \rightarrow \text{less beneficial} \\ \text{if } U \rightarrow \text{unknown} \end{cases} \tag{2}$$

Experimental results

As mentioned, the proposed methodology is implemented in hierarchical fashion. The implementation framework for conducting different experiments is described in Figure 4. Using this implementation framework, three types of experiments are conducted where two of the experiments are pertinent to QRM and one of them is related to CAG:

Experiment 1. Demonstration of QRM performance on development test data set for four different features and their combinations.

Experiment 2. Comparison of QRM performance on publication type feature in default (non-standardized) and in a standardized form.

Experiment 3. Contextual grading results CAG method on the basis of "physician interested in treatment" case study.

Experiment 1: QRM performance on development test data set

We here present the classification results obtained in the 10-fold cross-validation performed on the training set of 5682 documents and development test set of 1300 documents. In Table 6, accuracies on different features are presented. Using publication type feature, QRM produced second better results (79%) for testing documents. At training stage, the combination of three features (title, abstract, and standard publication type) stand second with 89.7percent accuracy. Title feature remains the lowest in both training and testing cases and abstract feature remains second lowest. At training stage, MeSH feature performed better than standardized publication type (SPT), while at testing stage, it is reversed. Overall, in both training and testing, QRM performed exceptionally well on the combination of all features (title, abstract, MeSH, and standard publication type).

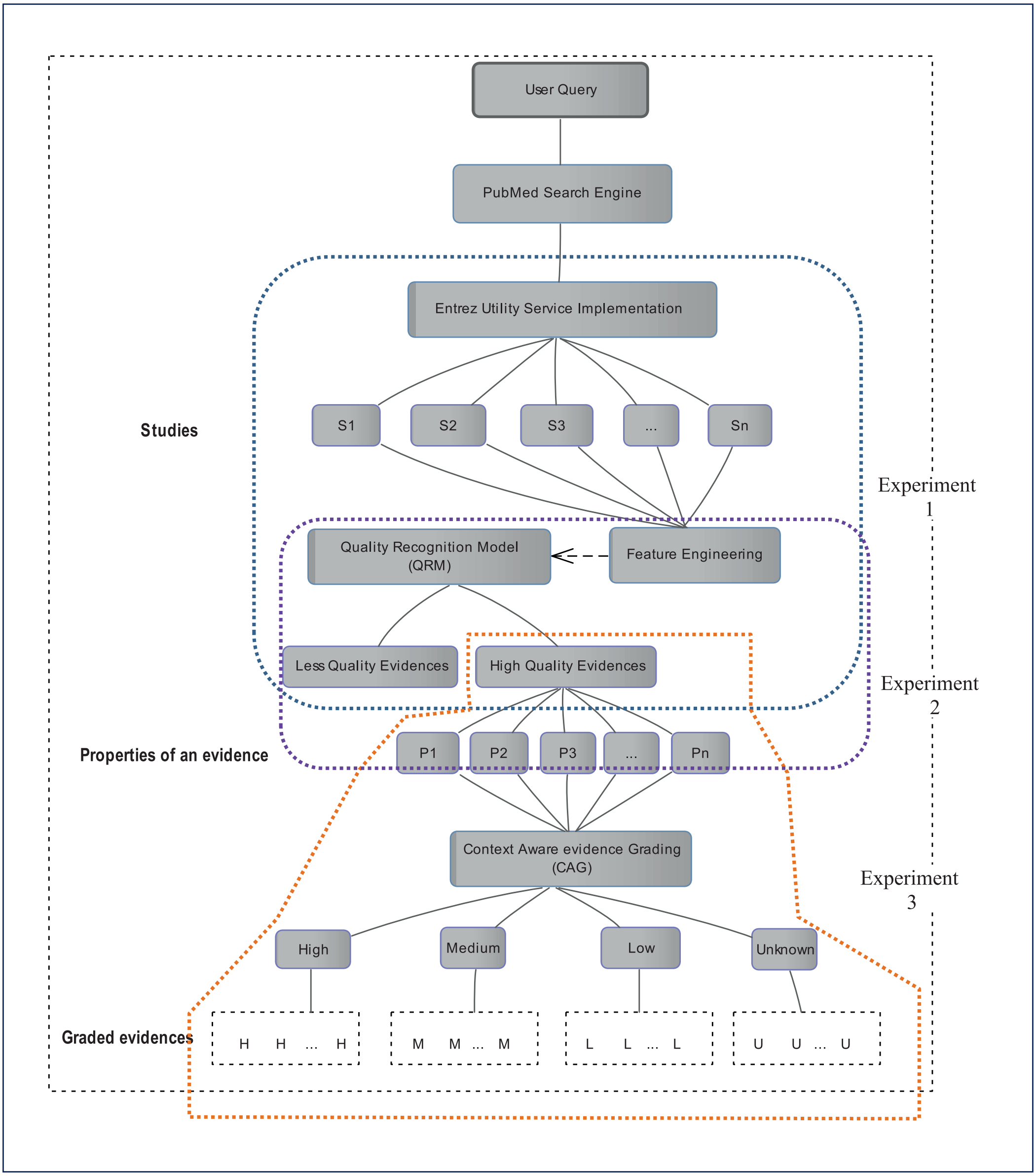

Figure 4. Proposed methodology implementation framework for conducting experiments.

Table 6. QRM accuracy on features separately and their overall combination.

Features	Title (%)	Abstract (%)	MeSH (%)	SPT (%)	Title, abstract, SPT (%)	All (%)
Training	76.28	82.81	86.4	85.71	89.7	92.14
Testing	73.31	75.46	76.9	79	78.15	80.15

QRM: quality recognition model.

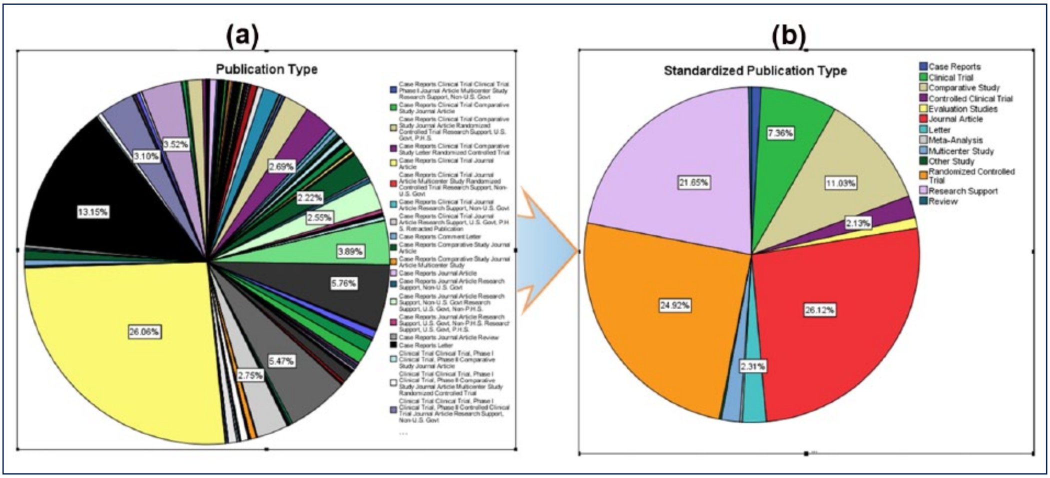

Figure 5. (a) Publication types and (b) standardized publication types.

Table 7. QRM performance on standard and non-standard publication types.

Recall (%)		Precision (%)		Accuracy (%)	
Non-standard	Standard	Non-standard	Standard	Non-standard	Standard
66.07	68.27	40.81	80.52	61.56	85.71

QRM: quality recognition model.

Experiment 2: QRM performance on standardized and non-standardized publication types

With Entrez eUtils service, we get the publication types for the 5682 articles in our training data set. Overall, 249 different variations are found in publication types as shown in Figure 5(a). Using algorithm 1, we normalized the 249 variations into 13 standard publication types having different frequencies as shown in Figure 5(b). We experimented the performance of QRM on 5682 documents on publication type both in default and standard form. Journal article, RCTs, and research reports are in the higher distributions of 1484, 1416, and 1230, respectively, depicted in Figure 5(b). The standard form publication type produced better results as described in Table 7. QRM performed exceptionally on standard publication type. The recall value showed about 2 percent, precision about 40 percent, and accuracy about 24 percent increase in the standardized form.

Experiment 3: CAG results for "physician interested in treatment" case study

The QRM model predicted 1355 out of 5682 documents as Rigor. Using equations (1) and (2), all 1355 documents are assigned aggregate value for the contexts as; user type = physician and user goal = treatment. As shown in Table 8, Out of 1355 documents, about 60 percent documents are graded as *H* which means highly beneficial for the physician to benefit in treatment-related clinical decisions. Other approximately 20 percent are graded as *M* (moderate beneficial), 8 percent as *L* (low beneficial), and 13 percent as *U* (unknown).

Table 8. Evidence grading distribution among high, moderate, low, and unknown.

Grade	*H*	*M*	*L*	*U*
No. of evidences	808 (59.63%)	266 (19.63%)	110 (8.12%)	170 (12.55%)

Table 9. Queries designed for retrieving evidentiary documents from PubMed database.

Query no.	Query terms
Q1	(Oral Cavity) AND (cancer AND head neck) AND (Therapy/Broad [filter])
Q2	(Oral Cavity) AND (T1 OR Clinical Stage 1) AND (cancer AND head neck) AND (Therapy/Broad[filter])
Q3	(Oral Cavity) AND (T3 OR Clinical Stage 3) AND (Squamous cell carcinoma) AND (cancer AND head neck) AND (Therapy/Broad[filter])

Table 10. eUtils web service URLs for retrieving evidentiary documents from PubMed database.

Query No	PubMed service URLs
Q1	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eSearch.fcgi?/db=pubmed term = Q1
Q2	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eSearch.fcgi?/db=pubmed term = Q2
Q3	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eSearch.fcgi?/db=pubmed term = Q3

URLs: universal resource locators.

The higher number of *H* graded evidence complements the QRM performance and also it confers the definitions of quality (Definition 2). Moreover, these evidences need to be evaluated from the experts in particular domains. In this study, since the documents are not related to any specific domain so human evaluation is not feasible to conduct.

Case study: results evaluation

To assess the performance of the models on a field test data, we perform experimentation on a real-world case study. The study is related to the retrieval of evidentiary documents pertinent to head neck cancer treatment decision-making. We utilize eUtils functions of PubMed and run the three types of queries described in Table 9. The reason of having three different queries is to test queries of various scales, small, medium, and large.

In the next step, universal resource locators (URLs) are generated as described in Table 10 for execution of these queries on the PubMed search service.

All the queries are executed on PubMed database and retrieve the evidentiary documents. The retrieved documents are processed and the processed documents are passed through the trained QRM. On the average, 17.53 documents are filtered out from the final set of evidentiary document list as shown in Table 11.

Discussion

QRM model

Considering appraisal using the SORT scale as performed in a very recent article,[15] we compared the results on the basis of same feature set. Unlike,[15] our appraisal model evaluates articles on two

Table 11. Performance results of QRM in terms of filter rate.

Query	No. of documents retrieved	Pred. rigor	Pred. non-rigor	Filter rate (%)
Q1	2218	1771	447	20.15
Q2	228	192	36	15.79
Q3	168	140	28	16.66

QRM: quality recognition model.

classes "rigor" and "non-rigor." The gold standard data set is not the same; however, we here present the comparison on the basis of feature set and machine learning algorithm equivalency. MeSH terms feature set is not included in their experiment, which produced better results for most of the feature sets in our experiment. Repeating the same method with SVM classifier proposed in Sarker et al.[15] on our data set and comparing the results, we obtained approximately 3 percent better results (89.7% increases to 92.14%) for the feature set that includes MeSH terms at the training stage as described in Table 6. At the testing stage, QRM showed 2 percent improved results (78.15% increases to 80.15%) as shown in Table 6.

CAG model

The proposed approach is different from existing approaches in terms of user context consideration for evidence grading. The existing approach[15] uses SORT[9] taxonomy to grade the evidences. SORT taxonomy is a strong system to determine a grade for an evidence; however, it may not decide whether the evidence fits in user context or not. Our approach introduced CAG; a scalable and robust method to include contexts applicable in a particular domain. The only requirement for the extension is the identification of values for the contextual mapping tables. The aggregation contextual vector (section "Context aggregation") parsing method is independent of contextual mappings' identification and population in the tables.

Feature significance

During evaluations, we noticed that publication type is the most influential feature to contribute in determining the quality of an article. This publication type feature has the highest accuracy level among the pool of evaluated features especially when it is transformed into a standard form as shown in experiment 2. In addition to publication type, the metadata feature "MeSH terms" has also produced good results. By combining both publication type and MeSH term features with data features; title and abstract produced the best and stable results across majority of the machine learning algorithms.

Limitations

The proposed CAG method requires prior contextual mappings for the aggregate vector generation. The proposed method will not be able to grade evidences where mappings of user context against the properties of evidences are not available. This limitation can be overcome by conducting a survey on a larger scale to cover multiple user contexts with maximum evidence properties and store the contextual mappings in a global repository or provide access for local utilization.

Conclusion

Getting high-quality evidence from a large volume of diverse literature is an important task in clinical care. Automation to improve the evidence appraisal process is still required for clinical efficiency. We demonstrate the automation at the evidence appraisal stage by developing a supervised classification model called QRM and CAG mechanism. This approach assists medical practitioners and other stakeholders making evidence informed clinical decisions in clinical setups.

We plan to extend context-aware evidence grading task creating domain-specific data set in order to make the evaluation more consistent and precise to a particular domain. The approval of evidence by domain experts will be a step toward generating domain-specific training data having characteristics of relevance and high-quality acquired by the methods presented in this study.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Industrial Core Technology Development Program (10049079, Development of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

References

1. Dalrymple PW, Lehmann HP, Roderer NK, et al. Applying evidence in practice: a qualitative case study of the factors affecting residents' decisions. *Health Informatics J* 2010; 16: 177–188.
2. Evidence-Based Medicine Working G. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA* 1992; 268: 2420–2425.
3. Kilicoglu H, Demner-Fushman D, Rindflesch TC, et al. Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Inform Assoc* 2009; 16: 25–31.
4. Ely JW, Osheroff JA, Ebell MH, et al. Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999; 319: 358–361.
5. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, et al. Text categorization models for high-quality article retrieval in internal medicine. *J Am Med Inform Assoc* 2005; 12: 207–216.
6. West SL, King V, Carey TS, et al. *Systems to rate the strength of scientific evidence*. Agency for Healthcare Research and Quality, US Department of Health and Human Services, 2002, https://archive.ahrq.gov/clinic/epcsums/strengthsum.htm
7. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336: 924–926.
8. Group GW. Grading quality of evidence and strength of recommendations. *BMJ* 2004; 328: 1490.
9. Ebell MH, Siwek J, Weiss BD, et al. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *Am Fam Pract* 2004; 17: 59–67.
10. Cimino JJ. An integrated approach to computer-based decision support at the point of care. *Trans Am Clin Climatol Assoc* 2007; 118: 273–288.
11. Fowler SA, Yaeger LH, Yu F, et al. Electronic health record: integrating evidence-based information at the point of clinical decision making. *J Med Libr Assoc* 2014; 102: 52–55.
12. Perez-Rey D, Jimenez-Castellanos A, Garcia-Remesal M, et al. CDAPubMed: a browser extension to retrieve EHR-based biomedical literature. *BMC Med Inform Decis* 2012; 12: 29.
13. Haynes RB and Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ* 2004; 328: 1040.

14. Wilczynski NL, Morgan D and Haynes RB. An overview of the design and methods for retrieving high-quality studies for clinical care. *BMC Med Inform Decis* 2005; 5: 20.
15. Sarker A, Molla D and Paris C. Automatic evidence quality prediction to support evidence-based decision making. *Artif Intell Med* 2015; 64: 89–103.
16. Ruiz-Rico F, Vicedo J-L and Rubio-Sanchez M-C. Medline abstracts classification based on noun phrases extraction. In: Fred A, Filipe J and Gamboa H (eds) *Biomedical engineering systems and technologies*. Berlin: Springer, 2008, pp.507–519.
17. Kim SN, Martinez D, Cavedon L, et al. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics* 2011; 12: S5.
18. Kim S and Choi J. An SVM-based high-quality article classifier for systematic reviews. *J Biomed Inform* 2014; 47: 153–159.
19. Tang T, Hawking D, Sankaranarayana R, et al. Quality-oriented search for depression portals. In: Boughanem M, Berrut C, Mothe J, et al. (eds) *Advances in information retrieval*. Berlin: Springer, 2009, pp.637–644.
20. Dobrow MJ, Goel V and Upshur REG. Evidence-based health policy: context and utilisation. *Soc Sci Med* 2004; 58: 207–217.
21. Dobrow MJ, Goel V, Lemieux-Charles L, et al. The impact of context on evidence utilization: a framework for expert groups developing health policy recommendations. *Soc Sci Med* 2006; 63: 1811–1824.
22. Sackett DL and Rosenberg WMC. The need for evidence-based medicine. *J R Soc Med* 1995; 88: 620–624.
23. Sim I, Gorman P, Greenes RA, et al. Clinical decision support systems for the practice of evidence-based medicine. *J Am Med Inform Assoc* 2001; 8: 527–534.
24. Afzal M, Hussain M, Ali T, et al. Knowledge-based query construction using the CDSS knowledge base for efficient evidence retrieval. *Sensors* 2015; 15: 21294–21314.
25. Afzal M and Lee S. Relevant evidence acquisition and appraisal using knowledge-intensive queries. In: *Proceedings of the 2016 18th international conference on advanced communication technology (ICACT)*, Pyeongchang, South Korea, 31 January–3 February 2016, pp.703–709. New York: IEEE.
26. Sayers E and Wheeler D. Building customized data pipelines using the entrez programming utilities (eUtils), NCBI, 2004.
27. Ramos J. Using tf-idf to determine word relevance in document queries. In: *Proceedings of the first instructional conference on machine learning*, Piscataway, NJ, USA, 3–8 December 2003, pp.133–142. Piscataway, NJ: Rutgers University.
28. Hofmann M and Klinkenberg R. *RapidMiner: data mining use cases and business analytics applications*. Boca Raton, FL: CRC Press, 2013.
29. Porter MF. An algorithm for suffix stripping. *Program* 1980; 14: 130–137.
30. PubMed Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. [Table, Publication Types]. https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.publication_types/
31. Ho PM, Peterson PN and Masoudi FA. Evaluating the evidence is there a rigid hierarchy? *Circulation* 2008; 118: 1675–1684.
32. Evans D. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *J Clin Nurs* 2003; 12: 77–84.
33. Sebastiani F. Machine learning in automated text categorization. *ACM Comput Surv* 2002; 34: 1–47.
34. Yang Y. An evaluation of statistical approaches to text categorization. *Inform Retrieval* 1999; 1: 69–90.
35. Lewis DD. Naive (Bayes) at forty: the independence assumption in information retrieval. In: Nédellec C and Rouveirol C (eds) *Machine learning: ECML-98*, vol. 1398. Berlin, Heidelberg: Springer, 1998, pp.4–15.
36. Guo G, Wang H, Bell D, et al. KNN model-based approach in classification. In: Meersman R, Tari Z and Schmidt DC (eds) *On The move to meaningful internet systems 2003: CoopIS, DOA, and ODBASE*, vol. 2888. Berlin, Heidelberg: Springer, 2003, pp.986–996.

37. Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Nédellec C and Rouveirol C (eds) *Machine learning: ECML-98*, vol. 1398. Berlin, Heidelberg: Springer, 1998, pp.137–142.
38. Apte C, Damerau F and Weiss SM. Automated learning of decision rules for text categorization. *ACM T Inform Syst* 1994; 12: 233–251.
39. Ling CX, Huang J, Zhang H, et al. AUC: a statistically consistent and more discriminating measure than accuracy. In: *Proceedings of the IJCAI*, Acapulco, Mexico, 9–15 August 2003, pp.519–524. San Francisco, CA: Morgan Kaufmann Publishers Inc.
40. Huang J, Lu J and Ling CX. Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. In: *2003 ICDM 2003 third IEEE international conference on data mining*, Melbourne, FL, 22 November 2003, pp.553–556. New York: IEEE.
41. Verbert K, Manouselis N, Ochoa X, et al. Context-aware recommender systems for learning: a survey and future challenges. *IEEE Trans Learn Technol* 2012; 5: 318–335.
42. Rycroft-Malone J. The PARIHS framework a framework for guiding the implementation of evidence-based practice. *J Nurs Care Qual* 2004; 19: 297–304.