Thesis for the Degree of Doctor of Philosophy

AUTOMATIC EVIDENCE ACQUISITION AND APPRAISAL TO SUPPORT EVIDENCE-BASED MEDICAL DECISION MAKING

Muhammad Afzal

Department of Computer Science and Engineering
Graduate School
Kyung Hee University
South Korea

Feb 2017

AUTOMATIC EVIDENCE ACQUISITION AND APPRAISAL TO SUPPORT EVIDENCE-BASED MEDICAL DECISION MAKING

Muhammad Afzal

Department of Computer Science and Engineering
Graduate School
Kyung Hee University
South Korea

Feb 2017

AUTOMATIC EVIDENCE ACQUISITION AND APPRAISAL TO SUPPORT EVIDENCE-BASED MEDICAL DECISION MAKING

by

Muhammad Afzal

Supervised by

Prof. Sungyoung Lee

Submitted to the Department of Computer Science and Engineering and the Faculty of Graduate School of Kyung Hee University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Dissertation Committee:

Prof. Tae-Seong Kim Kim, TaeSeong

Prof. Soo-Yong Shin Shin

Prof. TaeChoong Chung

Prof. Hyon Woo Seung Seung

Prof. Sungyoung Lee

Dedicated to my

Family and Teachers

During my educational career, my family and teachers play a pivotal role to keep me encouraged to pursue higher studies. Without their moral and financial support at different stages, it was not easy to step into a graduation level study. Particularly, my father who keep supporting to continue my study while he was not in the position to fulfill the financial requirements of my study. Although, he only saw my career accomplishment till the mid of my PhD study, however, it was his dream that ultimately come true in 2016-17. My mother at the age of 60s needs my support to be with her, but oppositely, she provides all the required support to keep me in the position to complete my PhD degree. Moreover, all my brothers and sisters are very kind and supportive during the entire period of my study with their sincere encouragement and soulful prayers. My wife particularly provided full support and accompanied me in Korea while taking care of me and my kids.

This thesis is dedicated to my family and teachers. I will always remember the affection, enthusiasm, and encouragement that they entrusted to me that gave me strength to face all the difficulties and challenges during my course of research.

Abstract

An overwhelming amount of biomedical information is available in the form of text documents which can be retrieved intelligently to support the process of evidence-based decision making. However, there exists a large gap between the user space and evidentiary document space reaching a better clinical decision. Successful clinical practice demands best available evidence to find from the research literature for a better clinical action. However, clinicians face many barriers in order to access best available evidence. The main barriers are the lack of automated and reliable methods to finding and recognizing the best evidence among a huge array of evidentiary documents. Best evidence is the one which is relevant, is of high quality, and fits contextually.

Current approaches are not sufficient to cover all the three, aforementioned, aspects of the best evidence. The existing approaches of query construction are largely information driven, non-structured, and manual. Similarly, for the quality assessment, Boolean approaches are largely in practice. Even though modern approaches have shifted focus towards statistical approaches for quality assessment, however, dataset availability and reliability in addition to feature engineering are among the major challenges. Determining quality is not sufficient to establish an evidence is contextually fit for a user. Currently, user context is taken to the level of user goal and other aspects such as varied user role (physician, nurse, researcher) and environment (clinical setup, research unit) are majorly overlooked.

The main goal of this thesis is to minimize human efforts getting best research evidence for better clinical decision making. This goal is achieved through satisfying the objectives of: to develop and evaluate methods/models for finding relevant evidentiary documents, to develop and evaluate methods/models for recognizing quality evidences, and to develop and evaluate methods/models for evidence contextual fitness. To achieve these objectives, this thesis proposed a three-fold methodology for acquiring a best available evidence: relevant evidence acquisition, quality assessment, and contextual fitness determination.

For relevant evidence acquisition, an automatic method called *Task Aware PICO (Problem, Intervention, Comparison, and Outcome)-compliant question construction* is proposed, which has two sub-parts: PICO compliancy and task awareness. PICO-compliant question construction involves knowledge of a clinical decision support system as a source of query contents. A mapping model called KAP (knowledge alignments to PICO) is constructed for correct mapping of different parts of a knowledge representation scheme to appropriate part of PICO. The mappings are achieved at two levels: structure level and concept level. For structure level mappings, a set of specialized models are proposed in order to cover the diversified knowledge representations such as Arden syntax medical logic module (MLM), production rules, and others. For concept level mappings, an algorithm called STI (salient term identification) is developed that identifies important terms for the final query on the basis of term matching using standard terminology system. PICO-compliant question is considered as the initial query, which is augmented with a clinical task information making it more concentrated on a specific user goal such as treatment, diagnosis, etiology, and prognosis. For clinical task awareness, the concepts belonging to the parts ”I” (Intervention) and ”C” (Comparison) are utilized to recognize the top level semantic category of a particular hierarchy in SNOMED CT. For quality assessment, a statistical-based quality assessment is proposed which is based on a classification model called quality recognition model (QRM). QRM is a support vector machine (SVM) based binary classification model which is trained on a dataset annotated by a team of professional experts. QRM utilizes two types of features: data features (title and abstract) and metadata features (medical subject heading (MeSH) terms and publication type). All of these features are engineered automatically by involving text processing functions of tokenization, stop words removal, case changing, stemming, and token filtration. For contextual fitness, this work proposed a method called context-aware evidence grading, which aggregates the user context with evidence context. The aggregation of user and evidence contexts are derived from the contextual matrices designed for user and evidence. The contextual matrices are initially constructed through two mechanisms: literature-based context acquisition and expert-driven context determination. Final grading of evidence is instantiated at three levels: high, medium, and low according to its fitness to the context.

The proposed methodology presented in this thesis is evaluated at different levels by performing multiple experiments on different evaluation criteria. First of all, the correctness of automatically constructed PICO-compliant question is evaluated with four types of measurements: precision at ten retrieved documents (P10), mean precision (MP), total document reciprocal rank (TDRR), and mean reciprocal rank (MRR). Secondly, the QRM model performance is evaluated on a set of expert annotated evidentiary documents using 10-fold cross-validation technique. Thirdly, the evidence contextual fitness is duly verified from the physicians. Finally, the results obtained from these evaluations showed significant improvements in terms of accuracy and time efficiency. Moreover, this work has been realized for head and neck cancer treatment domain where its importance has been recognized by the physicians involved in evidence-based clinical practice.

Acknowledgement

First and foremost, I render my humble and sincere thanks to the ***Almighty Allah*** for showering HIS blessings upon me. The Almighty gave me the strength, courage, and patience during my doctoral study.

Special thanks to my advisor Prof. Sungyoung Lee who provided me guidance, strength, support, and courage in overcoming the difficult challenges throughout my time as his student. I have learned a lot from him in becoming a productive person in diverse situations. He inspired me with his dynamic personality and his unreserved help and guidance lead me to finish my thesis. He has great role in polishing my skills such as thinking, creativity, and technical soundness which are key ingredients for high quality research. Additionally, I would like to acknowledge the valuable guidance and support of Prof. Byeong Ho Kang from University of Tasmania, Australia, who closely worked with me and refined the problem statement as well as streamlined the research direction.

I am grateful to Prof. R. Brian Haynes (McMaster University, Canada) whom I collaborated to collect data for statistical-based quality model. Moreover, he reviewed my research work and provided valuable comments to improve the overall quality of work.

I am extremely grateful to some of my colleagues who have always provided me time, expertise, and encouragement in my course of research. They were always present to guide me in difficult situations of my PhD duration. I would like to thank Dr. Maqbool Hussain, Dr. Wajahat Ali Khan, Dr. Rahman Ali, Taqdir Ali, Jamil Hussain, Syed Imran Ali, Shujaat Hussain, Dr. Asad Masood Khattak, Dr. Zeeshan Pervaiz, Dr. Muhammad Fahim, Dr. Muhammad Bilal Amin, and Dr. Oresti Baos (University of Twente). They have contributed enormously in successfully performing various academic and personal tasks that confronted me during my stay at South

Korea.

I am very thankful to all of my current and former Ubiquitous computing lab fellows and colleagues for their kind support to my personal and academic life at Kyung Hee University. I am highly obliged to brilliant researchers - Tae Ho Hur, Sung Hoo, Jae Hun Bang, Bilal Ali, Kifayat Ullah, Dr.Waqas Nawaz, Omar Farooq, Dildar Hussain, Mahmood Ahmad, Saeed Ullah, Abdul Sattar, Ahsan Raza Kazmi, Zain Abbas, Maqbool Ali, Muhammad Asif Razaq, Usman Akhtar, Shehryar, Muhammad Telha, Ammad ud Din, Dr. Hanif Sidiqqi, Muhammad Jalil Piran, Waseem ul Hassan, Ahsan Raza, Aftab Alam, Musarrat Hussain, Muhammad Sadiq, and Anees ul Hassan. This journey would not have been possible without their support. They contributed to my personal and academic life to polish myself. Also, I appreciate all my Korean and international friends who worked as a team with me and developed my team work skills and provided me wonderful memories during my stay in South Korea.

I would like to thank Mrs. Kim for being available and helping me in many regards. Even though she has always been extremely busy, but always been kind when asked for help.

I am also very thankful to all collaborators from Shuakat Khanum Memorial Cancer Hospital (SKMCH) Lahore Pakistan for their enthusiasm and efforts in developing clinical decision support system for head and neck cancer and shared with us the clinical domain knowledge. I am especially thankful to Dr. Arif Jamshid, Dr. Hassan Iqbal, Dr. Irfan, Dr. Sadaf, Mr. Manzar Abbas, Mr. Idress Kokhar, and other HMIS team members for their kind support. Moreover, I am thankful to MS advisor and co-advisors in SEECS, NUST, Islamabad especially Dr. Arshad Ali, Dr. Hafiz Farooq Ahmad, Dr. Khalid Latif, and Mr. Majjid Maqbool for their guidance and continuous support during my study in Korea.

I would like to extend my sincere thanks to my kind and respected teachers at primary and high school level, who support and encourage me to pursue my higher study. To mention few in the list, I am highly thankful to Mr. Jamil Hussain Kazmi, Mr. Zafar Hussain, Mr. Asghar Hussain, Mr. Sajjad Hussain, Mr. Iftikhar Hussain, Mr. Ashiq Hussain, Mr. Wazir Ali, Mr. Ali Muhammad, and Mr. Nowroz Ali.

Last but not the least, I would like to express my sincere gratitude to my close friends, colleaques, and relatives especially Mr. Habib Hussain Fani, Mr. Naib Hussain, Dr. Zahid Hassan, Mr. Ashiq Hussain, Mr. Muhammad Ejaz, Mr. Ajmal Hussain, Mr. Mumtaz Ali, and Mr. Najm ul Hassan), my father-in law, brothers-in law and other relatives who support me and my family morally and financially during long journey of my study.

Muhammad Afzal
Feb, 2017

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Overview

Evidence-based systems have long been used in the clinical domain as a support for clinicians to make better clinical decisions. Individual clinical expertise and best available external evidence complement each other and doctors are encouraged to use them in combination [3]. Without clinical expertise, even an excellent external evidence report may be inapplicable to, or inappropriate for, an individual patient. Similarly, without current best evidence, clinical expertise alone cannot provide clear justification for a clinical decision. As highlighted in [4], health systems fail to optimally use evidence, which results in deficiencies that adversely affect the patients quality of life.

In the medical literature, there is a huge amount of credible medical resources which suffice to provide new research and knowledge in the domain. However, there is a large gap between actual practice and what evidence shows can be done [5] [6]. It is not surprising that most clinicians consider the research literature to be unmanageable [3] and of limited applicability to their own clinical practices. Finding relevant and high quality clinical evidence is essential to successful practice [2] but, clinicians face many barriers in order to access the relevant information which is suitable to the context. If done at all, most of the time, seeking best evidence is done manually [7]. Clinicians may spend a lot of time searching for required information using conventional search engines such as PubMed, Google, and others. Moreover, when they find the relevant information, they have no established way to manage the results conveniently for efficient use in future scenarios. The big challenge is the design of searching query from the user context in an automatic and intelligent manner to save clinicians time spent on query construction. The low quality of documents from where the evidence is fetched for the decisions adds further to the challenge of an automated acquisition of evidence.

To resolve these challenges, a comprehensive methodology is designed that is consisted of a set of methods for biomedical documents' acquisition and appraisal in order to find relevant and high-quality research evidence. More specifically, the contributions are made in three parts.

- Evidence acquisition through knowledge-based query formulation and augmentation. There are three important steps are performed in evidence acquisition: i) automated formulation of the primary queries from the rules of a clinical decision support system (CDSS), ii) automatic identification of clinical task and translation of the primary query making it augmented query, iii) transformation of augmented queries into searchable URLs.

- Evidence appraisal through quality prediction model so-called quality recognition model (QRM). In this method, three fundamental steps are carried out: i) automated feature engineering for corpus preparation, ii) preprocessing of features for improved accuracy and efficiency, and iii) machine learning algorithm selection and parameter setting.

- Contextual fitness through context-aware evidence grading, which aggregates user context with evidence context. The aggregation of user and evidence contexts are derived from the contextual matrices designed for user and evidence. The contextual matrices are initially constructed through two mechanisms: literature-based context acquisition and expert-driven context determination. Final grading of evidence is instantiated at three levels: high, medium, and low according to its fitness to the context.

With proposed methodology, clinicians have a better chance to get support in evidence informed clinical decisions, keeping themselves up-to-date with newly published research in the domain, and minimizing the time spent on manual methods of searching and finding the best available evidence from the literature.

1.1 Motivation

For evidentiary support, clinicians mostly rely on the publicly available searching services such as PubMed, Google, UpToDate, and others. These searching services are worthy, but need to be integrated with a user system to make the evidence retrieval manageable. Also it is required to evaluate the retrieved evidence for quality rather to rely on a search engines in-built evaluation system.

Some of todays health systems are equipped with the knowledge base (KB) of a clinical decision support system (CDSS). The KB of the CDSS provides additional support to help in automation of the process of evidence support. It helps in query construction automation, since it has knowledge rules that consist of patient information with established logical connections. An additional opportunity we get from the existence of a CDSS in a health system is the purpose or query type information. The query type information shows the purpose for which a CDSS is developed, such as, treatment plan or diagnosis recommendations, and others. Adding purpose information to the user query has a great impact on the evidence retrieval process which we will show in subsequent sections; Methods and Results.

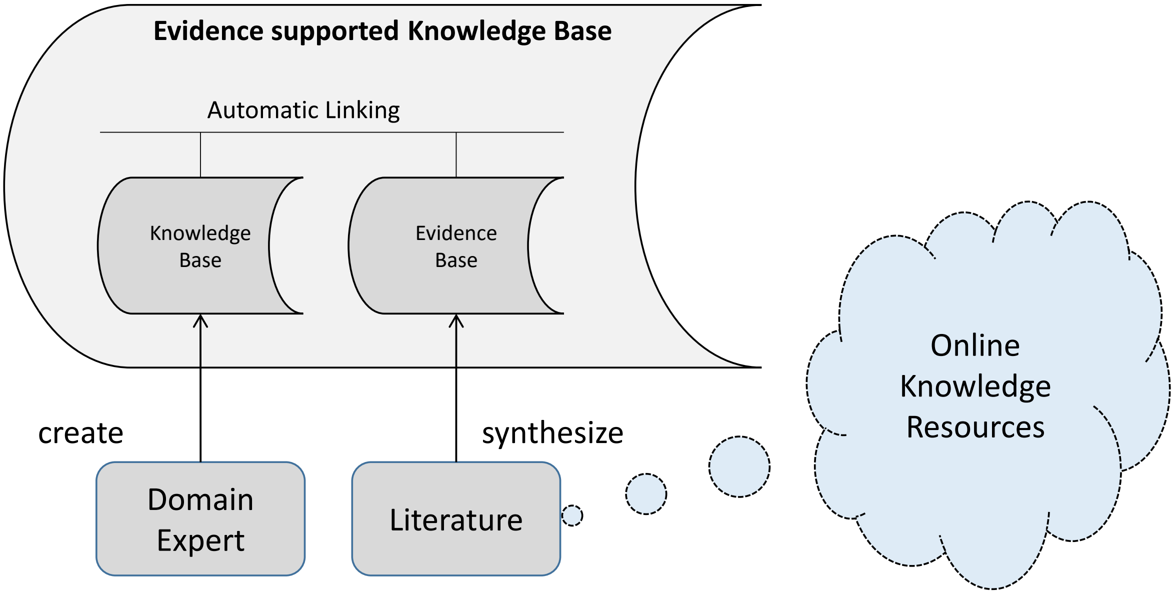


Figure 1.1: Motivations for evidence supported clinical decision making

Putting a health system, a CDSS, and an evidence support system together, a conceptual model is designed as shown in Figure 1.1, which describes the linkage among a health system, KB of CDSS, and evidence base retrieved from external resources. The health system manages the patient records to be used by the clinician and knowledge base of a CDSS is created with the support of expert clinicians either through directly authored rules or machine learning approaches. The evidence base shows the appraised evidence synthesized from literature through automatic methods of acquisition and appraisal. In this thesis, the scope is limited to synthesizing evidence from online resources for facilitating clinicians in the clinical decision making process. The proposed methodology utilizes information contents from health system and CDSS knowledge base in order to construct the query for searching and retrieving the best available evidence.

1.2 Problem Statement

Acquiring evidence which is relevant, methodologically rigorous, and contextually fit from a sheer amount of literature is a complex problem. Without a well formulated question and a content quality recognition, it is highly time consuming for a clinician to identify relevant and quality evidence.

This problem is highlighted in the existing literature on more than one occasion but still it leaves many research challenges to be focused. To resolve the time consuming issue of finding relevant and evidence from a sheer amount of literature, the existing work can be categorized into two parts: pre-retrieval and post-retrieval efficiency of the research evidences. Evidence pre-retrieval efficiency address the challenges in query part i.e. how to improve the query contents to get relevant results, while the post-retrieval efficiency consider to check the document contents more rigorously in order to identify the quality of documents.

In evidence-based medicine, what matters the most is the reliability of the evidence and relevancy to the user question. Relying only on keywords matching techniques cannot serve the purpose of evaluating the quality of the evidence. Thus, it is required to construct a well-built query in order to successfully retrieve only the relevant set of evidentiary documents and also, an approach to assess the quality of the contents beyond matching of query elements with documents contents. On the contrary, there is lack of studies that focus both on pre- as well as post-retrieval efficiency in order to assure the relevancy as well as quality of the evidence.

The main goal of this thesis is develop a comprehensive methodology to support automatic approaches for constructing a well-built question and quality assessments to minimize human efforts getting best research evidence for better clinical decision making. Reaching this goal, we have to achieve two objectives:

- To develop and evaluate methods/models for finding relevant evidentiary documents.
- To develop and evaluate methods/models for recognizing quality evidences.

To achieve the aforementioned goal and objectives, the candidate challenges for this research work are as follows:

- How to develop a well-built question automatically and what structure to use for a well-built question?
- How to identify the clinical task for a well-built question in order to segregate user required documents from unwanted set of documents?
- How to assess the quality of contents in the documents with what method?
- How to rank the quality assessed documents making them contextually applicable for the user who makes the request?

To address these challenges, this work provides a solution consisting of two parts;

- Automatic Evidence Acquisition: A PICO-compliant well-built question is constructed from the knowledge elements. The PICO-compliant question is augmented with a clinical task determined automatically through the intervention concepts of PICO (Chapter 4).
- Automatic Evidence Appraisal: A statistical model called quality recognition model (QRM) is designed and trained on a set of documents annotated by a team of expert professionals in the domain for the purpose to assess the quality of the contents and for appropriate ranking, the quality-assessed documents are graded on the basis of user context (Chapter 5).

1.3 Contributions

The main contribution of this research work is divided mainly into two areas: automatic construction of PICO-compliant question and recognition of quality evidences.

Construction of PICO-compliant question

PICO is a well-established template to frame a well-built question consisting of P (problem/patient/population), I (intervention), C (comparison), and O (outcome). Identifying and extracting PICO parts from a knowledge is a big hurdle. This work propose a mapping model called KAP (knowledge alignments to PICO) which maps appropriate part of a knowledge to appropriate part of PICO. On the top of KAP, different implementation models can be developed. This thesis discusses two such models: MAP (MLM alignments to PICO) and PRAP (production rules alignments to PICO). MLM is a short form of medical logic module that appears as a leading standard knowledge representation for CDSS (clinical decision support system).

Recognition of clinical task

A searching query will retrieve a huge set of evidentiary documents if executed in the as-is form without constraining to focus more towards user goal. Usually, the search services such as PubMed provides the facility of constraining a user query by providing filter support in the form of ”advanced search” facility. In a human made query it is quite possible that a user take advantage of using filter to get more target oriented results. It, however, becomes a great challenge to find out the user task correctly. To recognize the user task correctly, this work has utilized concepts belong to two of the parts of PICO i.e. I and C as input and identify the top-level semantic category of particular hierarchies in SNOMED CT. The identified semantic category is rightly translated to appropriate terminologies which are recognized in the target search engine (i.e. PubMed).

Recognition of quality evidence

Critical appraisal is required to filter out the studies of less quality. There are two possibilities to achieve the appraisal objective: manually through the domain experts or automatically through the system. Involving domain experts in the first phase of the appraisal is highly time consuming, which we avoided in this study through machine learning approaches. For machine learning models, the great challenge is related to the selection of training data and automated preparation of features. In order to resolve the first challenge, i.e. selection of training data, a collection of MEDLINE documents is acquired which is created by highly qualified specialists for the purpose of finding high-quality articles. For the second challenge, i.e. automated preparation of features, the data as well as metadata features for offline as well as online experiments are engineered automatically.

Contextual grading of evidence

Assessment on the basis of relevancy and methodologically rigor is not sufficient to establish whether evidence contextually fits to serve the user appropriately or not. Conventionally, the contextual factors were evidence focused and very less attention is provided to the user contextual factors. This thesis focuses to involve both user and evidence contextual factors while evaluating the contextual fitness of an evidence. Prior to aggregation of contexts, contextual matrices are acquired on the basis of literature-driven and expert-driven context acquisition approaches.

1.4 Thesis Organization

This dissertation is organized into chapters as following.

- **Chapter 1: Introduction**. Chapter 1 provides a brief introduction of the research work on evidence acquisition and appraisal. It highlights the motivation for evidence-based medical decision making. Moreover, it discusses the problem in brief with research challenges and goals and objectives.

- **Chapter 2: Background and Related Work**. A background detail is provided in this chapter about the evidence-based medicine including evidence acquisition and appraisal. Furthermore, the different approaches in the relevant areas are described to explore the potential work in the domain. In addition, it also discusses the preliminaries in the research domain.

- **Chapter 3: Proposed Methodology**. This chapter provides the abstraction of proposed solution in the form of functional flow. It connects the elements of solution with respect to their input and output. It provides the holistic view of the overall proposed methodology and provides a high level description of major function of each solution.

- **Chapter 4: Automatic Evidence Acquisition**. This chapter describes the details of solution 1, i.e., automatic evidence acquisition that is composed of two parts: PICO-compliant question construction and clinical task aware query formulation.
- **Chapter 5: Automatic Evidence Appraisal**. In this chapter, the evidence appraisal methods are demonstrated. It includes the quality assessments on the basis of a statistical-based model called quality recognition model (QRM) and grading of the quality evidences on the basis of user context.
- **Chapter 6: Results and Evaluation**. In this chapter results of the proposed solution are presented and evaluated. Finally, the proposed solution is compared with existing approaches in the domain.
- **Chapter 7: Conclusion and Future Directions**. This chapter concludes the thesis and also provides future directions in this research area.
- **Appendix A: System Implementation**. This is an auxiliary chapter composed to describe the design and implementation of a system developed using the proposed methods in order to highlight the practical implication of the proposed research.

Chapter 2

Related Work

2.1 Background and Related Work

Evidence-based Practice (EBP) [8–11] and CDSS [12–14] have long been used in the clinical domain to enhance clinical efficacy. EBP and CDSS share clinical expertise as a source of data. EBP utilizes clinical expertise along with research evidence and other factors for a clinical decision. A CDSS knowledge base (KB) is the representation of clinical expertise of one or more clinical experts. EBP is defined as the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients integrated with clinical expertise and patient values to optimize outcomes and quality of life [15]. Three-circle model in [16] explains three sources of data for evidence-based clinical decisions: clinical expertise, research evidence, and patients preferences. This model lacks explicit guidance in how the sources of data are to be integrated when making decisions. Also, the scope, relative value, and appropriate applications of clinical expertise remain unclear. The same model is revised in [17] by adding a transdisciplinary perspective. It incorporates an ecological framework and emphasizes shared decision making. The model adds a new external frame of environment and organizational factors to create a cultural context that moderates the acceptability of an intervention, its feasibility, and the balance between fidelity and adaptation that is needed for effective implementation. In [15], the process of systematically assessing the need for information during routine clinical care is summarized as the five As of EBP - assess, ask, acquire, appraise and apply. This 5A cycle is also called an evidence-based decision-making cycle and has inspired others to implement for their purposes. The Evidence Based Library and Information Practice (EBLIP) is one of such examples which provides a sequential, structured process for integrating the best available evidence into making important decisions [18].

For any evidence-based system to efficiently work in a domain, the context of that domain plays a critical role. Context provides the features for query generation in order to seek relevant information. The source and format of data are crucial to consider for automatic or semi-automatic query generation. Cimino JJ. presented the idea of Infobuttons [19] and Infobutton Manager (IM), which attempt to determine the information needs based on the user context. Infobuttons are mainly topic specific with a question facility for the users to tune the query more towards the context. The main focus of the Infobuttons approach is to establish context-specific links to health information resources. It is based on simple topic based linkage to the resource from within the context of EMR/EHR and is not suitable to develop complex queries. Fowler et al [20] were successful in creating and integrating a diagnostic decision support tool (DDST) and connecting it to the differential diagnostic tool. The authors described that the DDST lifts terms from standard, coded fields in the EHR and sends them to the diagnosis tool Isabel, which produces a list of possible diagnoses. CDAPubMed [21] is a browser extension aiming to provide a tool to semi-automatically build complex queries. It provides additional information to the contents of EHR for improving and personalizing biomedical literature searches. HL7 Clinical Document Architecture (CDA) is used as a main source to extract clinical terms for query generation. It loads the patient clinical documents (CDA documents), identifies relevant terms for scientific search, and generates and launches literature search queries to a major search engine, i.e. PubMed, to retrieve citations related to the EHR under examination.

Similarly, finding high-quality articles in MEDLINE, Wilczynski et al. [22] developed clinical query filters which was later adapted by PubMed for use as Clinical Queries (CQ). The data collection used in CQ filters is annotated across four dimensions: format, human health care, purpose, and scientific rigor. Aphinyannaphongs et al. [7] performed experiments to compare the results of clinical query filters and machine learning methods. The authors demonstrated that the machine learning approach applied to categorizing high-quality articles in internal medicine performed better than the Boolean methods. Kilicoglu, Halil, et al. [2] used the same collection that was manually created to develop the CQ filters for PubMed with semantic features to find high-quality evidence. The authors confirm that statistical approaches using semantic features performed better than merely used Boolean methods. The work in [2, 7] provided the baseline to consider statistical methods for evidence appraisal in this study rather to focus only on Boolean methods. In a very recent article on evidence quality prediction [1], the authors addressed the problem of automatic grading of evidences on a chosen discrete scale. The authors experimented many features such as publications year, avenue, and type to evaluate the quality of an evidence. They found that publication type is the most eminent feature to consider for evaluation the evidence quality results.

The existing approaches discussed in the above paragraphs can be summarized to focus on automation of evidence processing to overcome the most important problem of time saving for the clinicians while practicing evidence-based medicine. The inclusion of research evidence in clinical decisions varies with respect to domain context and the purpose for which the evidentiary support is required. Conceptually, the evidence adaption follows the same 5As cycle as mentioned in [15], however, implementation makes the scenario different. A user in a clinical setup with CDSS implementation needs to approach the evidence differently than a user having no CDSS implementation. The feature selection and automation at different levels brings uniqueness to the approach and poses challenges at the same time. For instance, in [1] the authors did not consider Medical Subject Heading (MeSH) terms as a feature which is of high importance according to [2,7]. Our proposed approach takes into account the challenge of evidence adaption by supporting the automation at two levels: evidence acquisition and evidence appraisal. The automation in evidence acquisition is supported by automated query formulation and augmentation with Boolean methods, while the automation in evidence quality appraisal is supported by statistical methods applied on data as well as metadata features.

Finding high quality evidence is essential for successful practice [2], but medical practitioners face many barriers in using evidence-based answers at point-of-care [23]. If done at all, most of the time, seeking best evidence is done manually [7]. 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 [24]. 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) [25, 26] and Strength Of Recommendation Taxonomy (SORT) [27] 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 [19–21] 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 [22, 28], but the statistical approaches proposed in [2, 7] presented a proof of better accuracy in recognizing quality articles as compared to Boolean approaches. Very recently, Abeed Sarker, et al. [1] presented an approach of evidence quality prediction through supervised classification model. The approach uses the strength of recommendation taxonomy (SORT) [27] to grade the evidences. A number of other approaches [29–31] are proposed in the area of text classification. Authors of 16 combine the existing techniques innovatively for the classification of Medline abstracts based on a noun phrase extraction. Kim et al. [30] provides 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 [31]. Domain specific post-retrieval re-ranking approach [32] is proposed in the domain of depression that attempts to re-rank the articles returned by the search engine.

2.2 Preliminaries

This section presents a brief overview of evidence-based medicine (EBM) and current advancements in EBM, clinical decision support system (CDSS), evidence-adaptive CDSS, Boolean and statistical approaches for evidentiary documents' quality assessment, and different machine learning methods used for statistical evaluations. It provides an overview of these approaches to make a solid ground for relevant and quality evidence acquisition and appraisal.

2.2.1 Evidence-based medicine

Evidence-based medicine (EBM) was introduced as a new paradigm for the practice of clinical medicine [33] and was intended to develop and promote an explicit and rational process for clinical decision making. After important critical exchanges within the medical community, EBM was more explicitly defined as the conscientious and judicious use of current best evidence from clinical care research in the management of individual patients [16].

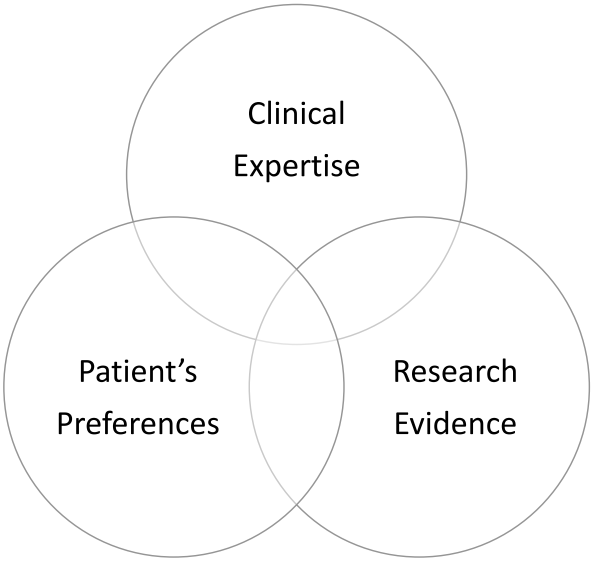


Figure 2.1: Three-Circle Model of Evidence-Based Clinical Decisions

As shown in 2.1, the three circles illustrate the distinct but overlapping sources of data that might be used when making clinical decisions. Moreover, the authors explicitly stated that under certain circumstances, clinical expertise and/or the patient's preferences may override research evidence. Note that the three circles are of equal size or weight, with clinical expertise occupying the top, central position. The authors also were careful to state (and restate) that EBM is not cook book medicine, a means of cutting costs by limiting care, or a subversive means for clinical researchers to overemphasize the value of randomized-controlled trials

Although conceptually appealing, the original model lacked explicit guidance in how the circles or sources of data were to be integrated when making decisions, particularly when the research evidence was at odds with either clinical experience or the patients preferences [17]. Furthermore, the scope, relative value, and appropriate applications of clinical expertise remained unclear. An updated model then attempted to address these concerns by changing the clinical expertise circle to clinical state and circumstances and moving clinical expertise to the intersection points of the new three circles, as shown in 2.2.

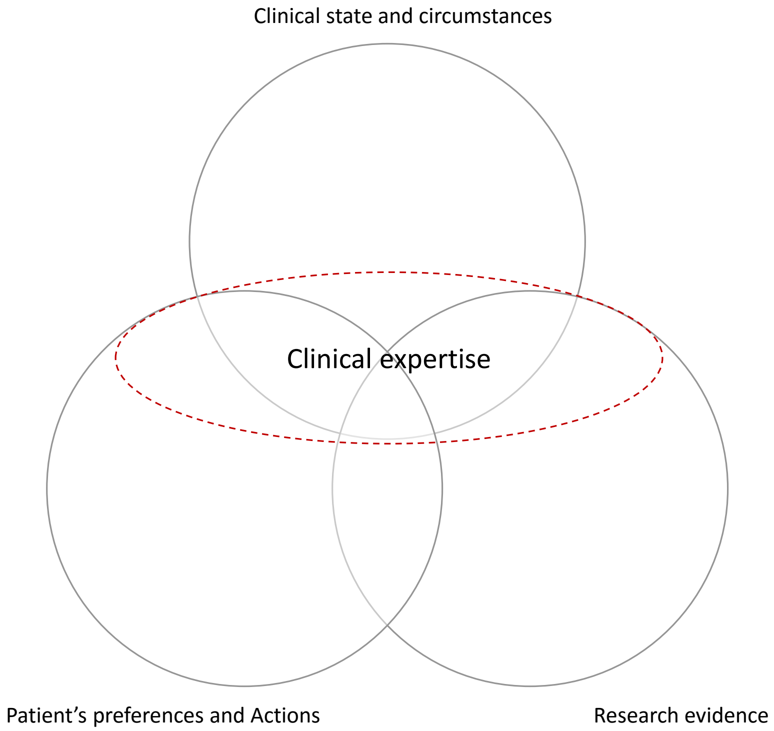

Clinical state and circumstances
Clinical expertise
Patient’s preferences and Actions
Research evidence

Figure 2.2: An Updated Three-Circle Model of Evidence-Based Clinical Decisions

The central placement of clinical expertise highlights the value of clinical experience in guiding the EBM decision-making process and offers a noteworthy concession regarding the importance of the individual practitioner.

The updated three circle model is criticized for uncovering the role and value of practitioners and their expertise are unclear; resources and/or contextual factors are ignored; and not enough attention is paid to clients’ preferences. This criticism led the model to be defined more broadly. A new transdisciplinary evidence-based practice (EBP) model is proposed as shown in 2.3, that incorporates each discipline’s most important advances and attempts to address remaining deficiencies.

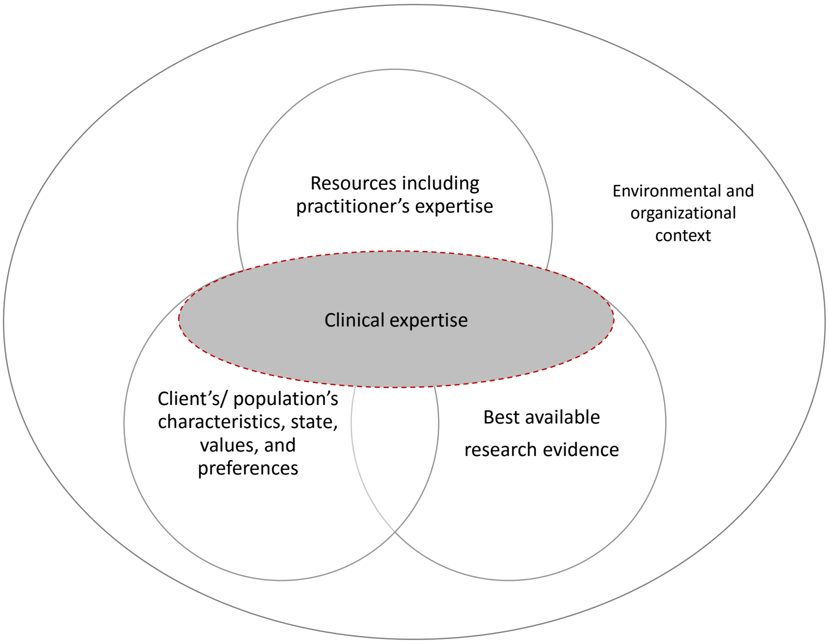

Resources including practitioner's expertise
Environmental and organizational context
Clinical expertise
Client's/ population's characteristics, state, values, and preferences
Best available research evidence

Figure 2.3: Revised EBP Transdisciplinary Model:

The most sophisticated practice of EBM requires, in turn, a clear delineation of relevant clinical questions, a thorough search of the literature relating to the questions, a critical appraisal of available evidence and its applicability to the clinical situation, and a balanced application of the conclusions to the clinical problem. The evidence cycle incorporates the facet of EBM including asking the question, acquiring the information, appraising its quality, applying the results, and ultimately acting on the patient [34]. More formally, this cycle is named as 5 A's evidence cycle.

- Assess
- Ask
- Acquire
- Appraise
- Apply

Inspired from the 5 A's evidence cycle, Evidence Based Library and Information Practice *EBLIP* [18] achieved attention. EBLIP provides a sequential, structured process for integrating the best available evidence into making important decisions. The practitioner applies this decision making process by using the best available evidence while informed by a pragmatic perspective developed from working in the field, critical thinking skills, and an awareness of different research designs, which is further modulated by knowledge of the affected user population's values or preferences. The EBLIP process provides structure for reaching important decisions. The EBLIP process resembles evidence based processes in other professions such as education, health, management, or public policy analysis. The steps in the EBLIP process can be summarized as follow.

- Formulate an answerable question on an important issue
- Search for the best available evidence to answer the question
- Critically appraise the evidence
- Make a decision and apply it
- Evaluate one's performance

2.2.2 Knowledge Representation Languages

Clinical decision support system (CDSS) is a computer system designed to support clinician decision making about individual patients at the point of care where these decisions are made [13,35]. A CDSS is consisted of multiple components where some of them are mandatory such as knowledge base and inference engine. The knowledge base of a CDSS can be represented in a variety of formats depending on the domain and organizational context and priorities. For the sake of this thesis scope, the onward description is focused on different schemes of representation currently practices as industry standards.

Production Rule

A production rule is a statement of programming logic that specifies the execution of one or more actions in the case that its conditions are satisfied. Production rules therefore have an operational semantic (formalizing state changes, e.g., on the basis of a transition system formalism) [36]. The Production Rule Representation (PRR) is a proposed standard of the Object Management Group

(OMG) to provide a vendor-neutral rule-model representation in UML for production rules as used in forward-chaining rule engines. More formally, a production rule is a two part structure where the first part is referred as antecedent part and later is called the consequence part [37]. The antecedent part is comprised of conditions which if true triggers the consequence part for taking some action. The two-part structure is formally written as follow.

IF conditions THEN actions

If there is more than one condition in the antecedent part, they are connected together. Each condition may be positive or negative, however, action is taken when all of them are understood to be true.

Arden Syntax Medical Logic Module

HL7 have provided Arden Syntax formalism for representing clinical knowledge in order to facilitate the sharing of computerized health knowledge. Knowledge bases encoded in Arden Syntax are represented as a set of discrete and independent module called Medical Logic Modules (MLMs) [38]. HL7 Arden Syntax intent is to specify knowledge representation that is sharable while the contents should be readable by both human and machines [39]. Each MLM is containing three categories (maintenance, library, and knowledge) and each category is made of slots. The maintenance and library categories describe the meta-information on the knowledge such as title, keywords, author, explanation, version) and the knowledge category provides the logic of an MLM [38].

GuideLine Interchange Format

The GuideLine Interchange Format (GLIF) is a combination of two aspects: GLIF model and GLIF syntax. The GLIF model consists of a set of classes represented in an object-oriented representation. The GLIF syntax provides the specification of the test file format that contains the encoding [40]. GLIF main intention is to represent different types of guidelines for screening, diagnosis, and treatment, at the time of primary or specialty care, and in acute or chronic problems [41].

2.2.3 Medical Terminology Standards

Terminology standards are broadly utilized for encoding the clinical concepts in standard codes for the purpose of shared and non-conflicted meaning. A number of terminology standards are currently in use to serve the said purpose. To the scope of this thesis, three standards i.e. SNOMED CT, UMLS, and MeSH are presented.

SNOMED CT

SNOMED CT [42, 43] stands for Systematized Nomenclature of Medicine – Clinical Terms. It is is a multilingual standardized vocabulary of clinical terminology that is used for the electronic exchange of clinical health information. SNOMED CT is managed and maintained by the International Health Terminology Standards Development Organization (IHTSDO). It is conducive to improve the patient care by enforcing the development of Electronic Health Records of clinical information for contextual retrieval. The analytical and decision support system get all essential and effective information through SNOMED CT for patients benefit in terms of standardize communication and quality of care. SNOMED CT is a set of standardized concepts arranged in a hierarchical structure. The whole ontology is organized through these hierarchies with the top level concepts which represent broad semantic types. The top level concepts are of much importance to recognize the semantics of different concepts. Some of these top level concepts include ”clinical findings,” ”procedure,” ”body structure,” and ”observable entities.”

Unified Medical Language System

The Unified Medical Language System (UMLS) [44–46] is a repository of biomedical vocabularies having over 2 million names for some 900, 000 concepts, and 12 million relations among these concepts [44]. UMLS is developed by the US National Library of Medicine in 1986. It is pertinent to highlight that there are three major components (Semantic Network, metathesaurus, and lexical resources) of the UMLS that are distributed annually to interested researchers. The Semantic Network provides high level categories used to categorize the metathesaurus concepts which is a repository of inter-related biomedical concepts, and lexical resources include the SPECIALIST lexicon and programs for obtaining the lexical variants of different biomedical terms.

Medical Subject Headings

Medical Subject Headings (MeSH) [47–49] is a controlled vocabulary used for indexing and searching of journal articles in MedLine database, books, and other printed and non-printed materials and catalogs. MeSH terms constitute a thesaurus and are arranged in a hierarchical, tree-like structure by subject categories. Currently, there are over 13,000 main MeSH terms and a list of 80 possible concept terms, e.g., diagnosis, drug therapy, surgery, etiology, and so on is associated with MeSH [50].

2.2.4 Query Construction

Query construction is a wide topic that include it method of generation, sources of information, and style/structure. Many efforts have been made in this area and still in progress making the query process more intelligent and contextual-rich. In the sections below, few aspects of query construction are highlighted in the area of biomedical searching.

Information retrieval (IR) systems including search engines depends on the term importance that appear in the query and in the documents [51]. The concept of similarity i.e. sim(q, d) is used that determines the similarity between query q and document d. The similarity function is id defined in Eq. 2.1.

$$sim(q,d) = \sum_{t \epsilon q \cap d}^{n} w_{t,q}.w_{t,d} \tag{2.1}$$

where $w_{t,q}$ is the weight of term t in query q and $w_{t,d}$ is the weight of term t in document d, according to the systems weighting function. Further progress has been made to compute the weights beyond the scope of search engines by modifying the original query q to expanded query $q^{'}$ as shown in Eq. 2.2.

$$sim(q^{'},d) = \sum_{t \epsilon q^{'} \cap d}^{n} w_{t,q^{'}}.w_{t,d} \tag{2.2}$$

This part of research is accumulated to the concept of automatic query expansion (AQE). A decent amount of literature is available on AQE on its different aspects. Applications of AQE include question answering [52–54], multimedia information retrieval [55–57], information filtering [58, 59], cross-language information retrieval [60–62], and some other applications [63–65]. The process of AQE is divided in four steps: preprocessing of data source, expansion features generation and ranking, features selection, and query reformulation.

Preprocessing of data source [66, 67] include text extraction from the source documents, database, web pages, tokenization, stop word removal, word stemming, and word weighting. For candidate feature generation and ranking, a number of approaches are followed in the literature. The simplest form of feature generation is achieved through one-to-one association [68]. Sometimes one-to-many association is not applicable then the features are generated through one-to-many assication [69] technique. There are other approaches such as analysis of feature distribution in top-ranked documents [70, 71] and query language modeling [72, 73]. After generating and ranking the expansion features, the selection of features function is carried out. In feature selection, a body of research [74–76] is available. Finally, the query is reformulated to prepare a final searchable query. In this area, the most popular technique is modeled on Rocchio formula [77], which is subsequently revised in [70].

Automatic Query Generation

An effort has been made for automated medical literature retrieval [78] that retrieves medical articles by automatically generating query terms using patient record. It adds weights to each term in the query based on the importance calculated through the logic they developed. A tool called SmartQuery [79] is designed for the provision of context-sensitive links from electronic medical record to knowledge resources for the better access of clinical information in the clinical setting. For the facilitation of integration of clinical information system and the online knowledge resources, Health Level Seven International (HL7) CDS working group has developed context-aware knowledge retrieval standards called infobutton standards [80–83]. Infobuttots are enhanced with infobutton managers [84, 85], which anticipate the information needs a healthcare user is interested in a particular context.

In biomedical domain, PubMed search engine is used as a foundation search engine which itself has passed through many advancements in the last decade. A number of tools have been developed over PubMed to improvise the search results with respect to accuracy and performance efficiency. Zhiyong Lu has conducted a survey on PubMed and beyond [86] that enlist 28 web-based tools and services. These tools are categorized in four categories.

- Ranking search results (RefMed, MedlineRanker, SemanticMedeline, etc.)
- Clustring results into topics (GoPubMed, ClusterMed, XplorMed, etc.)
- Extracting and displaying semantics and relations (MedEvi, EBIMed, MEDIE, etc.)
- Improving search interface and retrieval experience (iPubMed, askMedLine, PICO, etc.)

A subset of systems from the list is selected that has some resemblance to the proposed approach in this thesis. The list shown in Table 2.1 provides the systems information with respect to thier profile information, major features, and limitations.

Table 2.1: PubMed derivatives with their features and limitations

Systems	**Year**	**Provider Profile**	**Major Features**	**Limitations**
RefMed [87]	2010	Academic	Featuring multi-level relevance feedback for ranking	1) User dependency for relevancy finding 2) Query non-automated
askMEDLINE [88]	2005	Govt	Converting questions into formulated search as PICO	Query non-automated
iPubMed [89]	2010	Academic	Allow fuzzy search and approximate match	Query non-automated

2.3 Summary of Related Work

On the basis of underline methodology and the working strategies, three systems enlisted in Table 2.2 in comparison of KnowledgeButton. KnowledgeButton is the system developed on the basis of proposed methods to be discussed throughout this thesis.

Table 2.2: Methodological comparison of proposed system with existing systems

	RefMed
How it works	A. [user to pubmed] Submit a keyword query B. [pubmed to user] Return initial results C. [user to pubmed] Feedback relevance D. [pubmed to user] Return ranked results Induces a relevance function using RankSVM and ranks the results according to the function
Methodology	Takes User clickthrough data for training, The method RankSVM ranks the documents according to the user preferences
	askMedline
How it works	1. User enters clinical question, search engine retrieve relevant journal articles. 2. for the user to check correctness of the PICO, a link is provided to the PICO interface. Round 1: i. query clean up. ii. send the query to pubmed iii. xml is returned, indicating category of each term used in query iv. Un-related terms are removed from the xml, remaining terms are sent back to the pubmed. Round 2: the search may proceed to Round 2, if: i. no results are returned in the first round ii. more than limit (50, 000) results returned, indicates that search term is too broad.
Methodology	PICO-based query generation
	iPubMed
How it works	provides a searching mechanism which is represented by two unique features: i. interactive: providing instant feedback to users as the query is being typed ii. fuzzy: allowing approximate search Interactive and fuzzy searching algorithms or exploring MEDLINE are implemented in a system called iPubMed
Methodology	i. Finding the predicted words of each keyword and the list of records that contains the predicted words. ii. identifying the predicted records by computing the intersection of the lists corresponding to different query keywords iii. ranking the answers.
	KnowledgeButton
How it works	It works in two models: Push and Pull Push Model: Phase I: A. Receives a clinical decision (and fired rules) as an input\ B. It extracts clinical terms from the fired rules C. Prepare the PICO question on the basis of standard concepts D. Augment the PICO question with clinical task (type of question) E. Run the query using PubMed Clinical Queries feature and get the articles Phase II: A. Assess the quality of the returned articles statistically. B. Rank the articles on the basis of user and article context aggregately Pull Model: Phase I: A. User provides input query in the form of pretext or a rule B. It extracts clinical terms from the pretext/rule C. Convert the terms into PICO format D. Augment the PICO question with clinical task (type of question) E. Run the query using PubMed Clinical Queries feature and get the articles Phase II: A. Assess the quality of the returned articles statistically. B. Rank the articles on the basis of user and article context aggregately
Methodology	A. PICO-Compliant Automated Query Construction based on KAP (Knowledge Alignment to PICO) B. Automated determination of clinical task using standard vocabulary service C. QRM-based Quality Assessment - QRM (quality recognition model) is based on SVM machine learning method. D. Evidence Contextual Grading -The user and article evidence contexts aggregation to grade an evidence in provision of contextual fitness.

Chapter 3

Proposed Methodology

To integrate health systems and CDSS with evidentiary support, we devised a two-step methodology as shown in 3.1. In step 1, the potentially relevant evidence reports are retrieved from the literature while in step 2, the retrieved evidence reports are evaluated to find scientifically rigorous or quality studies or systematic reviews.

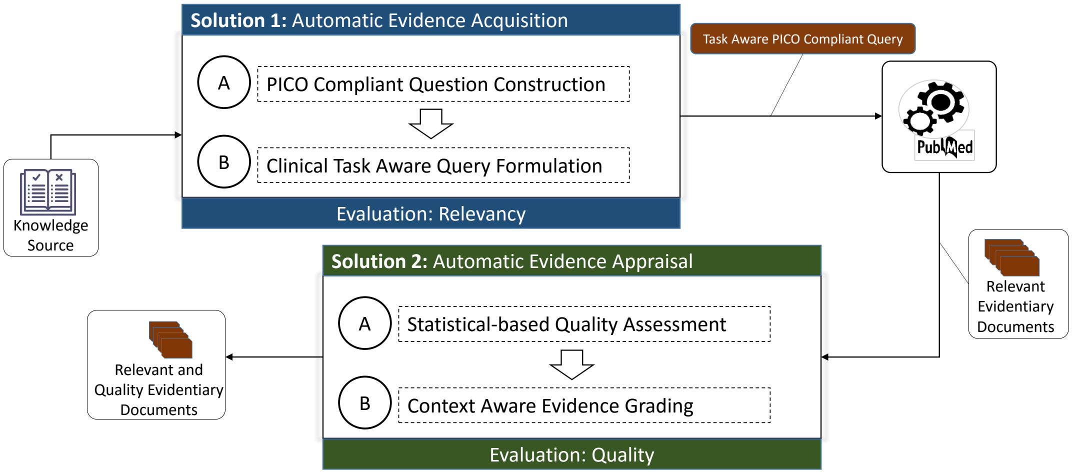

Solution 1: Automatic Evidence Acquisition
A
PICO Compliant Question Construction
B
Clinical Task Aware Query Formulation
Evaluation: Relevancy
Task Aware PICO Compliant Query
PubMed
Knowledge Source
Relevant Evidentiary Documents
Solution 2: Automatic Evidence Appraisal
A
Statistical-based Quality Assessment
B
Context Aware Evidence Grading
Evaluation: Quality
Relevant and Quality Evidentiary Documents

Figure 3.1: Abstract view of proposed methodology: two-step process of acquisition of potentially relevant evidences and appraisal of rigorous (quality) evidence recognition

Formally, step 1 is titled as ”solution 1” and step 2 as ”solution 2. Each of these solutions has two parts A and B. Solution 1 part A solve the problem of ”how to automatically construct a well-built question?”. The proposed solution for this problem is given as ”PICO-compliant question construction.” Solution 1 part B adds the clinical context to the question by identifying a clinical task and reformulate the query that is built in solution 1 part A. Solution 2 deals with post-retrieval evaluation of the documents for quality. In part A, the quality of the retrieved studies is assessed and in part B they are graded according to the user context in association with study context.

The details of this two-steps methodology is described comprehensively in 3.2. Each of the solutions is described by elaborating the internal functions and their relationships.

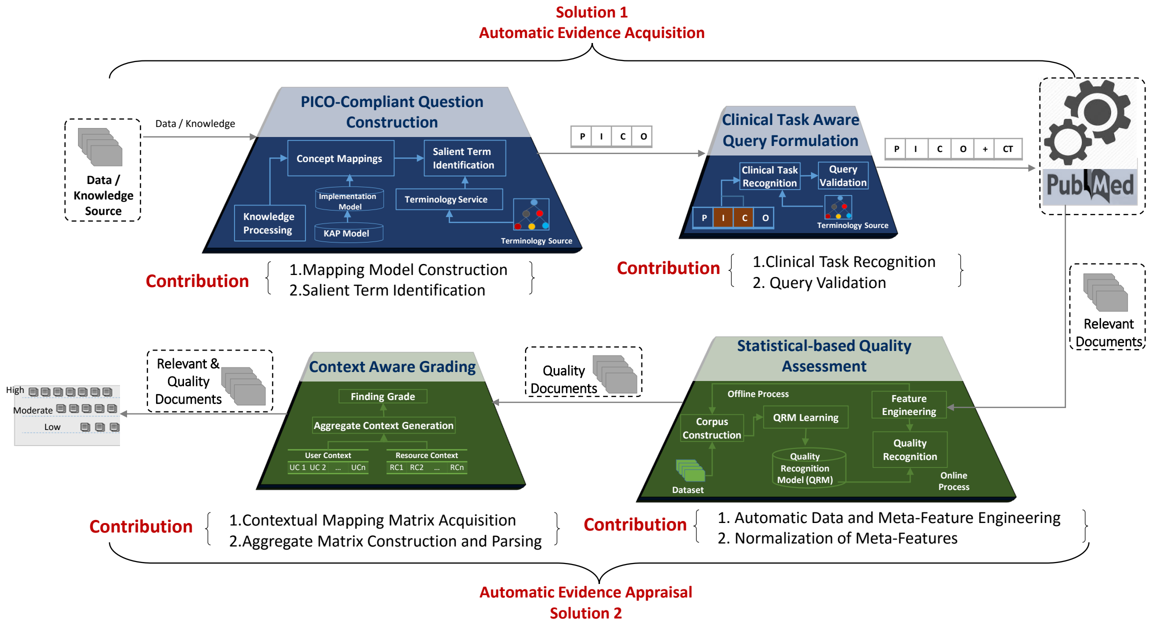

Solution 1
Automatic Evidence Acquisition
Data / Knowledge
Data / Knowledge Source
PICO-Compliant Question Construction
Concept Mappings
Salient Term Identification
Implementation Model
Terminology Service
Knowledge Processing
KAP Model
Terminology Source
P I C O
Clinical Task Aware Query Formulation
Clinical Task Recognition
Query Validation
P I C O
Terminology Source
P I C O + CT
PubMed
Contribution
1.Mapping Model Construction
2.Salient Term Identification
Contribution
1.Clinical Task Recognition
2. Query Validation
Relevant Documents
High
Moderate
Low
Relevant & Quality Documents
Context Aware Grading
Finding Grade
Aggregate Context Generation
User Context
UC 1 UC 2 ... UCn
Resource Context
RC1 RC2 ... RCn
Quality Documents
Statistical-based Quality Assessment
Offline Process
Corpus Construction
QRM Learning
Feature Engineering
Quality Recognition Model (QRM)
Quality Recognition
Dataset
Online Process
Contribution
1.Contextual Mapping Matrix Acquisition
2.Aggregate Matrix Construction and Parsing
Contribution
1. Automatic Data and Meta-Feature Engineering
2. Normalization of Meta-Features
Automatic Evidence Appraisal
Solution 2

Figure 3.2: Detailed view of proposed methodology: a two-step process of acquisition of potentially relevant evidences and appraisal of rigorous (quality) evidence recognition.

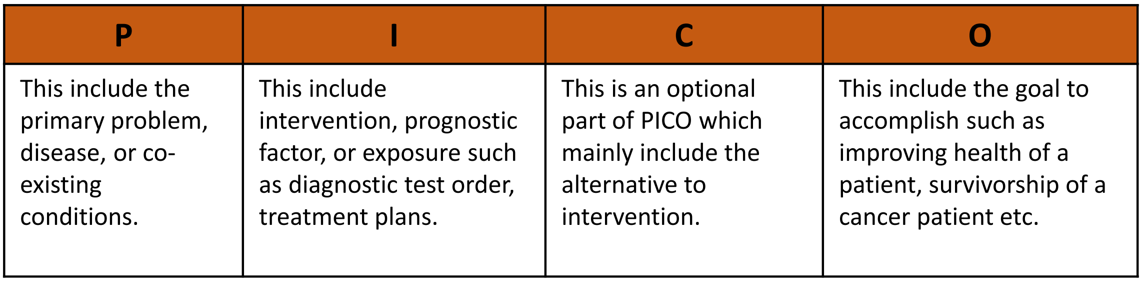
In the following section, solution 1 and solution 2 are briefly described with respect to the internal functions, input/output, and flow order of the functions.

3.1 Automatic Evidence Acquisition

Evidence acquisition search the possibly relevant evidence reports in the literature. The two essential sub-steps of evidence acquisition are; question construction and query reformulation. In question construction, a well-built question is framed in PICO format, while in query reformulation, the question is augmented with a clinical task filter in order to focus the retrieval set of results more target oriented. In proposed approach, the question is constructed automatically from the knowledge enclosed in a rule or rules of a clinical decision support system (CDSS).

3.1.1 What is PICO?

In EBM, it is very hard and time consuming to find appropriate evidentiary resources and recognize a relevant evidence. For that reason, practitioners of EBM are very much convinced to use a well-focused and well-formulated question while searching for a research evidence. For well-formulated questions, they use a specialized framework, called PICO which is stands for **P**atient problem, **I**ntervention, **C**omparison, and **O**utcome [90]. The PICO framework can be expanded to PICOTT to include **T**ype of question (therapy, diagnosis, prognosis, etiology, etc.) and **T**ype of study (clinical trial, randomized clinical trial, meta-analysis, etc.). Short description of P, I, C, and O are provided in Figure 3.3.

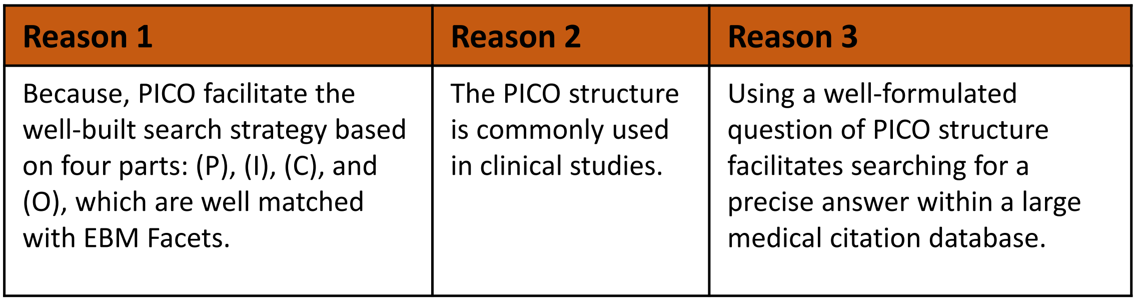


P	I	C	O
This include the primary problem, disease, or co-existing conditions.	This include intervention, prognostic factor, or exposure such as diagnostic test order, treatment plans.	This is an optional part of PICO which mainly include the alternative to intervention.	This include the goal to accomplish such as improving health of a patient, survivorship of a cancer patient etc.

Figure 3.3: PICO descriptions

3.1.2 Why PICO?

With PICO framework, the clinician understand and work with the important parts of the clinical question. The clinicians can easily identify the most applicable parts of the question to the patient and can manage the searching process by including the key concepts for an effective search strategy [91,92]. Some of the reasons why PICO is a preferable choice to use for question preparation are highlighted in [93] and are shown in Figure 3.4.



Reason 1	Reason 2	Reason 3
Because, PICO facilitate the well-built search strategy based on four parts: (P), (I), (C), and (O), which are well matched with EBM Facets.	The PICO structure is commonly used in clinical studies.	Using a well-formulated question of PICO structure facilitates searching for a precise answer within a large medical citation database.

Figure 3.4: Reasons to use PICO framed question

3.1.3 Clinical task importance in query

PICO-comliant question is a well-formulated question, however, making it more focused to a specific user task, type of the question is important to be recognized. The type of question is interchangeably used with a term ”clinical task”. PubMed has included this as a feature of ”Clinical Queries”. The Clinical Queries are based on the research carried out at McMaster university by the team of Hedges Study [94]. It has been envisioned that PICO framework together with the PubMed Clinical Queries can improve the retrieval efficiency. In this thesis, clinical task is automatically determined through the concepts used in I and C parts of PICO as they are matched in the SNOMED CT vocabulary using standard terminology service (STS). More details are described in chapter 4 on Automatic Evidence Acquisition.

3.2 Automatic Evidence Appraisal

Using the PubMed search engine, the results are retrieved based on the PubMed in-built information retrieval strategy, however, most times the results include a large number of documents that need to be checked again to reduce number of retrieved results by filtering out the less important ones. In other words, the appraisal process has to identify the quality evidence reports that are scientifically rigorous and remove the non-rigorous from the final set. For this purpose, a supervised classification model is learned, called appraisal model on training dataset.

3.2.1 Evidence Quality

Evidence quality is a subjective term. In EBM, it has been discussed from multiple aspects. This thesis uses the definition for quality at general level defined in the work of Wilczynski et al. [22].

Definition 1: *Evidence is considered scientifically rigorous if its analysis is consistent with the study design. The scientifically rigorous or methodologically rigorous articles are considered to be of high quality.*

For treatment related evidence quality, the definition of quality is further defined at a more granular level as described 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 major analysis at any given follow up assessment, and analysis consistent with study design.*

Based on these definitions, the quality of evidence is assessed through machine learning methods. Four machine learning methods are experimented in this work, where support vector machine (SVM) is finally chosen because of its better performance over the competing methods. The candidate selection of these four methods was based on the previous experiment carried out in the same domain [2, 7].

3.2.2 Contextual Evidence Grading

Contextual grading of evidence is necessary for re-ranking of the evidentiary documents. Current approaches such as [1] use SORT grading system. The SORT grading system is based on the evaluation of evidence characteristics, including publication type, avenue, year of publishing, authors, etc. User related contextual factors are missing while grading an evidence, which in this thesis has particularly focused. For derivation of user contextual factors, a context framework described in [95] and the PARIHS framework [96] are being followed. Figure 3.5 shows a general overview of user and evidence contexts.

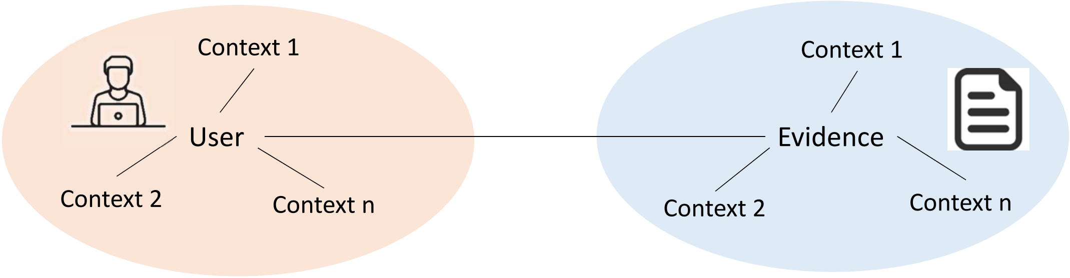


Figure 3.5: User and evidence contextual factors for grading

Chapter 4

Automatic Evidence Acquisition

Evidence acquisition search the possibly relevant evidentiary documents in the literature. This thesis present two functions for evidence acquisition.

- Query Construction where PICO-compliant question is constructed,
- Clinical Task Awareness where type of the question is identified.

4.1 Knowledge-based PICO-compliant question construction

PICO has well-defined structure with proper identification of patient problem related terms/concepts encapsulated in P part, intervention related terms/concepts in I part and so on. The issue raised when a knowledge encoded in a different structure that needs to be correctly map to corresponding part of PICO. As described previously, the proposed methodology support mechanisms to map different knowledge representation schemes such as production rules, medical logic modules (MLMs), and others through the use of mapping models. The complete process for constructing a PICO-compliant question is described in Figure 4.1

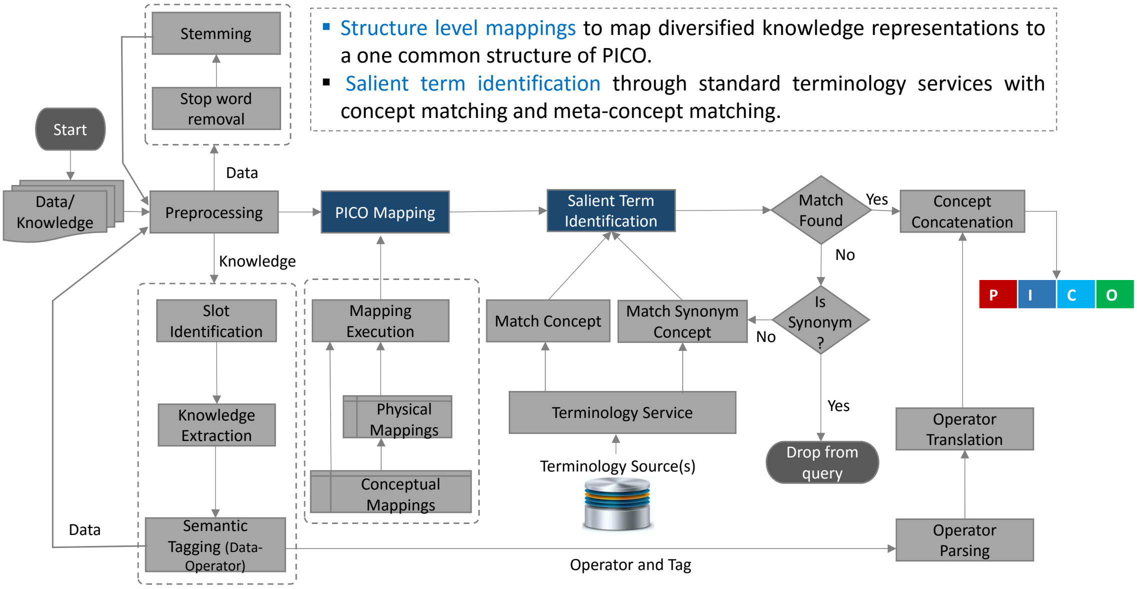


Figure 4.1: PICO-compliant question construction process

Generally, this process accepts clinical data and knowledge as an input. If the input is data, it is preprocessed to remove stop words and normalize it to the stem word. However, if the input is knowledge, it is preprocessed to identify the correct slot in the given knowledge representation scheme such as MLM, production rule, etc. Based on the slots, the concepts are extracted keeping in view the structure mapping information provided by the KAP (knowledge alignments to PICO) model. While extracting knowledge, there is need to parse the control structures and operators used in the knowledge representation scheme. For instance, MLM has a complex structure that involves different kind of control structures in developing logics for the knowledge representation. We developed control structure and operator parsing methodology described below.

4.1.1 Control Structure Parsing

Control structure used in the logic is parsed according to the grammar rules given in Table 4.1 and are described below. According to Arden Syntax, a number of variations of If-Then statements are used in the logic slot, such as Simple If-Then statement, If-Then-Else statement, If-Then-Elseif statement, Switch statement, Call statement, and others. Each of these statements is parsed according to the methods described below.

- Simple If-Then statement: The parsing process divides such statements into two sentences, the If sentence and the Then sentence. The concepts found in the If sentence are recognized as condition concepts, and concepts in the Then part are recognized as decision concepts, (see parsing example in Table 1 A).

- If-Then-Else statement: Such statements are parsed into three sentences. If a condition is satisfied in the If part, then the parsing is like a Simple If-Then statement. However, if a recommendation is found in the Else part, then the associated concepts are considered to be decision concepts, while the condition concepts in the If part are negated, (see parsing example in Table 1 B).

- If-Then-Elseif statement: Unless it is a last Else part, such statements are treated similar to a simple If-Then statement, with Elseif similar to If. The last Else part is handled similar to an If-Then-Else statement by considering Elseif as similar to If. For simplicity and to avoid multiple negations due to more than one ElseIf statement, we scoped the parsing to include immediate Elseif only, (see parsing example in Table 1 C).

- Nested If-Then statements: Sometimes an If statement occurs inside another If statement. In such cases, we consider the inner and outer statements as two conditions. For example, if a Simple If-Then statement occurs in another Simple If-Then statement, it is parsed into three sentences, If, If, and Then. Concepts in both if sentences are included in condition concepts, while concepts of the then sentence are included in decision concepts, (see parsing example in Table 1 D).

- Switch statement: The only case involving recommendation is the required segment where the concept value of case is considered as a condition, while the concepts in the body of that case are considered decisions, (see parsing example in Table 1E).

- Call statement: If a decision originates from the sub MLM, then the sub MLM is first parsed in reference to caller MLM through the ID. The executed paths of both caller and called MLMs are concatenated into one path, and the conditions are connected to each other accordingly, (see parsing example in Table 1 F).

Table 4.1: Control Structure Parsing Examples.

Example	Logic	Explanation
A	IF (C = v1) THEN	Condition sentence: C = v1 Decision sentence: D = d1
	D = d1	
	Output: d1 is recommended	
	END IF	
B	IF (C = v1) THEN	For CDSS output d1 is recommended: Condition sentence: C = v1 Decision sentence: D = d1
	D = d1	
	Output: d1 is recommended	
	ELSE	For CDSS output d2 is recommended: Condition sentence: C != v1 Where ! represents the negation (not). Decision sentence: D: d2
	D = d2	
	Output = d2 is recommended	
	END IF	
C	IF (C = vl) THEN	For CDSS output d1 is recommended: Condition sentence: C = v1 Decision sentence: D = d1
	D = d1	
	Output: d1 is recommended	
	ELSEIF (C in (v2, v3)) THEN	For CDSS output d2 is recommended: Condition sentence: C in (v2, v3) Decision sentence: D = d2
	D = d2	
	Output: d2 is recommended	
	ELSEIF (C = v3) THEN	For CDSS Output d3 is recommended Condition sentence: C = v3 Decision sentence: D = d3
	D = d3	
	Output = d3 is recommended	
	ELSE	For CDSS output d4 is recommended Condition sentence: C != v3 Decision sentence: D = d4
	D = d4	
	Output = d4 is recommended	
	END IF	
D	IF (C1 = v1) THEN	Condition sentence: C = v1 AND C2 = v2 Decision sentence: D = d1
	IF (C2 != v2) THEN	
	D = d1	
	Output = d1 is recommended	
	END IF	
	END IF	
E	Switch C	For CDSS output d1 is recommended: Condition sentence: C = v1 Decision sentence: D = d1
	case v1	
	D = d1	
	Output = d1 is recommended	For CDSS output d2 is recommended: Condition sentence: C = v2 Decision sentence: D = d2
	case v2	
	D = d2	
	Output = d2 is recommended	
	EndSwitch	
F	IF (C1 = v1) THEN	Condition sentence: C1 = v1 AND C2 = v2 Decision sentence: D = d2
	Call subMLM1	
	END IF	
	subMLM	
	IF (C2 = v2) THEN	
	D = d2	
	Output: d2 is recommended	
	END IF	

4.1.2 Operator Parsing and Concept Extraction

Operator parsing is the next step after structure parsing. In Arden Syntax, there is a pool of operators. For this work, we only parse the commonly used operators such as $or, and, =$

$, eq, is, isnotequal, <>, ne, isin, andin$. The operators $=, eq, andis$ are all parsed as equivalent to $=$. Similarly, $isnotequal, <>, andne$ are parsed as equivalent to the $!=$ operator. The binary operators $isinandin$ are parsed by including OR among the operands. Finally, the logical operators $and, or, andnot$ are parsed in the same order in which they occurred. The logical operator not has a key role in excluding the undesired elements from the retrieval set. Based on operator parsing, concepts are extracted as operands of the parsed operators for query construction. Also, the semantic tagging is performed to sustain the correct position of clinical terms and operators while concatenating the terms for the final query construction.

4.1.3 KAP Model

As illustrated in Figure 4.2, KAP is a mapping model providing mappings at two levels: at structure level and at concepts level.

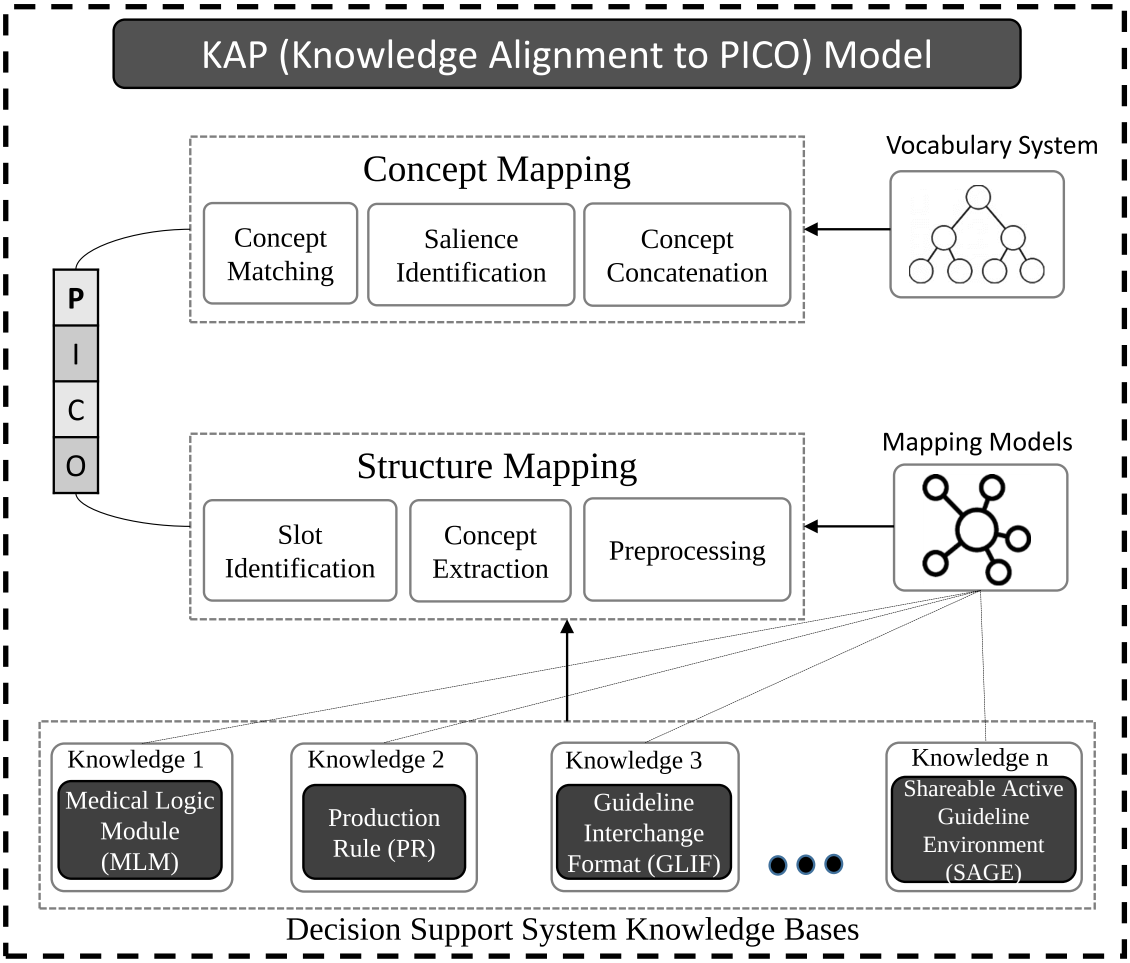


Figure 4.2: KAP: a mapping model for knowledge structure and concepts.

The structure level mappings provide the mechanism of different knowledge slots mapped to the corresponding part of PICO. Conceptually, KAP supports any knowledge representation provided that it has some defined structure. However, for real implementation, KAP is dependent on the implementation level mappings for each knowledge representation. The conceptual mapping model of KAP is provided in Eq. 4.1.

$$D \rightarrow P,\ A \rightarrow I,\ E \rightarrow C(optional),\ and\ P \rightarrow O(optional) \tag{4.1}$$

Where,

$D = Data$ (condition part of the rule),

$A = Action$ (decision part of the rule),

$E = Event$ (the intervention concept for invocation of the rule), and

$P = Purpose$ (the ultimate goal or outcome of the rule).

Each of these parts D, A, E, and P is a collection of concepts as represented in Eq. 4.2.

$$D = \bigcap_{i=1}^{n} DC_i,\ A = \bigcap_{i=1}^{n} AC_i,\ E = \bigcap_{i=1}^{n} EC_i,\ and\ P = \bigcap_{i=1}^{n} PC_i. \tag{4.2}$$

In structure level mappings, first of all required part (D, A, E, and P) are located in the input using slot identification function followed by concept extraction function that extracts all the concepts encoded in the identified slot. In order to map knowledge slots to PICO, KAP provides four type of mappings as shown in Eq. 4.3 in order to fill the PICO contents.

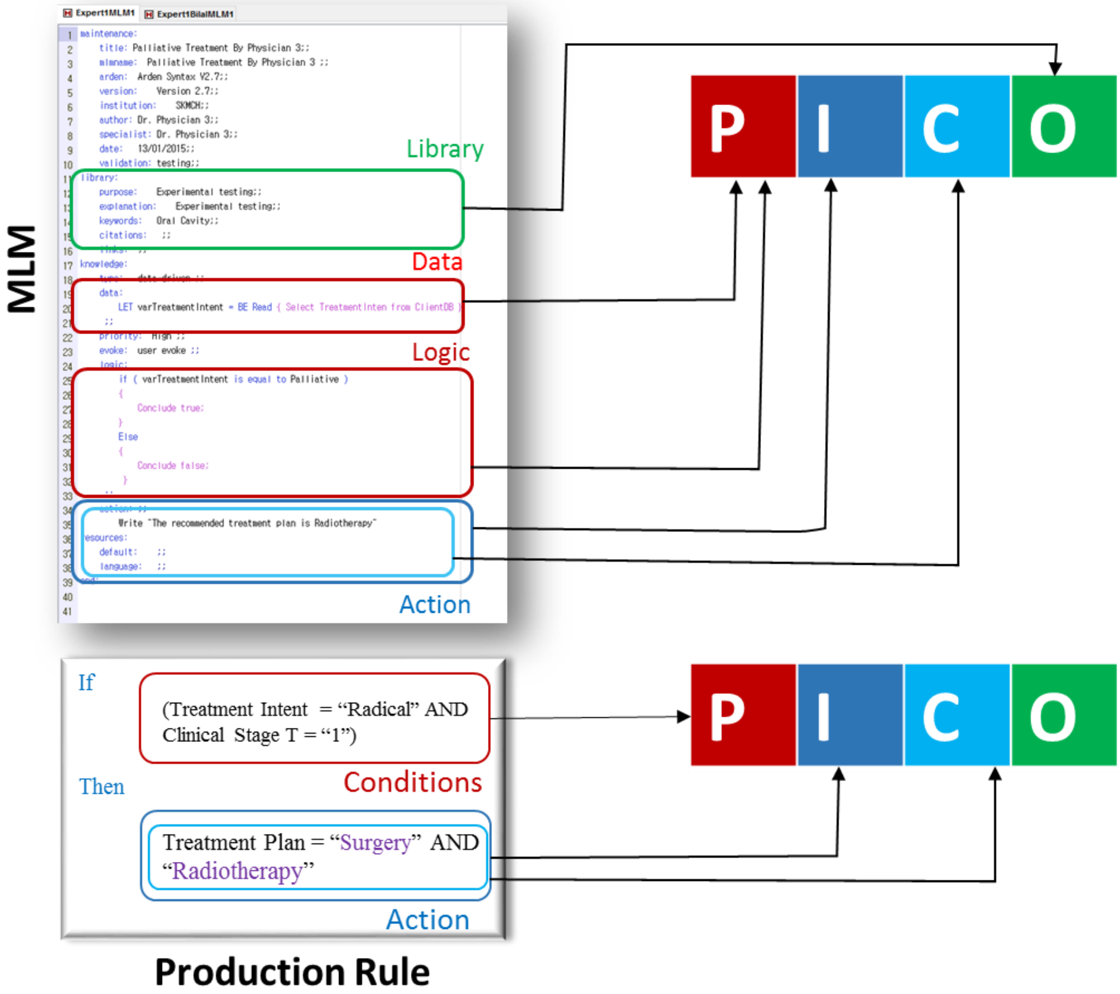
$$\begin{aligned} D &\to P \\ A &\to I \\ A &\to C \\ P &\to O \end{aligned} \tag{4.3}$$

Based on these mappings, a complete PICO structured is prepared consisting of D, A, E, and P sets of concepts. Each of these sets are logically Anded as shown in Eq. 4.4

$$\begin{gathered} PICO = D \wedge A \wedge E \wedge P \\ D = \bigcap_{i=1}^{n} DC_i\ \wedge\ A = \bigcap_{i=1}^{n} AC_i\ \wedge\ E = \bigcap_{i=1}^{n} EC_i\ \wedge\ P = \bigcap_{i=1}^{n} PC_i \end{gathered} \tag{4.4}$$

This is important to mention that each knowledge representation scheme has a different level of granularity to represent medical knowledge. MLM, for instance, is a more comprehensive knowledge representation scheme and has explicit slots for metadata, data, and knowledge, and is capable to provide information for all the slots of PICO. However, every knowledge scheme is not necessary to fill all the slots of PICO. Production rule, for instance, can provide only three parts:

P, I, and C. The outcome part O of PICO may not be able to map from conventionally represented production rule structure. In Figure 4.3, an example of MLM and production rule is provided to illustrate their structural differences in mapping to PICO different parts.



```
maintenance:
    title: Palliative Treatment By Physician 3;;
    mlmname:  Palliative Treatment By Physician 3 ;;
    arden:  Arden Syntax V2.7;;
    version:     Version 2.7;;
    institution:     SKMCH;;
    author: Dr. Physician 3;;
    specialist: Dr. Physician 3;;
    date:   13/01/2015;;
    validation: testing;;
library:
    purpose:    Experimental testing;;
    explanation:    Experimental testing;;
    keywords:   Oral Cavity;;
    citations:   ;;
    links:   ;;
knowledge:
    type:   data-driven ;;
    data:
        LET varTreatmentIntent = BE Read { Select TreatmentInten from ClientDB }
     ;;
    priority:  High ;;
    evoke:  user evoke ;;
    logic:
        if ( varTreatmentIntent is equal to Palliative )
        {
            Conclude true;
        }
        Else
        {
            Conclude false;
         }
     ;;
    action: ;;
        Write "The recommended treatment plan is Radiotherapy"
resources:
    default:    ;;
    language:   ;;
end:
```

(Treatment Intent = "Radical" AND Clinical Stage T = "1")

Treatment Plan = "Surgery" AND "Radiotherapy"

Figure 4.3: MLM and production rule structure mapping to PICO illustration example.

4.1.4 Concept Mapping

Concept mapping is the second level of mappings in KAP model where concepts extracted from the input sources, which are then mapped to PICO slots are checked for their importance to be included in the final PICO query. We propose an algorithm called STI (salience term identification) for concept matching in the standard vocabulary in order to find the most important terms in the sets of terms. We used IHTSDO SNOMED CT Restful API [97] that supports concept matching of user terms on SNOMED CT vocabulary. We develop our STI algorithm to perform text search over the database, selecting mode and other limits. Formally, STI methodology is represented in Algorithm 1.

Algorithm 1. Salient term identification (STI)

```
Begin
        inputs: $C - \{p, i, c, o\}$  # set of p, i, c, o concepts each with its n terms and each term with its initial weight = 0.0
        output: $picoQ$  # the set n returned results

        Initialize query
        foreach pico in C
            terms = pico.getTerms()
            type = pico.getType ()       # returns type of concept e.g. P, I, C or O
            foreach term in terms
                if exact_matched(term) then   # if concept is exact matched
                    term.weight = 1.0
                elseif partial_matched(term)
                        term.weight = 0.5
                endif
            endfor
            sort_by_weight_desc(pico)
            picoQ.concate( build_pico_query(pico, type) )
        endfor
        return picoQ
End
```

```
Procedure build_pico_query(T, Type)
Begin
        inputs: $T - \{c_1, c_2, ..., c_n\}$     # set of terms of  sorted by weight in DESC order
                Type   # the type of concept e.g. P, C, I or O
        output: $Q -$ optimized pubmed query

        counter = 0;
        if count_exact_match(T) greater_than 0 then  # e.g.  either exactly or partial matched
            foreach c in T
                if c.weight equals 1.0 OR c.weight equals 0.5 AND counter lessThan c.limit
                    Q.concate("(").concate(c.value)
                    Q.concate("OR")
                    Q.concate(c.matched_value).concate(")")
                    Q.concate("AND")
                    increment counter
                endif
            endfor
        else
            If Type notEquals C OR O
                Q.concate(c.value)
            endif
        endif
        return Q    # partial pico query
End
```

Description of Algorithm

The STI algorithm takes an input C that contains four sets of terms: p set of terms, i set of terms, c set of terms, and o set of terms. Initially, all of these terms are assigned with 0.0 weight value. The output of this algorithm is a PICO query consisting of standardized concepts matched from standard vocabulary. There three kinds of matching functions supported in the IHTSDO API: regex matching, partial matching, and full text matching.

- **Regex matching,** matches the query term exactly as a one unit. For example, a query term ”oral cavity” will return the concept of ”oral cavity cancer” but will not return the concept of ”oral and nasal cavity” because oral cavity in the latter case is not present as a one unit.
- **Partial matching,** matches the query term completely, but not necessary appeared as a one unit. Taking the same example as above, i.e. ”oral cavity” query term will rightly return both ”oral cavity cancer” and ”oral and nasal cavity” because both concepts have oral cavity term. However, it will not be able to return the concept ”oral cancer” as it misses the cavity part of the term.
- **Full Text matching,** matches the query term either in parts, completely, or exactly. The same ”oral cavity” term will return all the three concepts discussed in partial matching examples. Particularly the last term, i.e. ”oral cancer” will be successfully returned as it has at least one part of ”oral cavity” matched. The rule of thumb is this, a concept matched with Regex matching is also matched with partial matching as well as full text matching. Similarly, a concept matched with partial matching is also matched with full text matching. In both of these cases, the reverse is not true.

In algorithm, exact_matched function is used for regex matching and partial matching while partial_matched function is used for full text matching. The terms found with exact_matched are given weight value 1 and for the partial_matched terms 0.5 weight value is assigned. As all parts of PICO are not at the same level with respect to their importance and necessity in a query, so they needed to be dealt accordingly. An upper limit for each P, I, C, and O is provided in order to control the unnecessary lengthy queries. As provided in PICO guidelines, P and I parts are mandatory while C and O are optional parts. There is lacking of guidelines available for exact limit on the number of terms from each part of PICO to be included in the query, however, generally, a balance query shall not exceed the limit of 10 concepts. For the scope of thesis case study implementation, the limit for P terms is kept 3, followed by I terms 2. The C terms are dependent on I terms because they their alternatives so same limit is given to C as that of I , and finally O terms are given limit value of 1.

Each matched concept is concatenated with the corresponding query term using logical **OR** operator by putting parenthesis on the sides. This type of concatenation provides the expansion to the base terms for increasing the chance of retrieval. On the contrary, each part of PICO is concatenated using **AND** logical operator in order to form the final PICO query represented as **picoQ** in the algorithm.

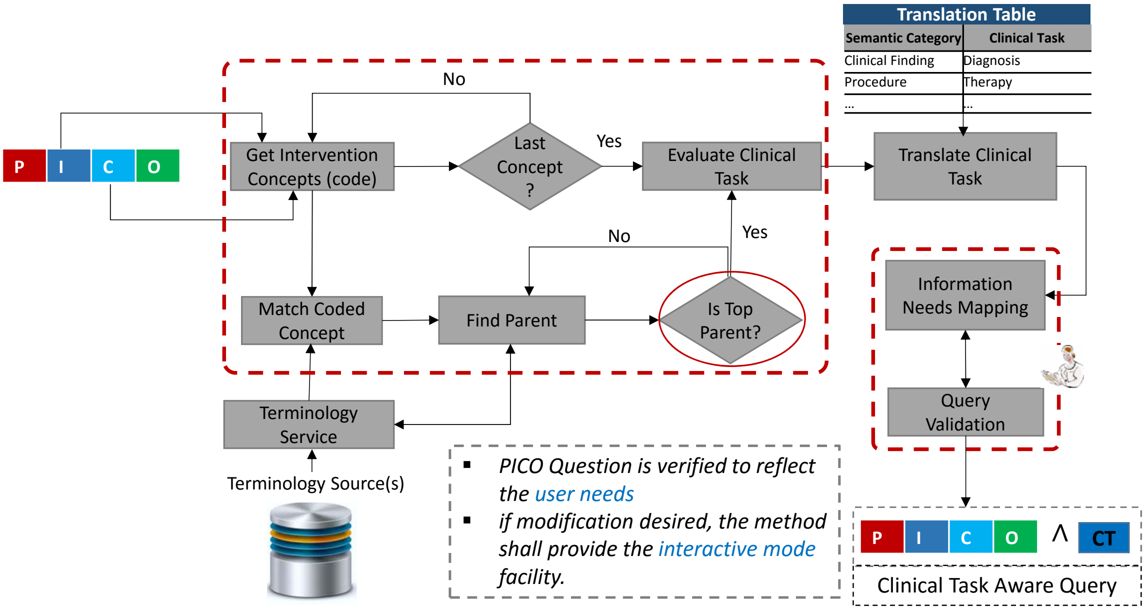
An important aspect to mention that semantic filters are provided in some parts of PICO such as for I and C, it was ”procedure”, while for P, it was ”clinical finding”. Providing semantic filters reduce the search space and decrease the chance of matching in the wrong category of SNOMED CT.

4.1.5 Data-based PICO compliant question construction: a variation of STI algorithm

The above version of STI algorithm works fine when it is pre-determined among the query terms that what set of terms belongs to P, I, C, and O. It is applicable when a query terms are derived from the knowledge as input source. Alternatively, if the input source is data rather than knowledge, then it needs to be decided that which set of terms belongs to which part of PICO. In this case, semantic filters such as ”procedure” and ”clinical findings” cannot be provided at the initial level of matching. In this case, the matching functions will be executed without semantic filter at first. If a match is found, its semantic type need to be checked. If it is ”clinical finding”, the concept will be considered as P concept. Similarly, if it returns ”procedure” semantic category, the concept is considered as I or C concept.

4.2 Clinical Task Awareness

Clinical Task represents the user task. Objectively, clinical task refers to the scenario or purpose of the query such as diagnosis, treatment, etc. PubMed provides the purpose filter through the advanced search facility and PubMed clinical queries specified four major types of clinical tasks including diagnosis, treatment, prognosis, and etiology. Applying clinical task filter minimizes the number of documents to be retrieved from the wrong categories. It becomes an extra activity for a user to choose what clinical task he/she wants regarding the query. To make this activity automatic, this work proposes an algorithm so-called clinical task recognition (CTR). The workflow of this algorithm is described in Figure 4.4.



Translation Table

Semantic Category	Clinical Task
Clinical Finding	Diagnosis
Procedure	Therapy
...	...

P
I
C
O
No
Get Intervention Concepts (code)
Last Concept ?
Yes
Evaluate Clinical Task
Translate Clinical Task
No
Yes
Match Coded Concept
Find Parent
Is Top Parent?
Information Needs Mapping
Query Validation
Terminology Service
Terminology Source(s)
PICO Question is verified to reflect the user needs
if modification desired, the method shall provide the interactive mode facility.
P
I
C
O
Λ
CT
Clinical Task Aware Query

Figure 4.4: Automatic clinical task recognition process

The PICO query constructed in the last step provides the concepts of its' parts I and C as input to the CTR algorithm. CTR takes the code of each concept repeatedly and find its parent concept unless it reaches the ultimate parent. Finally, when ultimate parents for all the concepts included in I and C are determined, they are evaluated to recognize the final clinical task for the given query. The identified clinical task determined using the standard terminology system such as SNOMED CT and UMLS may not the same to the labels used by the PubMed clinical queries. A translation table is utilized to translate the clinical task into a translated clinical task with possible values of diagnosis, treatment, etiology, and prognosis. The proposed CTR algorithm is formally represented in Algorithm 2.

Algorithm 1. Clinical Task Recognition (CTR)

```
Begin
        inputs: $\boldsymbol{IC} - \{c_1, c_2, \dots, c_n\}$ # concepts included in I and C of PICO query
        output: $\boldsymbol{CT}$ # the semantic category/ultimate parent of I and C concepts on majority basis

  1.    Let $\boldsymbol{TP} - \{p_1, p_2, \dots, p_n\}$
  2.    foreach $c$ in $IC$
  3.        parent = sts.getParent(c)
  4.            if topParent(parent) then   # if concept is exact matched
  5.                add parent to TP
  6.            else
  7.                Repeat step 3
  8.            endif
  9.    endfor
  10.   CT = findClinicalTask(TP)
  11.   return CT
End
```

```
Procedure findClinicalTask(TP)
Begin
        inputs: $\boldsymbol{TP} - \{p_1, p_2, \dots, p_n\}$ # set of top parent concepts
        output: $\boldsymbol{CT} - Clinical\ Task$

  1.    Let $\boldsymbol{countDist} - \{pCount_1, pCount_2, \dots, pCount_m\}$
  2.    foreach p in TP
  3.            pCount = countDistinct(p)
  4.            add pCount to countDist
  5.    Endfor
  6.    CT = getHighestCountValue(countDist)
  7.    return CT   # clinical task found out based on majority function
End
```

4.2.1 Clinical Task Translation

The clinical task is translated to the broad category which is interpreted by the PubMed Clinical Queries using a clinical task translation (CTT) algorithm. The parent term clinical finding and ”Disorder” concepts are translated to Diagnosis while procedure is translated to Therapy. CTT is formally represented in Algorithm 3.

Algorithm 3. Clinical Task Translation (CTT)

```
Begin
        inputs: $\beta$  # represents clinical task recognized from standard vocabulary system
        output: $\beta'$  # represents the clinical task used by PubMed Clinical Queries (diagnosis, treatment,..)

        if $\beta$ isEqual( clinical finding OR disorder )
                $\beta' \leftarrow Diagnosis$
        elseif
                $\beta' \leftarrow Procedure$
        endif
        return $\beta'$
End
```

4.3 Query Verification and Optimization

The automatically constructed PICO-compliant query along with clinical task is duly verified by the user for correctness. An interactive interface is developed to facilitate the user for the verification. If user wants to modify some concept in the query or adjust the clinical task, he/she can do so. The system adjusts the modification without losing the PICO semantics. The user verification process is illustrated in Figure 4.5.

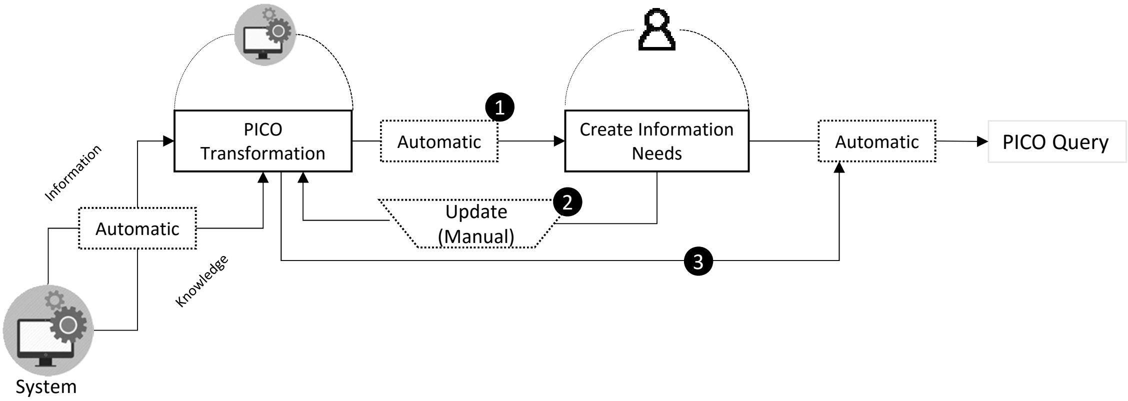


Figure 4.5: Query verification process by user

The verified query is ready to be converted into a format acceptable by PubMed. One more step is performed in case the executed query return zero document. In that case the query is optimized to truncate the least significant term. The least significant theory is formulated from the guidelines of PICO. In PICO, C and O are considered optional parts while P and I are mandatory. As shown in Figure 4.6 first, run the query with all terms included in the originally constructed query. If first run fails to return any document, truncate C term from the end and run the query. Repeat this process until getting at least one document in the retrieval set. Finishing C terms still zero document, start with O terms followed by I and P.

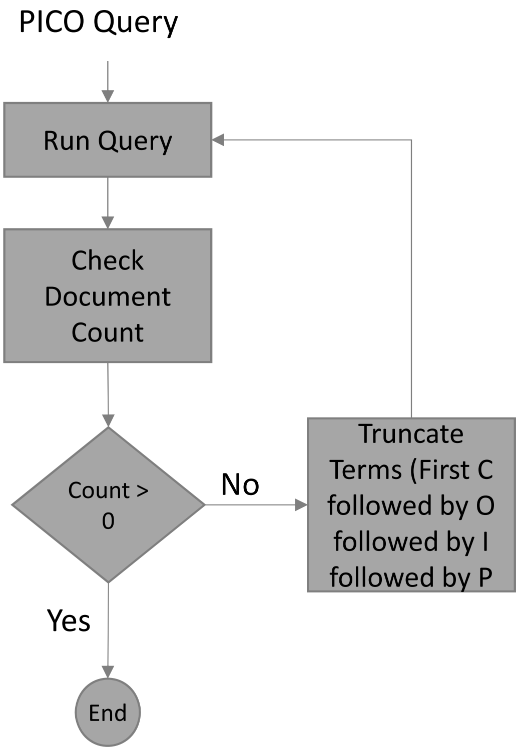

PICO Query
Run Query
Check Document Count
Count > 0
No
Yes
Truncate Terms (First C followed by O followed by I followed by P
End

Figure 4.6: Query optimization process

4.4 Evidence Searching

The context aware PICO compliant query represened as augmented query (AQ) is transformed formally to a universal resource locator (URL) acceptable to PubMed search engine. The formal structured of PubMed URL is specified in Eq. 4.12.

$$U ::=< BaseURL >< eUtils\ Method >< DB >< AQ > [< additionalFilters >] \tag{4.5}$$

Here, BaseURL is provided by PubMed as a compulsory part of the complete URL, DB represents the database name, i.e., PubMed. AQ is the augmented query generated with Eq. 4.11, and represent the filters that make the results more precise, such as journal and authors. These filters are pre-specified in system configuration, or the user can add/modify them at run time.

For the implementation of searching documents from within application, this thesis used Entrez API for the PubMed search service called Entrez Programming Utilities (eUtils) [98]. The eUtils provide a stable interface to the Entrez query and database system, including 23 databases on a variety of biomedical data. To access these data, a piece of software, first posts the eUtils URL to the database in order to retrieve the results. Using eUtils, we build a PubMed URL consisting of a Base URL and user query. We also employ the automatic term mapping (ATM) process provided by PubMed. ATM uses translation via MeSH for indexing and searching of the MEDLINE database of journal citations. A neglected term in the query is added to the MeSH term of the original query in order to access the MeSH field of MEDLINE documents. We implement three server functions of eUtils: ePost, eSearch, and eFetch. Using an ePost method, we create our own data set on the PubMed database. The eSearch method searches the relevant documents from the data set. Finally, using eFetch, the meta-information of each retrieved document is extracted, including title, author, journal name, publication year, identifier, and the link to the source document. These functions work in a sequence by using the output of one function as the input for another function. Figure 4.7 describes the step by step process of different functions involved in the query execution method in the form of a sequence diagram.

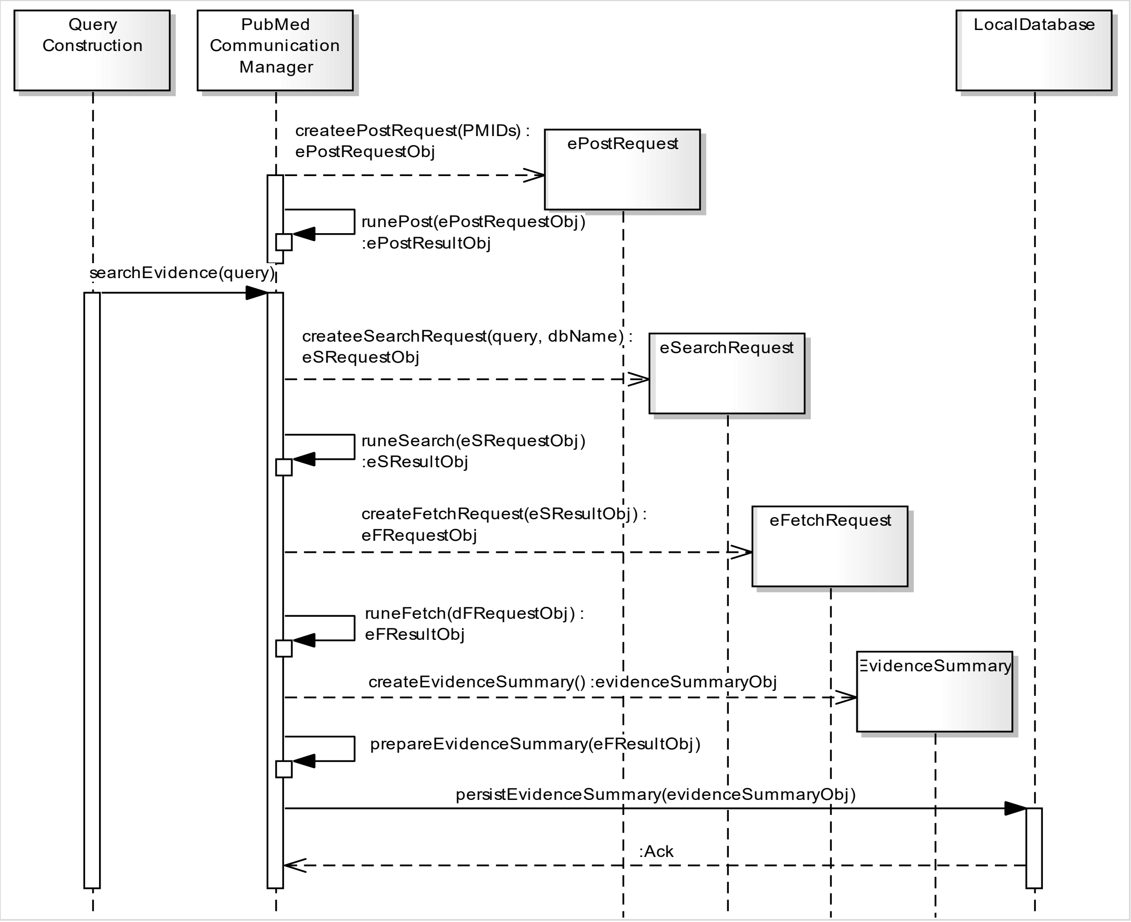

Query Construction
PubMed Communication Manager
LocalDatabase
createePostRequest(PMIDs) : ePostRequestObj
ePostRequest
runePost(ePostRequestObj) :ePostResultObj
searchEvidence(query)
createeSearchRequest(query, dbName) : eSRequestObj
eSearchRequest
runeSearch(eSRequestObj) :eSResultObj
createFetchRequest(eSResultObj) : eFRequestObj
eFetchRequest
runeFetch(dFRequestObj) : eFResultObj
EvidenceSummary
createEvidenceSummary() :evidenceSummaryObj
prepareEvidenceSummary(eFResultObj)
persistEvidenceSummary(evidenceSummaryObj)
:Ack

Figure 4.7: Process of Evidence Searching through eUtils service implementation

4.4.1 Running Examples

Suppose a clinical decision is made involving a rule R with the following information. This rule is derived from a decision tree which is developed for oral cavity patient treatment plan recommendation [99].

R = If (TreatmentIntent = Radical AND Clinical Stage T = (”T3” OR ”T4”) AND Histology= ”Squamous Cell Carcinoma”) Then (Treatment Plan = ”C CRT”)

In Rule R, there are five terms represented as QTerm. These terms include four condition terms

(Radical, T3, T4, and Squamous Cell Carcinoma) and one composite term, a decision term (C CRT). C CRT is a composite term because it consists of two terms, C for Chemotherapy and CRT for Chemoradiation Therapy. Among the four condition terms, three terms (T3, T4, and Squamous Cell Carcinoma) belong to the problem identified with ”STI” algorithm. The conclusion term C CRT is also identified with ”STI” algorithm. These four terms are assigned to α, as shown in Eq. 4.6.

$$\alpha = \{(T3\,OR\,T4)\,AND\,Squamous\,Cell\,Carcinoma\,AND\,C\,CRT\} \tag{4.6}$$

The clinical task is determined on the basis of intervention term ”C CRT” which is a composite term, where the term C is a root term and the second term CRT is an associated term. On the basis of the root term C, the parent term is determined. Using CTR algorithm, the ultimate parent of C is found, which the procedure is assigned to β as shown in 4.7.

$$\beta = \{Procedure\} \tag{4.7}$$

4.4.2 Query augmentation

The query is augmented with translation mechanisms by adding variants and alternative terms to each term of the query. PubMed provides automatic term mapping (MAP) for the augmentation of terms. MAP uses a MeSH translation table [100], which contains not only MeSH terms and MeSH Subheadings, but also the terms derived from the Unified Medical Language System (UMLS) that has equivalent synonyms or lexical variants in English. Using this option of PubMed saves the effort of adding other terminological variants such as SNOMED CT, which has normally been performed in the past. Similarly, for query type terms, Algorithm 3 (translate query type) is used.

Running Example: The augmentation process is applied on ”α” , transforming it into ”$\alpha^{‘}$”, as shown in Eq. 4.8.

$$\propto' ::= \left\{[t_{i'}] < logicalOp > t_{j'}\right\} \, where \; t_i{}' = t_i\left[\{OR\; et_i\}\right],\; t_{j'} = t_j\left[\{OR\; et_j\}\right] \; and \; t_i, t_j \; \in ST \; and \; et_i, et_j \; \in ET \tag{4.8}$$

The translated form of the example in Eq. 4.6 is presented in Eq. 4.9.

$$\alpha' = \left\{ \begin{array}{c} (T3[All\ Fields]\ OR\ T4[All\ Fields])\ AND\ (\text{"}carcinoma,\ squamous\ cell\text{"}[MeSH\ Terms] \\ OR\ \left(\text{carcinoma}\left[All\ Fields\right] AND\ \text{squamous}\left[All\ Fields\right] AND\ \text{cell}\left[All\ Fields\right]\right) \\ OR\ \text{"}squamous\ cell\ carcinoma\text{"}[All\ Fields]\ OR\ (\text{"}squamous\text{"}[All\ Fields] \\ AND\ \text{"}cell\text{"}[All\ Fields]\ AND\ \text{"}carcinoma\text{"}[All\ Fields])) \\ AND\ (C[All\ Fields]\ AND\ CRT[All\ Fields]) \end{array} \right\} \tag{4.9}$$

Using the translation algorithm, the term Procedure in example Eq. 4.7 is translated ”Therapy” as shown in Eq. 4.10.

$$\beta^{'} = \{Therapy\} \tag{4.10}$$

Finally, and are combined to form the augmented query AQ, as shown in Eq. 4.11.

$$AQ = \left\{ \begin{array}{c} (T3[All\ Fields]\ OR\ T4[All\ Fields])\ AND\ (\text{"}carcinoma,\ squamous\ cell\text{"}[MeSH\ Terms] \\ OR\ \left(\text{carcinoma}\left[All\ Fields\right] AND\ \text{squamous}\left[All\ Fields\right] AND\ \text{cell}\left[All\ Fields\right]\right) \\ OR\ \text{"}squamous\ cell\ carcinoma\text{"}[All\ Fields]\ OR\ (\text{"}squamous\text{"}[All\ Fields] \\ AND\ \text{"}cell\text{"}[All\ Fields]\ AND\ \text{"}carcinoma\text{"}[All\ Fields])) \\ AND\ \left(C\left[All\ Fields\right] AND\ CRT\left[All\ Fields\right]\right)\ AND\ \text{"}Therapy\text{"} \end{array} \right\} \tag{4.11}$$

The augmented query is transformed into a universal resource locator (URL) required by the target search engine. The URL generator sub-component structures the queries in a specified URL U in accordance with Eq. 4.12.

$$U \ ::=< BaseURL >< eUtils\ Method >< DB >< AQ > \ [< additionalFilters >] \tag{4.12}$$

Running Example: Extending the query Q in running example 2, the generated URL U is shown in Eq. 4.13.

$$U^{'} = \left\{ \begin{array}{c} http\text{://eutils. ncbi. nlm. nih. gov/entrez/eutils/eSearch. fcgi?/db} = \text{pubmed} \\ term = (\ (T3[All\ Fields]\ OR\ T4[All\ Fields])\ AND\ (\ carcinoma, squamous\ cell \\ [MeSH\ Terms]\ OR\ (\ carcinoma\ \ [All\ Fields] AND\ \ squamous \\ [All\ Fields] AND\ \ cell\ \ [All\ Fields])\ OR\ \ squamous\ cell\ carcinoma \\ [All\ Fields]\ OR\ (\ squamous\ [All\ Fields]\ AND\ \ cell\ [All\ Fields] \\ AND\ \ carcinoma\ [All\ Fields]))\ AND\ (C[All\ Fields]\ AND\ CRT[All\ Fields])\ ) \\ \#clincat = Therapy, Broad \end{array} \right\} \tag{4.13}$$

The URL U is executed on Entrez API [98] using its different services including ePost, eSearch, and eFetch. The ePost method is utilized to create a custom database in PubMed for local processing. eSearch is used for basic searches, while eFetch returns the documents (for the PMIDs already searched with the eSearch method). The eFetch method utilizes the output of eSearch by retaining the history and environment variables to maintain the previous histories and avoid the repeated retrieval of documents. The eSearch and eFetch methods are used in a pipeline approach of executing eSearch first, and the output is used as input to eFetch. The response is utilized by the Evidence Appraisal component to appraise the evidence reports for further application.

Chapter 5

Automatic Evidence Appraisal

We propose a hierarchical strategy for the assessment 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 quality recognition model (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 context aware grading (CAG) method.

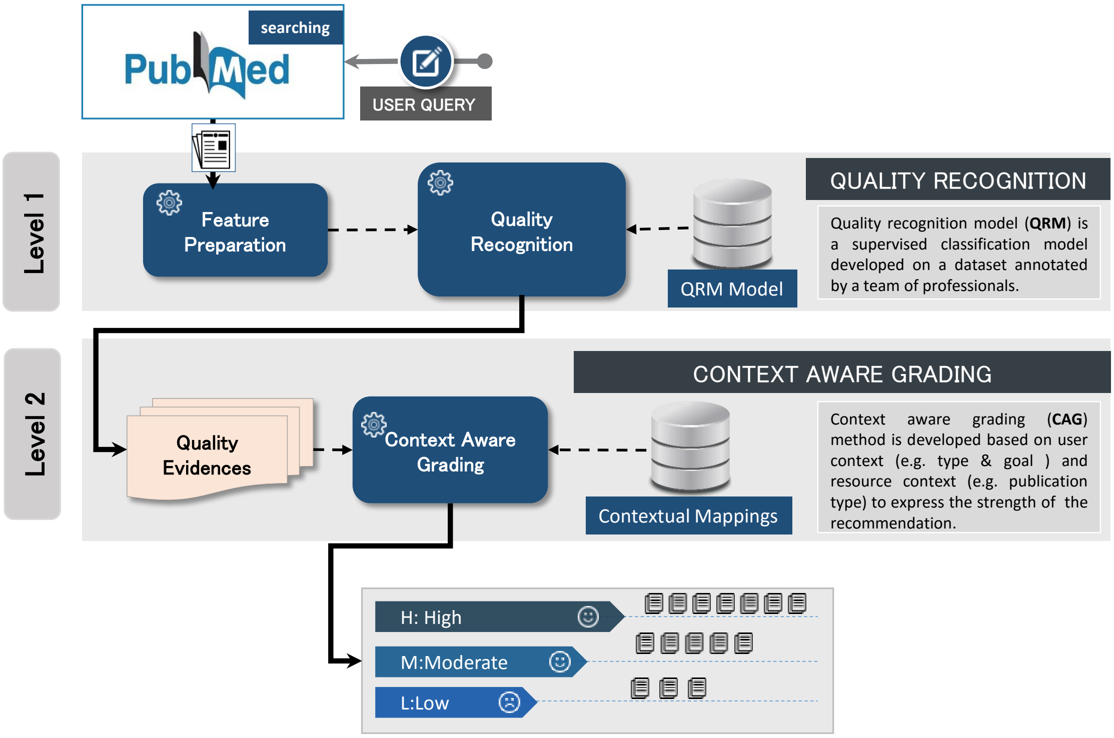

searching
PubMed
USER QUERY
Level 1
Feature Preparation
Quality Recognition
QRM Model
QUALITY RECOGNITION
Quality recognition model (QRM) is a supervised classification model developed on a dataset annotated by a team of professionals.
Level 2
Quality Evidences
Context Aware Grading
Contextual Mappings
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 5.1: Two level evidence evaluation: quality recognition and context aware grading

5.1 Level 1: Quality Recognition

Prior describing the method of quality evidence recognition, it is necessary to agree upon quality parameters. The 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 literature for clinical care. Strength of Recommendation Taxonomy (SORT) [27] 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 [27]. 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 with consistent findings or high quality individual randomized controlled trial [27].*

In a study protocol [22], 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 [22].*

For this study, definition 2 is considered for quality evaluation of the evidences. For quality evaluations, we develop a supervised classification model called quality recognition model (QRM). We follow the steps of: data collection; feature selection; corpus preparation; algorithm selection; and parameter tuning for QRM development. The complete process of quality recognition process is described in Figure 5.2.

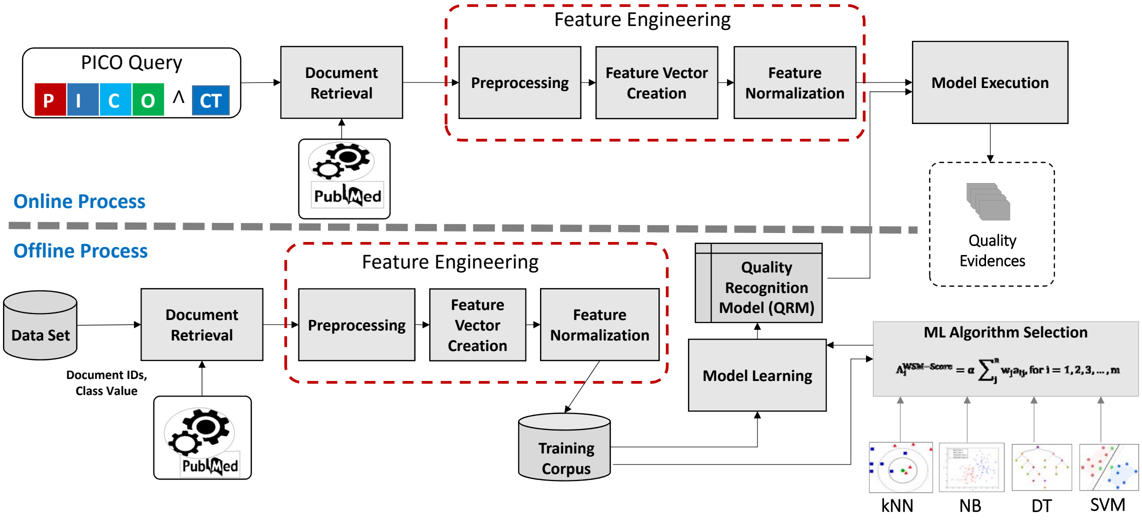
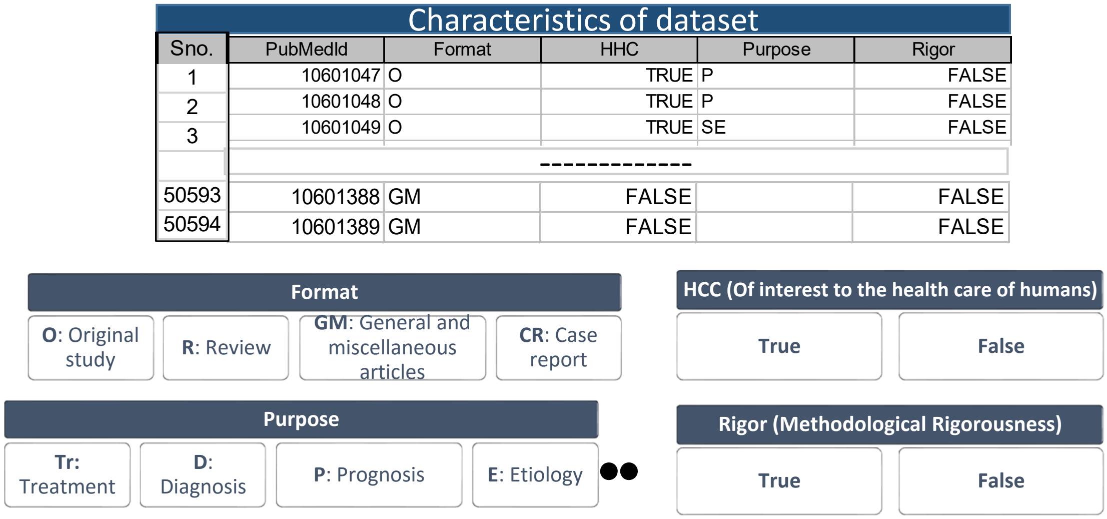


Figure 5.2: Quality recognition model (QRM) learning and execution process

5.1.1 Data collection for the appraisal model

We use the data that was 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 health care interest (yes/no), scientific rigor (yes/no), and purpose (diagnosis, etiology, prognosis, treatment, economic studies, reviews, and clinical predication guides). Among 50,594 documents, 3,363 are labeled as being scientifically rigorous. Brief description of the dataset is provided in Figure 5.3.



Characteristics of dataset

Sno.	PubMedId	Format	HHC	Purpose	Rigor
1	10601047	O	TRUE	P	FALSE
2	10601048	O	TRUE	P	FALSE
3	10601049	O	TRUE	SE	FALSE

50593	10601388	GM	FALSE		FALSE
50594	10601389	GM	FALSE		FALSE

Figure 5.3: Dataset descriptions

5.1.2 Document downloading and parsing

The collected dataset contains PubMed Identifiers (PMIDs), which we post to create a custom database in PubMed through the ePost service method of the eUtils API. We search the database using the eSearch service of the eFetch service by enabling history and environmental variables. Using eFetch, documents are downloaded and parsed to analyze the data features (title, abstract) and meta-data features (MeSH terms and article type). The processes of downloading and parsing the documents are described in Algorithm 6.1.

Algorithm 4: Downloading and parsing documents

```
Begin
        inputs: $PMIDs - \{id1, id2, \dots, idn\}$; //list of PubMed ids of training dataset
        output: $F - \{f_1, f_2, f_3, f_4\}$ /* where $f_1 = title, f_2 = abstract, f_3 = MeSH, and\ f_4 = publication\ type$ */

    Let $ePostResultRef$ is the reference to the database of uploaded IDs
    $ePostResultRef \leftarrow ePost(PMIDs)$; //upload the PMIDs list to PubMed database
    $eFetchResult \leftarrow eFetch(ePostResultRef)$; //download the documents

    $for\ i = 0\ to\ eFetchResult.count - 1$
        $f_1 \leftarrow i.title$;
        $F.add(f_1)$;
        $f_2 \leftarrow i.abstractText$;
        $F.add(f_2)$;
        $f_3 \leftarrow ""$;
        $for\ j = 0\ to\ i.MeSHHeading.count - 1$
            $f_3 \leftarrow f_3 + i.MeSHHeading$;
        $endfor$
        $F.add(f_3)$;
        $f_4 \leftarrow ""$;
        $for\ m = 0\ to\ i.publicationtype.count - 1$
            $f_4 \leftarrow f_4 + ", " + i.publicationtype$;
        $endfor$
        $F.add(f_4)$;
    $endfor$

    Return F;
End
```

5.1.3 Feature selection

Feature selection plays an important role in predicting performance. From the existing studies, we come across features including data features (title, abstract), 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 [2]. The data features are used in studies [1, 2] have proved their importance. Metadata features publication type is the most important feature reported from the same studies. MeSH terms is also reported in [2] 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.

5.1.4 Preprocessing

The data feature vector is created by tokenizing the titles and abstracts, changing the case to lower, eliminating the stop words, stemming the words using the Porter stemmer [101], and filtering the tokens by length, with a minimum of 2 characters and a maximum of 999 characters. The prune method is chosen based on absolute value, with a minimum absolute value of 2 and a maximum absolute value of 100. Pruning below absolute 2 means to ignore words that appear in less than 2 documents, and pruning above absolute 100 means to ignore words that appear in more than 100 documents. The pruning reduces the number of regular attributes (nave features created in tokenization steps) from 9518 to 5049 (dependent variables) with no significant impact on performance. Unlike data features, the metadata features are created by applying only tokenization and case transformation, as there is no need to remove stop words or perform stemming. The processes of vector creation for data features and metadata features are shown in Figure 5.4.

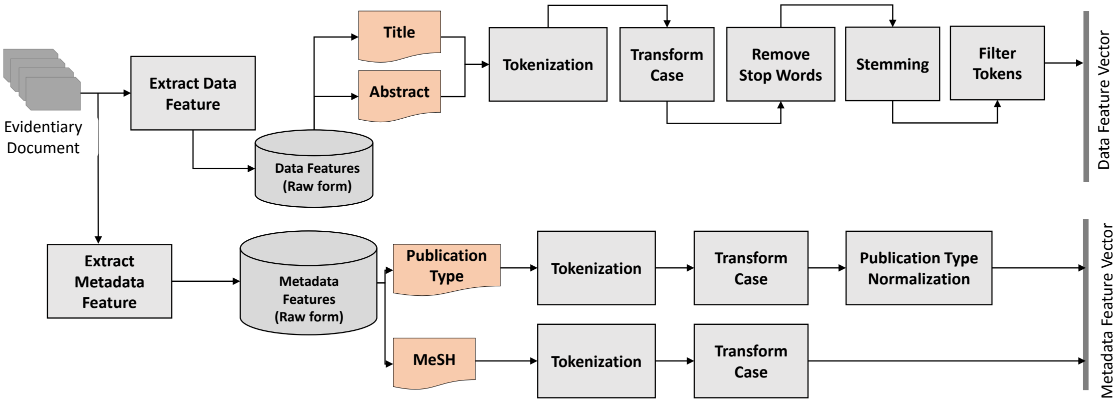


Figure 5.4: Data and metadata feature vector creation process

5.1.5 Standardizing language of publication type

The publication types text retrieved through eUtils API [98] are not consistent with the vocabulary of publication types provided by PubMed. Publication types found in PubMed are reported in count as 73 [102], which is quite less than the count 248 returned for the documents in our selected dataset. Algorithm 6.2 mapped the inconsistent publication types to standard publication types taking the list of articles as input.

Algorithm 1. Standardizing language of publication types

```
Begin
        inputs: $A - \{a_1, a_2, \dots, a_n\}$; //the list of articles
        output: $A' - \{a_1, a_2, \dots, a_n\}$; // the list of articles with standardized publication type

    Let;
        $\boldsymbol{pt}$ represents publication type;
        $\boldsymbol{rank}$ represents the rank of $\boldsymbol{pt}$;
        $\boldsymbol{tempRank} = 0$; // holds the previous rank temporarily for comparison
        $\boldsymbol{spt}$ $represents\ the\ standardized\ publication\ type$;
    for each a in A
        do
            $\boldsymbol{pt} \leftarrow \boldsymbol{a}.getPublicationType()$;
            $\boldsymbol{rank} \leftarrow getRank(\boldsymbol{pt}, \boldsymbol{R})$; //where R is the rank table for publication types.

            $\boldsymbol{if}\ (\boldsymbol{rank} > \boldsymbol{tempRank})$
                $\boldsymbol{tempRank} \leftarrow \boldsymbol{rank}$;
                $\boldsymbol{spt} \leftarrow \boldsymbol{pt}$;
            $\boldsymbol{endif}$
        $\boldsymbol{while}\ (\boldsymbol{a}.getPublicationType\ exists)$

        $\boldsymbol{a}.\boldsymbol{PublicationType} \leftarrow \boldsymbol{spt}$;
        $\boldsymbol{A}'.\boldsymbol{add}(\boldsymbol{a})$;
    $\boldsymbol{endfor}$
    $return\ \boldsymbol{A}'$;
End
```

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 is derived from the literature evidences in [22, 27, 103–105] 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 randomized controlled trial (RCTs) is considered the most important publication type for treatment so it is ranked on top by assigning value 1. Table 5.1 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.

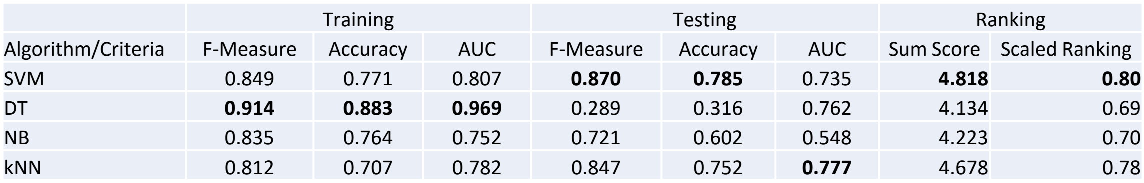
Table 5.1: Rank values of publication types (1 shows the highest and 4 is the lowest)

Publication Type	**Rank**
Meta-analysis of RCTs	1
Systematic Review of RCTs	2
Randomized Controlled Trials (RCTs)	3
Meta-analysis of CTs	4
Systematic Review of CTs	5
Control Trials (CT)	6
Cohort Study	7
Case-control study/report	7
Guidelines	8
Opinion	9
Observational Study	10
Any other publication type	11

5.1.6 Machine learning method selection

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 [106, 107]. For the chosen methods: Nave Bayes (NB) kernel [108]; k-Nearest Neighbor (kNN) [109]; support vector machine (SVM) linear [110]; and decision tree (DT) [111], we tested the performance at different parameter setting. NB is experimented with kernel values 5, 10, and 15 with a minimum 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 to 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 settings of parameter 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 almost similar results to C = 0.1. The kernel cache value is set to 200 and maximum iterations is set to 100000. Finally we were eft 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.

A subset of the selected dataset is chosen for the experiment to find quality evidence in treatment related documents. The subset includes 6882 documents out of which 4999 are labelled as Non Rigor and 1883 are labelled as Rigor. We determine the performance of chosen methods on precision and area under curve (AUC) criteria (Figure 5.5). Accuracy is 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 literature, it is reported that AUC is statistically consistent and more discriminant than accuracy [112,113]. 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.

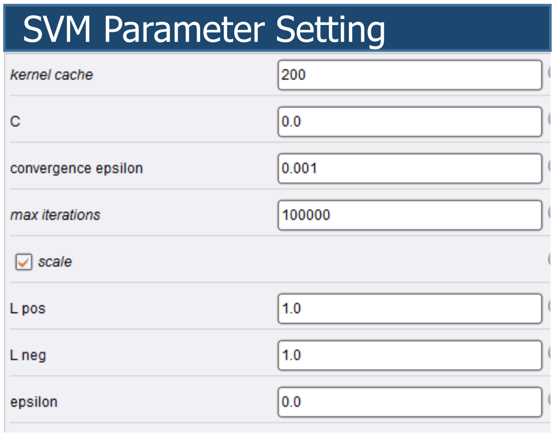


	Training			Testing			Ranking	
Algorithm/Criteria	F-Measure	Accuracy	AUC	F-Measure	Accuracy	AUC	Sum Score	Scaled Ranking
SVM	0.849	0.771	0.807	**0.870**	**0.785**	0.735	**4.818**	**0.80**
DT	**0.914**	**0.883**	**0.969**	0.289	0.316	0.762	4.134	0.69
NB	0.835	0.764	0.752	0.721	0.602	0.548	4.223	0.70
kNN	0.812	0.707	0.782	0.847	0.752	**0.777**	4.678	0.78

Figure 5.5: Best performing classifiers with accuracy and AUC on data + metadata features

5.1.7 SVM parameter setting

Because of the higher performance, SVM is chosen for the development of quality recognition model (QRM). In Figure 5.6, description of different parameters set for the algorithm is described.



SVM Parameter Setting

kernel cache	200
C	0.0
convergence epsilon	0.001
max iterations	100000
☑ scale	
L pos	1.0
L neg	1.0
epsilon	0.0

- Complex cost parameter C values less than 0.0 showed similar results to C = 0.0.
- Similarly, values greater than 0.1 produces almost similar results to C = 0.1.
- The kernel cache value is set to 200 and maximum iterations is set to 100000.
- 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.

Figure 5.6: SVM parameter setting

5.2 Context Aware evidence Grading (CAG)

Evidence recognition on the basis of user query and statistical methods may not fully determine the user preferred evidences. The statistical approach described in section 5.1 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. Katrien Vebert et al. [95] 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, users 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 [95, 96, 114], 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.), year of publication. For grading an evidence, we design a method as shown in Figure 5.7 and described in Algorithm 6.3, which evaluates an evidence on the basis of different user context elements.

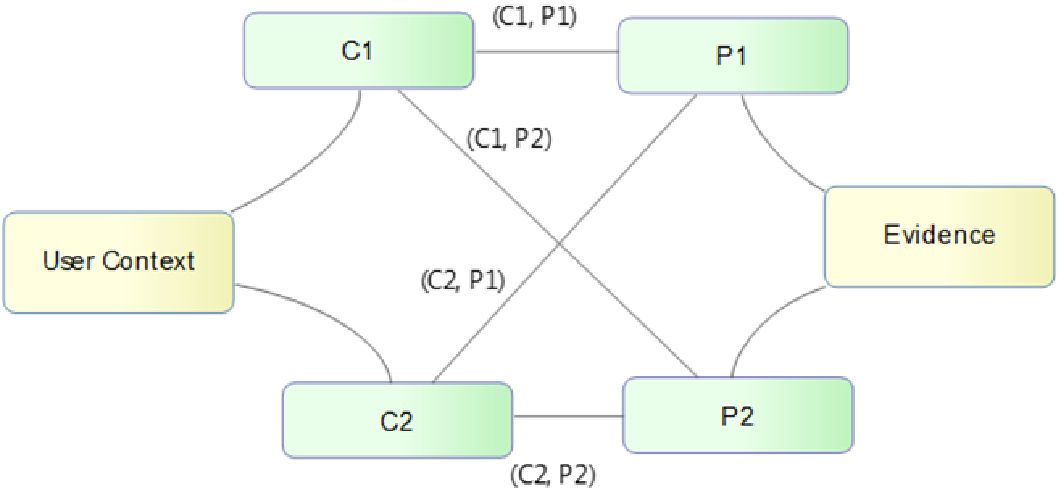


Figure 5.7: User context mapping with Evidence properties

Algorithm CAG. Grading evidence on the basis of user context

```
Begin

        input:  $E - \{e_1, e_2, \dots, e_n\}$; //the list of rigor evidences
        output: $GE - \{\{e_1, g_1\}, \{e_2, g_2\}, \dots, \{e_n, g_n\}\}$; // where g represents the grades h, m, l, u.

Let;
    $C - \{c_1, c_2, \dots, c_n\}$; //current context
    P - $\{p_1, p_2, \dots, p_n\}$; //properties of E
    G - $\{g_1, g_2, \dots, g_n\}$; //properties of E

for each e in E
    for each p in P
        for each c in C
            $grade \leftarrow computeGrade(p, c)$;
            $G.add(grade)$;
        endfor
    endfor
    $finalGrade \leftarrow getHighestGrade(G)$;
    $GE.add(e, finalGrade)$;
endfor
return GE;

End
```

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 possess the context 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 5.2.

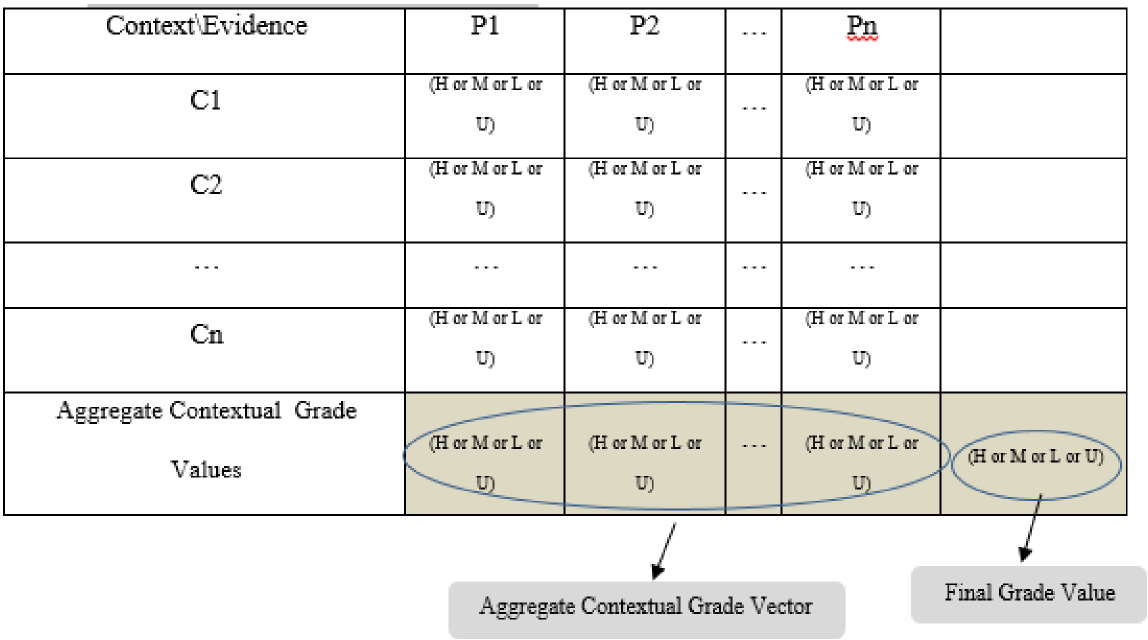
Table 5.2: Grade value population for an evidence with respect to contexts

Context/Evidence	P1	P2		Pn
C1	(H or M or L or U)	(H or M or L or U)		(H or M or L or U)
C2	(H or M or L or U)	(H or M or L or U)		(H or M or L or U)
Cn	(H or M or L or U)	(H or M or L or U)		(H or M or L or U)

5.2.1 Context aggregation

Based on the grade values, the aggregate contextual grade values are inferred from each column of Figure 5.8. The aggregate contextual grade values accumulatively makes the aggregate contextual vector. Figure 5.8 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 Eq. 5.1. For instance, between L and U, the aggregate value is assigned as L because L > U as in Eq. 5.1.

$$H > M > L > U \tag{5.1}$$



Context\Evidence	P1	P2	...	Pn	
C1	(H or M or L or U)	(H or M or L or U)	...	(H or M or L or U)	
C2	(H or M or L or U)	(H or M or L or U)	...	(H or M or L or U)	
...	...	...	...	...	
Cn	(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

Figure 5.8: Aggregate Contextual Grade Values and Vector

Final grade value (FGV) is inferred from the values of ACGV on the same rule as in Eq. 5.1. For the user explanation the FGV value is interpreted according to Eq. 5.2.

$$F(FGV) = \begin{matrix} if\ H \ \rightarrow Highly\ Beneficial \\ if\ M \rightarrow Moderate\ Beneficial \\ if\ L \rightarrow Less\ Beneficial \\ if\ U \rightarrow Unknown \end{matrix} \tag{5.2}$$

Chapter 6

Experimental Results and Evaluations

In order to evaluate the efficacy of the proposed solutions, different experiments are performed at different levels. Mainly the experiment is divided into two parts: offline and online experiments.

6.1 Offline Experiments

In the offline setup, first we collect documents on the basis of document identifiers in the training dataset. Second, documents are parsed to engineer features for constructing the training corpus. Third, experiments are performed on different combinations of features with different machine learning methods. Finally, the classification model is selected on the basis of classifier performance.

***Step 1**: Dataset Selection*

A subset of the collected data (relevant to treatment) is used for the offline experiments, with a training set that consists of 5,682 documents (1,683 rigorous and 3,999 non-rigorous). In the development test dataset, a total of 1,300 documents are included, 299 rigorous and 1,001 non-rigorous.

***Step 2**: Document downloading and parsing*

Algorithm 4 is used to download and parse the documents for the selected dataset. Four features are prepared from the parsed documents: title, abstract, MeSH, and publication type.

***Step 3**: Preprocessing*

Based on the procedures described in section 4.2.3, all of the features are processed to create feature vectors. For 5682 documents, a total of 9196 regular attributes are generated for the data feature vector, and 5468 regular attributes are generated for the metadata feature vector.

***Step 4**: Machine learning method performance using the development test dataset*

The experiments are performed using the RapidMiner tool and its text processing package [115]. Accuracies of different classifiers are presented in Table 6.1. The results on the development test set using 10-fold cross validation are evaluated on the four criteria of recall, precision, F1 measure, and accuracy using Eqs. 6.1, 6.2, 6.3, and 6.4, respectively.

$$Recall = TP/((TP + FN)) \tag{6.1}$$

$$Precision = TP/((TP + FP)) \tag{6.2}$$

$$F - Measure = 2 * (Recall * Precision)/(Recall + Precision) \tag{6.3}$$

$$Accuracy = (TP + TN)/((TP + TN + FP + FN)) \tag{6.4}$$

Where, $TP = TruePositive, FP = FalsePositive, FN = FalseNegative, andTN = TrueNegative$

Table 6.1: Experimental results for the development test set in recognizing scientifically rigorous studies using QRM

Feature Vector	Machine Learning	Recall (%)	Precision (%)	F1 Score (%)	Accuracy (%)
Data	Nave Bayes (Kernel)	39.46	18.61	25.29	46.38
	kNN (k = 5)	2.68	47.06	5.07	76.92
	SVM (C= 0.0)	21.4	33.68	26.17	72.23
	Decision Tree	14.05	4.33	6.62	8.85
Data + Metadata (Publication Type)	Nave Bayes (Kernel)	70.57	50.36	58.78	77.23
	kNN (k = 5)	62.21	67.99	64.97	84.69
	SVM (C= 0.0)	30.1	47.37	36.81	76.38
	Decision Tree	71.24	74.74	72.95	87.85
Data + Metadata (MeSH Terms)	Nave Bayes (Kernel)	41.47	24.17	30.54	56.62
	kNN (k = 5)	27.42	46.59	34.52	76.08
	SVM (C= 0.0)	32.11	51.06	39.43	77.31
	Decision Tree	14.38	4.46	6.81	9.46
Data + Metadata (Publication Type and MeSH Terms)	Nave Bayes (Kernel)	37.67	40.96	39.25	65.45
	kNN (k = 5)	33.78	51.53	40.81	77.46
	SVM (C= 0.0)	39.13	63.59	48.45	80.85
	Decision Tree	69.57	73.76	71.60	87.31

***Step 5**: Performance analysis of the models for quality appraisal* Each learning method was tested on four feature vectors: data, data + metadata (publication type), data + metadata (MeSH terms), and data + metadata (publication type and MeSH terms). The results in Table 1 show a distinct line between the data feature and data + metadata feature, as the latter outperformed the former in almost all evaluation criteria. Decision Tree performed well on the data + metadata (publication type) feature and data + metadata (publication type and MeSH terms) feature with 87.85% and 87.31% accuracies, respectively; however, on data and data + metadata (MeSH term) features, it showed equally poor performance. The linear SVM with a complexity constant value of 0.0 showed stable results across all the features, and it performed the best for the data + metadata (publication type and MeSH term) feature, with an accuracy of 80.85%. kNN showed better performance (76.92% accuracy) on the data feature, and it also performed well with the precision metric for all features, but it showed extremely low recall on the data feature. Nave Bayes showed better results in recall, with the highest value of 39.46% on the data feature, but showed lower precision and accuracy metrics. Based on the better performance on the development test dataset, we carried out an online experiment using DT.

6.1.1 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 dataset. Overall 249 different variations are found in publication types as shown in Figure 6.1 (a). Using algorithm 1, we normalized the 249 variations into 13 standard publication types having different frequencies as shown in Figure 6.1 (b). We experimented the performance of quality recognition model (QRM) on 5682 documents on publication type both in default and standard form. Journal article, randomized controlled trials, 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 6.2. QRM performed exceptionally on standard publication type. The recall value showed about 2%, precision about 40%, and accuracy about 24% increase in the standardized form.

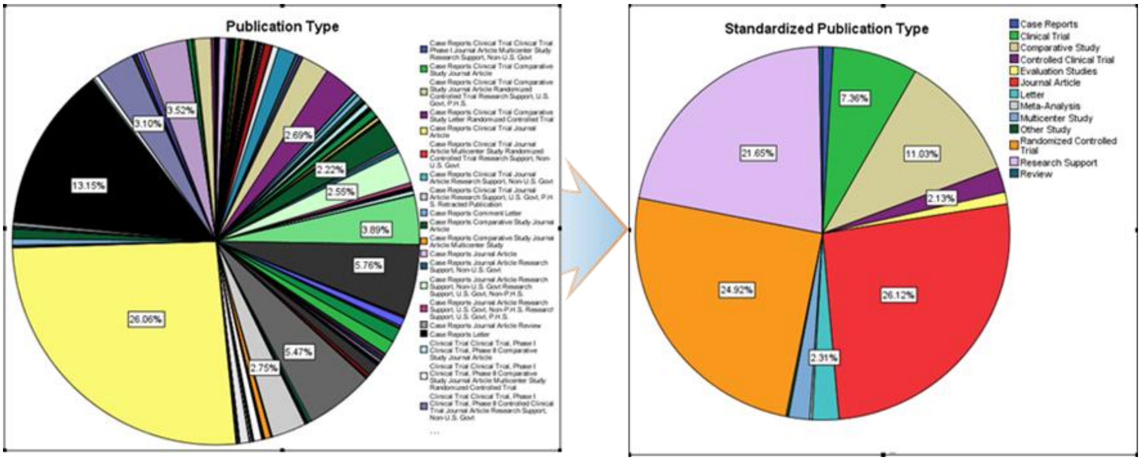


Figure 6.1: Publication types (a) and standardized publication types (b)l

Table 6.2: 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

6.1.2 CAG results for physician interested in treatment case study

The QRM model predicted 1355 out of 5682 documents as Rigor. Using Eq. (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 r6.3, Out of 1355 documents, about 60% documents are graded as H which means highly beneficial for the physician to benefit in treatment related clinical decisions. Other approximately 20% are graded as M (moderate beneficial), 8% as L (low beneficial), and 13% as U (unknown).

Table 6.3: 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%)

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.

6.2 Online Experiments

Level 1: Primary Queries (PQ) that are formulated from the rules. This is considered to be the baseline. Level 2: Augmented Queries (AQ) that are augmented with the clinical task or purpose filter. Level 3: Quality Recognition Model (QRM), which is a prediction model that filters out the non-quality studies. We first run all queries in the PubMed database using the proposed searching and downloading methods and record the results as baseline results. The same queries are repeated and augmented with the query type filter, and the results are recorded. The results of the augmented queries are passed through the QRM model, and the results of correctly predicted studies are recorded. We evaluated the results on the following four metrics: We evaluated the results on the following four metrics:

- Precision of 10 retrieved documents (P10) calculates the fraction of relevant and quality articles in the top 10 results using similar definitions to those used in [22].

$$P10 = a/(a+b); \tag{6.5}$$

where, a = true positives, articles found by the search term meet the criteria for treatment and methodological rigor (quality).
b = false positives, articles found by the search term do not meet the criteria for treatment and methodological rigor (quality). f there are fewer than 10 results, they are scaled to 10 using the scale formula in Eq. 6.6.

$$P10 = ((Precision * 10))/10; \tag{6.6}$$

- Mean Precision (MP) is the average precision of all queries.
- Total Document Reciprocal Rank (TDRR) is the sum of reciprocal ranks of all relevant-quality articles for a query.
  For example, if relevant and quality articles were found at ranks 3 and 8, the TDRR would be $1/3 + 1/8 = 0.45$. Mean Reciprocal Rank (MRR) measures the average of the TDRR for all queries.

The comparative results are displayed in Figure 6.2. In most cases, QRM results are better than those AQ and PQ. However, for some cases, AQ performance is equivalent to or better than that of QRM.

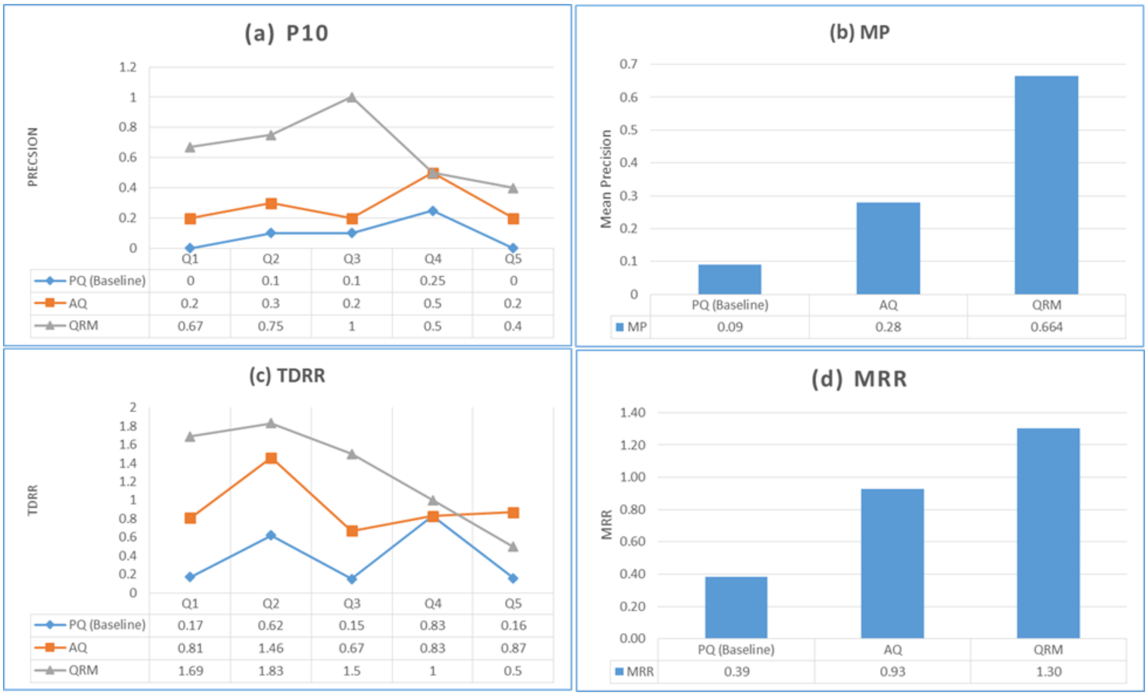


Figure 6.2: Performance of primary queries (baseline), augmented queries (AQ), and quality recognition model

We summarize the overall improvements of QRM and AQ over PQ (baseline) in the following points with reference to Figures 6.2 (a), (b), (c), and (d).

- Precision at 10 retrieved documents: the performance of QRM is better than that of PQ (baseline) and AQ for all except the fourth query, where its performance is equivalent to AQ. AQ performance is better than that of PQ (baseline) for all queries.

- Mean precision of QRM is found to be 137.14% and 637.78% improved compared to AQ and PQ (baseline), respectively, while AQ showed 211.11% better performance than PQ.

- Total document reciprocal rank for QRM is better than that for all queries in comparison to PQ; however, it performed poorly for the fifth query compared to AQ. PQ (baseline) performance was poor in all except the fourth query, where it showed equivalent performance to

AQ.

- Mean reciprocal rank value of QRM is 40.52% and 237.82% higher than those of AQ and PQ, respectively. Similarly, AQ showed 140.41% improved performance over PQ.

6.3 Comparison with Existing Approaches

In a very recent work [1], authors performed experiment with three features (Title, Abstract, Publication Type). We repeat the same method on our dataset and compare the results, we obtained approximately 4% better accuracy as shown in Table 6.4.

Table 6.4: Comparison with Sarker et al. [1]

System	**Accuracy**
Existing	76.38 %
Proposed System	80.85%

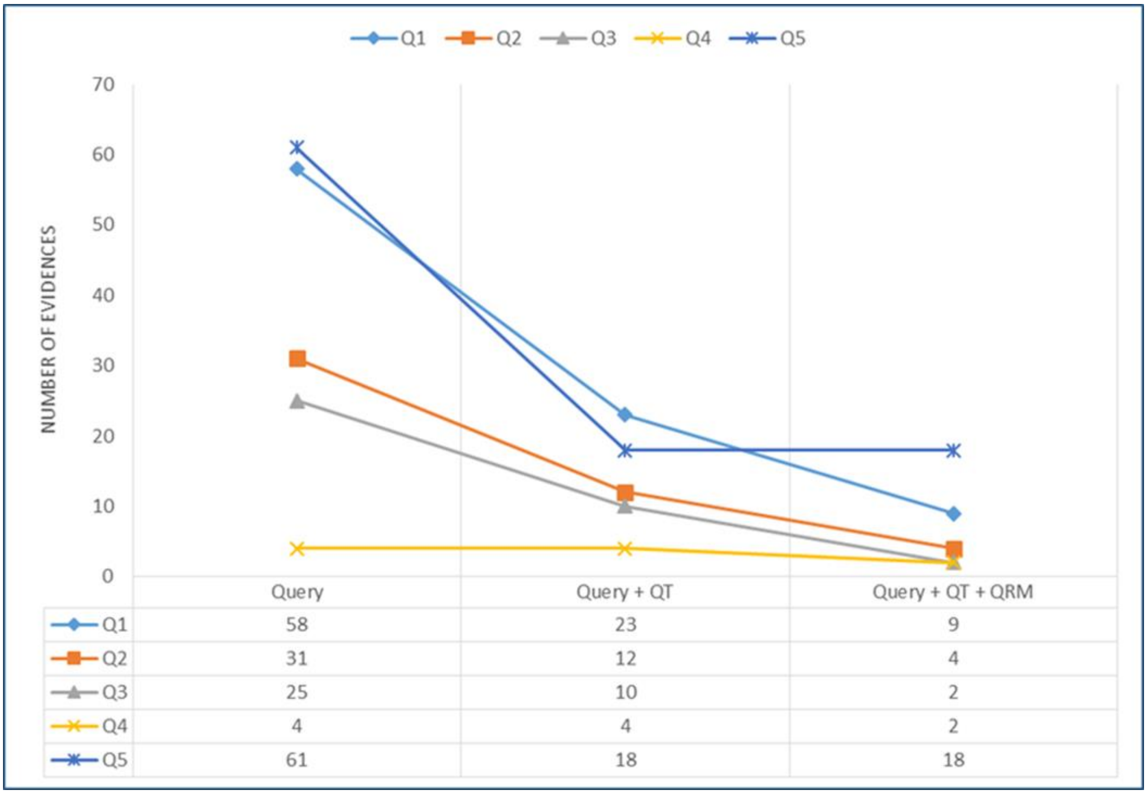
Previously authors in [2] performed experiment with different set of features (Title, Abstract, MeSH, Publication Type, entity, relationships). Data from the same collection using four features (Title, Abstract, MeSH, and Publication Type), we obtained about 5% better F-Score as shown in Table 6.5.

Table 6.5: Comparison with Kilicoglu et al. [2]

System	**F-Measure**
Existing	65.90 %
Proposed System	71.60%

6.4 Result evaluation for record reduction

When one relies on user queries without consideration of clinical task (CT) or quality recognition model (QRM), an overwhelming number of records from the literature are retrieved, which makes it hard for the busy clinicians to find the best evidence in a short period of time. The query type filter reduced the number of records, making it more concise. However, applying the QRM model reduced the set to its final level, consisting of more precise, relevant, high-quality evidence. In Figure 6.3, we present the retrieval results for the first five queries with a step-by-step reduction of the original query, applying QT and QRM filters. On average, 51% of the records are eliminated when the CT filter is applied. Further, 48% records are eliminated when QRM is applied. Overall, 75% of the records (on average) are filtered from the original query result by the application of QT and QRM filters.

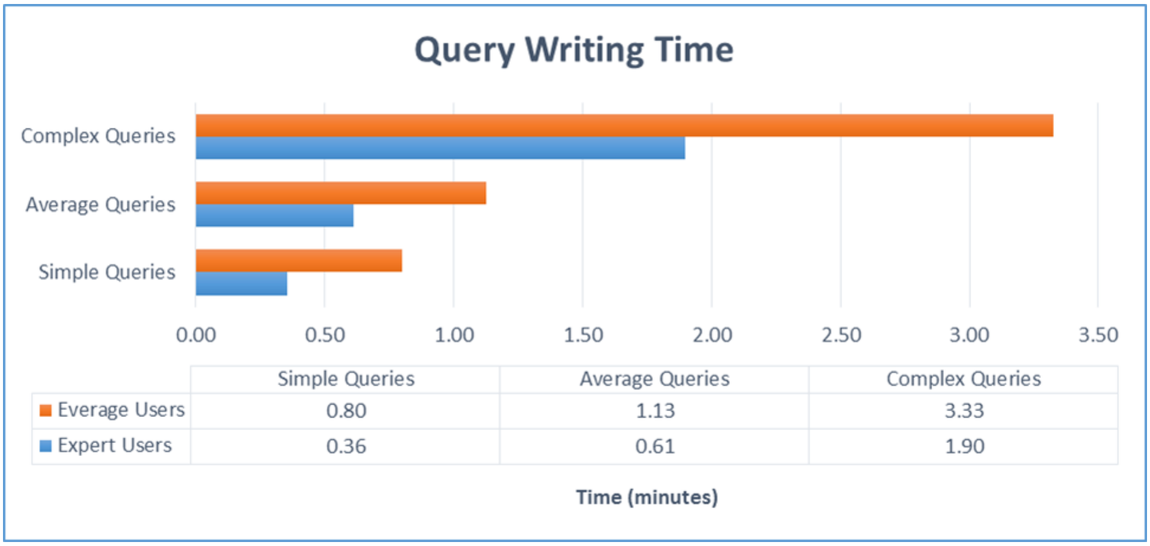


	Query	Query + QT	Query + QT + QRM
Q1	58	23	9
Q2	31	12	4
Q3	25	10	2
Q4	4	4	2
Q5	61	18	18

Figure 6.3: Record reduction results for primary query, clinical task augmented query, and quality recognition model

6.5 Result evaluation for query writing time

Automatic query construction saved considerable time for the clinicians. Queries written manually consumed much of the time of busy clinicians. To reflect the actual time of query writing, we categorized the queries into three types based on the complexity level in terms of length. The three types included simple queries (consisting of 4 or fewer terms), average queries (consisting of between 4 and 8 terms exclusive), and complex queries (consisting of 8 or more terms). The experiment was performed on two types of users: expert users (who had good domain knowledge and excellent computer typing skills) and average users (who had an average level of domain expertise and fair computer typing skills). Both of them used PubMed search browsers to type the queries. The query time for each user was recorded, and the compiled results in terms of query writing time are shown in Figure 6.4. During the experiment, we ignored the mistakes made while writing.

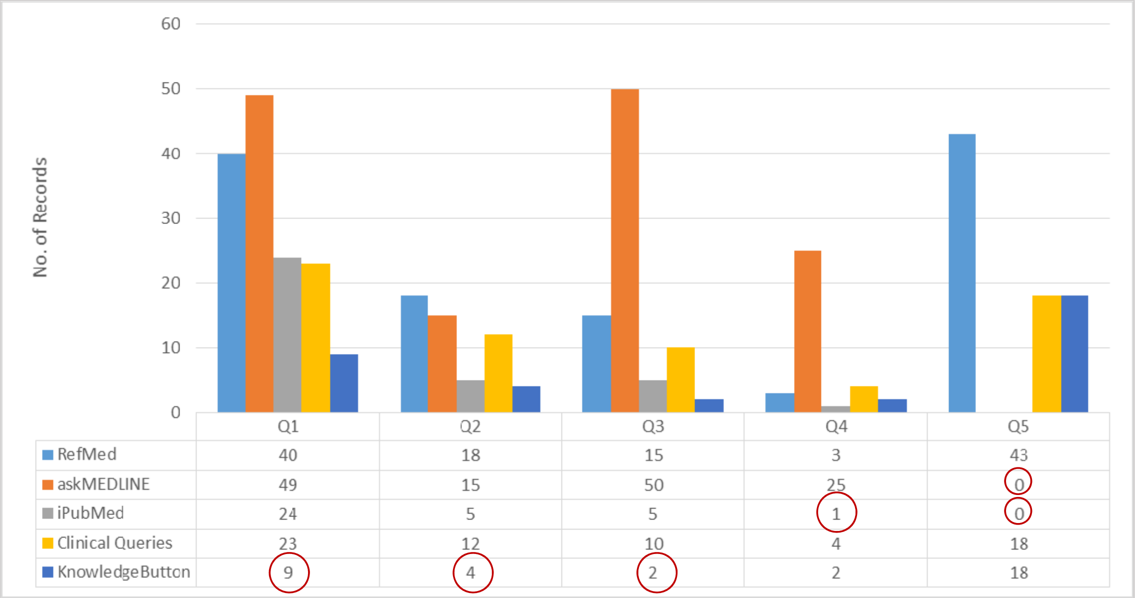


	Simple Queries	Average Queries	Complex Queries
🟧 Everage Users	0.80	1.13	3.33
🟦 Expert Users	0.36	0.61	1.90

Figure 6.4: Query writing time by an expert and an average-expert users

6.6 Comparison with PubMed derivative systems

There are overall 28 systems are reported that are developed as derivatives of PubMed [86]. Among these derivatives, four systems have some resemblance to the proposed approach. These four systems are, RedMed [87], askMEDLINE [88], iPubMed [89], and Clinical Queries [116]. The selected five queries are run on all these four derivative systems and recorded the results as shown in Figure 6.5.



	Q1	Q2	Q3	Q4	Q5
RefMed	40	18	15	3	43
askMEDLINE	49	15	50	25	0
iPubMed	24	5	5	1	0
Clinical Queries	23	12	10	4	18
KnowledgeButton	9	4	2	2	18

Figure 6.5: Comparison with PubMed derivatives

6.7 Discussion

Based on the experimental results in this study, we infer some important observations to consider while working in the area of evidence-based clinical practice and decision support.

6.7.1 Automatic query formulation

Currently, most of the studies in the area of automating the process of initial query generation use the electronic medical records (EMR) of patients [19–21]. The obvious limitation with queries generated from EMR records is the lack of contextual information such as the query type filter. Generating queries from the rules of a CDSS provides the mean in the form of an action part to find the type of the query. Secondly, the patient record by itself does not provide enough information to automatically fill some parts of the PICO format. For instance, the I (intervention) part is the most important part of PICO, which cannot be formulated from a patient record. However, in CDSS, it can be automatically formulated from the action part of a rule. Thirdly, automated query formulation saves considerable time for the clinicians compared to manually writing queries.

6.7.2 Automatic Appraisal

Critical appraisal is required to filter out the studies of low quality. There are two possibilities to achieve the appraisal objective: manually through domain experts or automatically through a system. Involving domain experts in the first phase of the appraisal is highly time consuming, which we avoided in this study through machine learning approaches. For machine learning models, the great challenge is related to the selection of training data and automated preparation of features. For the first challenge, i.e., selection of training data, we acquired a collection of MEDLINE documents created by highly qualified specialists for the purpose of finding high-quality articles [22]. For the second challenge, i.e., automated preparation of features, we programmed the automation of the retrieval process of data and metadata features for offline and online experiments.

6.7.3 Feature Significance

During evaluations, we noticed that publication type was the most influential feature contributing to determining the quality of an article. This publication type feature had the highest accuracy level among the pool of evaluated features. In addition to publication type, the metadata feature MeSH terms also produced good results using the machine learning algorithm. Combining both publication type and MeSH term features with data features produced the best and stable results across the majority of the machine learning algorithms.

6.7.4 Automation: efficiency and accuracy

It is hard to describe all sorts of automation quantitatively through experiments. However, it is important to highlight the benefits of automation with respect to clinician performance and content accuracy. Recalling the methodology proposed in this paper, we study automation at four levels: initial query construction from CDSS KB/patient record, query augmentation with query type and MeSH translations, automated preparation of data and meta-data features for corpus preparation, and finally automated appraisal of the articles retrieved with queries. This automation adds to the overall efficiency in term of time and effort. At the same time, an automated decision by the system cannot always be used as a final decision. The final approval of evidence by a human is crucial to avoid the entry of wrong evidence that might affect current and future clinical decisions into the system.

6.7.5 Limitation of the work

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 a survey is conducted 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.

Chapter 7

Conclusion and Future Directions

7.1 Conclusion

Obtaining high-quality evidence from a large volume of diverse literature is an important task in clinical care. Automation to improve the query formulation and appraisal process is required for clinical efficiency. The main goal of this thesis was to minimize user involvement in the research evidence acquisition and appraisal because manually doing the same takes precious time of their clinical schedule. This thesis achieved the aimed automation at the evidence acquisition stage by automatically constructing the query from the knowledge in the context and augmenting the query with a clinical task for more precise results. At evidence appraisal stage, the automation is achieved in automatically engineering the data and metadata features and aggregation of user and article context. At the same time, the proposed methods kept the accuracy level high because of the underline mappings are based on a well-formulated question framed in PICO format. Similarly, the selection of a reliable dataset which is annotated by a team of professional experts and the proper selection of machine learning method increase the level high of the quality assessment process. The presented methodology has the potential to produce several benefits in term of practical implementation in evidence-based medical practice.

- Clinician time saving on query construction and quality assessment.
- Contextual fitness of the evidence to the user clinical environment
- Improved patient care with availability of high quality research evidence
- Increase confidence level of physicians in clinical decisions
- Corpus preparation and availability for the future experiments

7.2 Future Work

We plan to provide an interface to clinicians for approving the system-appraised evidence. The approved evidence will be integrated with the knowledge base of CDSS. The approval of evidence will be a step toward generating domain-specific training data that is relevant and high quality, acquired from the methods presented in this study. There are three potential directions this work shall be extended.

- Rule mining from the quality studies obtained with the proposed approach.
- Investigating additional non-textual features such as image and figure information in order to increase the accuracy level of quality assessments.
- Extending the current towards precision medicine to assist the practitioners in practicing the precision evidence-based medicine.

7.2.1 Future work: rule mining

One of the potential areas that can be further extended is to mine rules from the recognized quality evidences. This will need another level of feature selection which can facilitate the process of rules learning in the form of conditions and actions. PICO query may of help to recognize such features where P can contribute to condition part of the rules and I part can contribute to the action part of the rule with possible alternative actions recognized with C. Rule mining from currently available quality evidentiary documents will keep the knowledge base more up-to-date and rightly evolved with time.

7.2.2 Future work: Non-textual features

In current work, the quality of the contents is assessed on textual features without considering image contents. Figures and images in an article carry an important piece of information. Adding non-textual features in the feature set for assessing quality may increase the chance of accuracy. Also, they are a good source for the quick presentation of the complex contents.

7.2.3 Future work: Precision Medicine

Precision medicine is an emerging approach to disease prevention and treatment that takes into account people's individual variations in genes, environment, and lifestyle. It is a hot topic very recently emerged around the world. Different countries start working to contribute to PM initiative. The proposed work can be further extended to identify cohort specific studies in the first part and then extract both textual and non-textual contents for strengthening the final PMI decision making process.

Bibliography

[1] A. Sarker, D. Mollá, and C. Paris, "Automatic evidence quality prediction to support evidence-based decision making," *Artificial intelligence in medicine*, vol. 64, no. 2, pp. 89–103, 2015.

[2] H. Kilicoglu, D. Demner-Fushman, T. C. Rindflesch, N. L. Wilczynski, and R. B. Haynes, "Towards automatic recognition of scientifically rigorous clinical research evidence," *Journal of the American Medical Informatics Association*, vol. 16, no. 1, pp. 25–31, 2009.

[3] D. L. Sackett, W. M. Rosenberg, J. M. Gray, R. B. Haynes, and W. S. Richardson, "Evidence based medicine: what it is and what it isn't," *Bmj*, vol. 312, no. 7023, pp. 71–72, 1996.

[4] S. E. Straus, J. M. Tetroe, and I. D. Graham, "Knowledge translation is the use of knowledge in health care decision making," *Journal of clinical epidemiology*, vol. 64, no. 1, pp. 6–10, 2011.

[5] R. B. Haynes, R. S. Hayward, and J. Lomas, "Bridges between health care research evidence and clinical practice," *Journal of the American Medical Informatics Association*, vol. 2, no. 6, pp. 342–350, 1995.

[6] W. S. Richardson, "We should overcome the barriers to evidence-based clinical diagnosis!" *Journal of clinical epidemiology*, vol. 60, no. 3, pp. 217–227, 2007.

[7] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, and C. F. Aliferis, "Text categorization models for high-quality article retrieval in internal medicine," *Journal of the American Medical Informatics Association*, vol. 12, no. 2, pp. 207–216, 2005.

[8] J. Houser, K. S. Oman *et al.*, *Evidence-based practice: An implementation guide for healthcare organizations*. Jones & Bartlett Publishers, 2010.

[9] B. M. Melnyk and E. Fineout-Overholt, *Evidence-based practice in nursing & healthcare: A guide to best practice*. Lippincott Williams & Wilkins, 2011.

[10] M. Dawes, P. Davies, A. Gray, J. Mant, K. Seers, and R. Snowball, *Evidence-based practice*. Emap Public Sector Management, 2000.

[11] S. Sackett and R. Richardson, "Evidence-based practice," *Found Evidence-Based Social Work Pract*, vol. 35, 2006.

[12] A. X. Garg, N. K. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. Devereaux, J. Beyene, J. Sam, and R. B. Haynes, "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review," *Jama*, vol. 293, no. 10, pp. 1223–1238, 2005.

[13] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems," in *Biomedical informatics*. Springer, 2014, pp. 643–674.

[14] R. A. Greenes, *Clinical decision support: the road ahead*. Academic Press, 2011.

[15] G. M. Leung, "Evidence-based practice revisited," *Asia-Pacific Journal of Public Health*, vol. 13, no. 2, pp. 116–121, 2001.

[16] R. B. Haynes, D. L. Sackett, J. M. Gray, D. L. Cook, and G. H. Guyatt, "Transferring evidence from research into practice: 2. getting the evidence straight," *Evidence Based Medicine*, vol. 2, no. 1, pp. 4–6, 1997.

[17] J. M. Satterfield, B. Spring, R. C. Brownson, E. J. Mullen, R. P. Newhouse, B. B. Walker, and E. P. Whitlock, "Toward a transdisciplinary model of evidence-based practice," *Milbank Quarterly*, vol. 87, no. 2, pp. 368–390, 2009.

[18] J. D. Eldredge, "The evolution of evidence based library and information practice, part i: Defining eblip," *Evidence Based Library and Information Practice*, vol. 7, no. 4, pp. 139–145, 2012.

[19] J. J. Cimino, “An integrated approach to computer-based decision support at the point of care,” *TRANSACTIONS-AMERICAN CLINICAL AND CLIMATOLOGICAL ASSOCIATION*, vol. 118, p. 273, 2007.

[20] S. A. Fowler, L. H. Yaeger, F. Yu, D. Doerhoff, P. Schoening, and B. Kelly, “Electronic health record: integrating evidence-based information at the point of clinical decision making,” *Journal of the Medical Library Association: JMLA*, vol. 102, no. 1, p. 52, 2014.

[21] D. Perez-Rey, A. Jimenez-Castellanos, M. Garcia-Remesal, J. Crespo, and V. Maojo, “Cdapubmed: a browser extension to retrieve ehr-based biomedical literature,” *BMC medical informatics and decision making*, vol. 12, no. 1, p. 1, 2012.

[22] N. L. Wilczynski, D. Morgan, and R. B. Haynes, “An overview of the design and methods for retrieving high-quality studies for clinical care,” *BMC Medical Informatics and Decision Making*, vol. 5, no. 1, p. 1, 2005.

[23] J. W. Ely, J. A. Osheroff, M. H. Ebell, G. R. Bergus, B. T. Levy, M. L. Chambliss, and E. R. Evans, “Analysis of questions asked by family doctors regarding patient care,” *Bmj*, vol. 319, no. 7206, pp. 358–361, 1999.

[24] N. C. R. T. I. E. based Practice Center North Carolina Central University (Durham and S. West, *Systems to rate the strength of scientific evidence*. AHRQ (Agency for Healthcare Research and Quality), 2002.

[25] G. H. Guyatt, A. D. Oxman, G. E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, and H. J. Schünemann, “Grade: an emerging consensus on rating quality of evidence and strength of recommendations.” *BMJ (Clinical research ed.)*, 2008.

[26] H. J. Schünemann, A. D. Oxman, J. Brozek, P. Glasziou, R. Jaeschke, G. E. Vist, J. W. Williams, R. Kunz, J. Craig, V. M. Montori *et al.*, “Grading quality of evidence and strength of recommendations for diagnostic tests and strategies,” *Bmj*, vol. 336, no. 7653, pp. 1106–1110, 2008.

[27] M. H. Ebell, J. Siwek, B. D. Weiss, S. H. Woolf, J. Susman, B. Ewigman, and M. Bowman, "Strength of recommendation taxonomy (sort): a patient-centered approach to grading evidence in the medical literature," *The Journal of the American Board of Family Practice*, vol. 17, no. 1, pp. 59–67, 2004.

[28] R. B. Haynes and N. L. Wilczynski, "Optimal search strategies for retrieving scientifically strong studies of diagnosis from medline: analytical survey," *Bmj*, vol. 328, no. 7447, p. 1040, 2004.

[29] F. Ruiz-Rico, J.-L. Vicedo, and M.-C. Rubio-Sánchez, "Medline abstracts classification based on noun phrases extraction," in *International Joint Conference on Biomedical Engineering Systems and Technologies*. Springer, 2008, pp. 507–519.

[30] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," *BMC bioinformatics*, vol. 12, no. 2, p. 1, 2011.

[31] S. Kim and J. Choi, "An svm-based high-quality article classifier for systematic reviews," *Journal of biomedical informatics*, vol. 47, pp. 153–159, 2014.

[32] T. Tang, D. Hawking, R. Sankaranarayana, K. M. Griffiths, and N. Craswell, "Quality-oriented search for depression portals," in *European Conference on Information Retrieval*. Springer, 2009, pp. 637–644.

[33] G. Guyatt, J. Cairns, D. Churchill, D. Cook, B. Haynes, J. Hirsh, J. Irvine, M. Levine, M. Levine, J. Nishikawa *et al.*, "Evidence-based medicine: a new approach to teaching the practice of medicine," *Jama*, vol. 268, no. 17, pp. 2420–2425, 1992.

[34] M. Bhandari and P. V. Giannoudis, "Evidence-based medicine: what it is and what it is not," *Injury*, vol. 37, no. 4, pp. 302–306, 2006.

[35] E. S. Berner, *Clinical decision support systems*. Springer, 2007.

[36] S. Tabet, G. Wagner, S. Spreeuwenberg, P. D. Vincent, and G. Jacques, "Omg production rule representation-context and current status," 2005.

[37] R. J. Brachman, H. J. Levesque, and R. Reiter, *Knowledge representation.* MIT press, 1992.

[38] P. A. De Clercq, J. A. Blom, H. H. Korsten, and A. Hasman, "Approaches for creating computer-interpretable guidelines that facilitate decision support," *Artificial intelligence in medicine*, vol. 31, no. 1, pp. 1–27, 2004.

[39] A. Wright and D. F. Sittig, "A framework and model for evaluating clinical decision support architectures," *Journal of biomedical informatics*, vol. 41, no. 6, pp. 982–990, 2008.

[40] L. Ohno-Machado, J. H. Gennari, S. N. Murphy, N. L. Jain, S. W. Tu, D. E. Oliver, E. Pattison-Gordon, R. A. Greenes, E. H. Shortliffe, and G. O. Barnett, "The guideline interchange format," *Journal of the American Medical Informatics Association*, vol. 5, no. 4, pp. 357–372, 1998.

[41] A. A. Boxwala, M. Peleg, S. Tu, O. Ogunyemi, Q. T. Zeng, D. Wang, V. L. Patel, R. A. Greenes, and E. H. Shortliffe, "Glif3: a representation format for sharable computer-interpretable clinical practice guidelines," *Journal of biomedical informatics*, vol. 37, no. 3, pp. 147–161, 2004.

[42] K. Donnelly, "Snomed-ct: The advanced terminology and coding system for ehealth," *Studies in health technology and informatics*, vol. 121, p. 279, 2006.

[43] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang, "Snomed clinical terms: overview of the development process and project status." in *Proceedings of the AMIA Symposium.* American Medical Informatics Association, 2001, p. 662.

[44] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, "The unified medical language system." *Methods of information in medicine*, vol. 32, no. 4, pp. 281–291, 1993.

[45] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.

[46] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, and G. O. Barnett, "The unified medical language system," *Journal of the American Medical Informatics Association*, vol. 5, no. 1, pp. 1–11, 1998.

[47] S. J. Nelson, W. D. Johnston, and B. L. Humphreys, "Relationships in medical subject headings (mesh)," in *Relationships in the Organization of Knowledge*. Springer, 2001, pp. 171–184.

[48] M. H. Coletti and H. L. Bleich, "Medical subject headings used to search the biomedical literature," *Journal of the American Medical Informatics Association*, vol. 8, no. 4, pp. 317–323, 2001.

[49] C. E. Lipscomb, "Medical subject headings (mesh)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[50] A. A. Chang, K. M. Heskett, and T. M. Davidson, "Searching the literature using medical subject headings versus text word with pubmed," *The Laryngoscope*, vol. 116, no. 2, pp. 336–340, 2006.

[51] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys (CSUR)*, vol. 44, no. 1, p. 1, 2012.

[52] S. Harabagiu, D. Moldovan, M. Paşca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morrescu, "The role of lexico-semantic feedback in open-domain textual question-answering," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2001, pp. 282–289.

[53] Z. Gong, C. W. Cheang *et al.*, "Multi-term web query expansion using wordnet," in *International Conference on Database and Expert Systems Applications*. Springer, 2006, pp. 379–388.

[54] N. Schlaefer, J. Ko, J. Betteridge, M. A. Pathak, E. Nyberg, and G. Sautter, "Semantic extensions of the ephyra qa system for trec 2007." in *TREC*, vol. 1, no. 2, 2007, p. 2.

[55] A. Singhal and F. Pereira, “Document expansion for speech retrieval,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 1999, pp. 34–41.

[56] M. L. Kherfi and D. Ziou, “Image retrieval based on feature weighting and relevance feedback,” in *Image Processing, 2004. ICIP’04. 2004 International Conference on*, vol. 1. IEEE, 2004, pp. 689–692.

[57] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan, “Semantic concept-based query expansion and re-ranking for multimedia retrieval,” in *Proceedings of the 15th ACM international conference on Multimedia.* ACM, 2007, pp. 991–1000.

[58] J. Arguello, J. L. Elsas, J. Callan, and J. G. Carbonell, “Document representation and query expansion models for blog recommendation.” *ICWSM*, vol. 2008, no. 0, p. 1, 2008.

[59] N. J. Belkin and W. B. Croft, “Information filtering and information retrieval: Two sides of the same coin?” *Communications of the ACM*, vol. 35, no. 12, pp. 29–38, 1992.

[60] P. McNamee and J. Mayfield, “Comparing cross-language query expansion techniques by degrading translation resources,” in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2002, pp. 159–166.

[61] W. Kraaij, J.-Y. Nie, and M. Simard, “Embedding web-based statistical translation models in cross-language information retrieval,” *Computational Linguistics*, vol. 29, no. 3, pp. 381–419, 2003.

[62] L. Ballesteros and W. B. Croft, “Phrasal translation and query expansion techniques for cross-language information retrieval,” in *ACM SIGIR Forum*, vol. 31, no. SI. ACM, 1997, pp. 84–91.

[63] J. Graupmann, J. Cai, and R. Schenkel, “Automatic query refinement using mined semantic relations,” in *International Workshop on Challenges in Web Information Retrieval and Integration.* IEEE, 2005, pp. 205–213.

[64] E. Agichtein, S. Lawrence, and L. Gravano, “Learning to find answers to questions on the web,” *ACM Transactions on Internet Technology (TOIT)*, vol. 4, no. 2, pp. 129–162, 2004.

[65] K. Church and B. Smyth, “Mobile content enrichment,” in *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 2007, pp. 112–121.

[66] R. Kraft and J. Zien, “Mining anchor text for query refinement,” in *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004, pp. 666–674.

[67] J. Xu and W. B. Croft, “Query expansion using local and global document analysis,” in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 4–11.

[68] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, “Introduction to wordnet: An on-line lexical database,” *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[69] J. Bai, D. Song, P. Bruza, J.-Y. Nie, and G. Cao, “Query expansion using term relationships in language models for information retrieval,” in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 688–695.

[70] G. Salton and C. Buckley, “Improving retrieval performance by relevance feedback,” *Readings in information retrieval*, vol. 24, no. 5, pp. 355–363, 1997.

[71] C. Carpineto, R. De Mori, G. Romano, and B. Bigi, “An information-theoretic approach to automatic query expansion,” *ACM Transactions on Information Systems (TOIS)*, vol. 19, no. 1, pp. 1–27, 2001.

[72] C. Zhai and J. Lafferty, “Model-based feedback in the language modeling approach to information retrieval,” in *Proceedings of the tenth international conference on Information and knowledge management*. ACM, 2001, pp. 403–410.

[73] V. Lavrenko and W. B. Croft, “Relevance based language models,” in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 120–127.

[74] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2008, pp. 243–250.

[75] S. Cronen-Townsend and W. B. Croft, "Quantifying query ambiguity," in *Proceedings of the second international conference on Human Language Technology Research.* Morgan Kaufmann Publishers Inc., 2002, pp. 104–109.

[76] Y. Chang, I. Ounis, and M. Kim, "Query reformulation using automatically generated query concepts from a document space," *Information processing & management*, vol. 42, no. 2, pp. 453–468, 2006.

[77] J. J. Rocchio, "Relevance feedback in information retrieval," 1971.

[78] K. Alexander, D. Hawking, R. Jones, T. Gedeon, and H. Greville, "Automated medical literature retrieval," *The Australasian medical journal*, vol. 5, no. 9, p. 489, 2012.

[79] S. L. Price, W. R. Hersh, D. D. Olson, and P. J. Embi, "Smartquery: context-sensitive links to medical knowledge sources from the electronic patient record." in *Proceedings of the AMIA Symposium.* American Medical Informatics Association, 2002, p. 627.

[80] J. J. Cimino, X. Jing, and G. Del Fiol, "Meeting the electronic health record meaningful use criterion for the hl7 infobutton standard using openinfobutton and the librarian infobutton tailoring environment (lite)," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 112.

[81] G. Del Fiol, V. Huser, H. R. Strasberg, S. M. Maviglia, C. Curtis, and J. J. Cimino, "Implementations of the hl7 context-aware knowledge retrieval (infobutton) standard: challenges, strengths, limitations, and uptake," *Journal of biomedical informatics*, vol. 45, no. 4, pp. 726–735, 2012.

[82] S. A. Collins, L. M. Currie, S. Bakken, and J. J. Cimino, "Information needs, infobutton manager use, and satisfaction by clinician type: a case study," *Journal of the American Medical Informatics Association*, vol. 16, no. 1, pp. 140–142, 2009.

[83] J. Cimino and G. Del Fiol, “Infobuttons and point of care access to knowledge,” *Clinical decision support-the road ahead*, pp. 345–372, 2007.

[84] J. J. Cimino, J. Li, S. Bakken, and V. L. Patel, “Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users.” in *Proceedings of the AMIA Symposium.* American Medical Informatics Association, 2002, p. 170.

[85] M. Guilherme Del Fiol, R. A. Rocha, and P. D. Clayton, “Infobuttons at intermountain healthcare: utilization and infrastructure,” 2006.

[86] Z. Lu, “Pubmed and beyond: a survey of web tools for searching biomedical literature,” *Database*, vol. 2011, p. baq036, 2011.

[87] H. Yu, T. Kim, J. Oh, I. Ko, and S. Kim, “Refmed: relevance feedback retrieval system fo pubmed,” in *Proceedings of the 18th ACM conference on Information and knowledge management.* ACM, 2009, pp. 2099–2100.

[88] P. Fontelo, F. Liu, and M. Ackerman, “ask medline: a free-text, natural language query tool for medline/pubmed,” *BMC Medical Informatics and Decision Making*, vol. 5, no. 1, p. 1, 2005.

[89] J. Wang, I. Cetindil, S. Ji, C. Li, X. Xie, G. Li, and J. Feng, “Interactive and fuzzy search: a dynamic way to explore medline,” *Bioinformatics*, vol. 26, no. 18, pp. 2321–2327, 2010.

[90] W. S. Richardson, M. C. Wilson, J. Nishikawa, R. S. Hayward *et al.*, “The well-built clinical question: a key to evidence-based decisions,” *Acp j club*, vol. 123, no. 3, pp. A12–3, 1995.

[91] R. Snowball, “Using the clinical question to teach search strategy: fostering transferable conceptual skills in user education by active learning,” *Health Libraries Review*, vol. 14, no. 3, pp. 167–172, 1997.

[92] E. V. Villanueva, E. A. Burrows, P. A. Fennessy, M. Rajendran, and J. N. Anderson, “Improving question formulation for use in evidence appraisal in a tertiary care setting: a ran-

domised controlled trial [isrctn66375463],” *BMC medical informatics and decision making*, vol. 1, no. 1, p. 1, 2001.

[93] C. Schardt, M. B. Adams, T. Owens, S. Keitz, and P. Fontelo, “Utilization of the pico framework to improve searching pubmed for clinical questions,” *BMC medical informatics and decision making*, vol. 7, no. 1, p. 1, 2007.

[94] N. L. Wilczynski, R. B. Haynes, H. Team *et al.*, “Robustness of empirical search strategies for clinical content in medline.” in *Proceedings of the AMIA Symposium.* American Medical Informatics Association, 2002, p. 904.

[95] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval, “Context-aware recommender systems for learning: a survey and future challenges,” *IEEE Transactions on Learning Technologies*, vol. 5, no. 4, pp. 318–335, 2012.

[96] J. Rycroft-Malone, “The parihs frameworka framework for guiding the implementation of evidence-based practice,” *Journal of nursing care quality*, vol. 19, no. 4, pp. 297–304, 2004.

[97] IHTSDO, “SNOMED CT Snapshot REST API,” http://docs.snomedctsnapshotapi.apiary.io/, 2016, [Online; accessed 07-Dec-2016].

[98] E. Sayers and D. Wheeler, “Building customized data pipelines using the entrez programming utilities (eutils),” 2004.

[99] M. Hussain, M. Afzal, T. Ali, R. Ali, W. A. Khan, A. Jamshed, S. Lee, B. H. Kang, and K. Latif, “Data-driven knowledge acquisition, validation, and transformation into hl7 arden syntax,” *Artificial intelligence in medicine*, 2015.

[100] PubMed, “Pubmed help,” 2015, [Online; accessed 01-July-2015]. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Clinical_Queries_Filters

[101] M. F. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[102] U. N. L. of Medicine, “Publication characteristics (publication types) with scope notes,” 2016, [Online; accessed 07-October-2016]. [Online]. Available: https://www.nlm.nih.gov/mesh/pubtypes.html

[103] A. D. Oxman, G. W. Group *et al.*, “Grading quality of evidence and strength of recommendations,” *Bmj*, vol. 328, no. 19, pp. 1490–4, 2004.

[104] P. M. Ho, P. N. Peterson, and F. A. Masoudi, “Evaluating the evidence is there a rigid hierarchy?” *Circulation*, vol. 118, no. 16, pp. 1675–1684, 2008.

[105] D. Evans, “Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions,” *Journal of clinical nursing*, vol. 12, no. 1, pp. 77–84, 2003.

[106] F. Sebastiani, “Machine learning in automated text categorization,” *ACM computing surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[107] Y. Yang, “An evaluation of statistical approaches to text categorization,” *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.

[108] D. D. Lewis, “Naive (bayes) at forty: The independence assumption in information retrieval,” in *European conference on machine learning*. Springer, 1998, pp. 4–15.

[109] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “Knn model-based approach in classification,” in *OTM Confederated International Conferences” On the Move to Meaningful Internet Systems”*. Springer, 2003, pp. 986–996.

[110] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in *European conference on machine learning*. Springer, 1998, pp. 137–142.

[111] C. Apté, F. Damerau, and S. M. Weiss, “Automated learning of decision rules for text categorization,” *ACM Transactions on Information Systems (TOIS)*, vol. 12, no. 3, pp. 233–251, 1994.

[112] C. X. Ling, J. Huang, and H. Zhang, “Auc: a statistically consistent and more discriminating measure than accuracy,” in *IJCAI*, vol. 3, 2003, pp. 519–524.

[113] J. Huang, J. Lu, and C. X. Ling, “Comparing naive bayes, decision trees, and svm with auc and accuracy,” in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 2003, pp. 553–556.

[114] M. J. Dobrow, V. Goel, and R. Upshur, “Evidence-based health policy: context and utilisation,” *Social science & medicine*, vol. 58, no. 1, pp. 207–217, 2004.

[115] M. Hofmann and R. Klinkenberg, *RapidMiner: Data mining use cases and business analytics applications*. CRC Press, 2013.

[116] R. B. Haynes and N. Wilczynski, “Finding the gold in medline: clinical queries,” *Evidence Based Medicine*, vol. 10, no. 4, pp. 101–102, 2005.

[117] M. Hussain, W. A. Khan, M. Afzal, and S. Lee, “Smart cdss for smart homes,” in *International Conference on Smart Homes and Health Telematics*. Springer, 2012, pp. 266–269.

[118] T. Ali, M. Hussain, W. A. Khan, M. Afzal, and S. Lee, “Authoring tool: acquiring sharable knowledge for smart cdss,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 1278–1281.

[119] M. Afzal, M. Hussain, W. A. Khan, T. Ali, S. Lee, and B. H. Kang, “Knowledgebutton: An evidence adaptive tool for cdss and clinical research,” in *Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 2014 IEEE International Symposium on*. IEEE, 2014, pp. 273–280.

Appendix A

System Implementation

In order to realize the research methodology described in Chapters 3-5 in terms of a one-unit implementation, this work designed a comprehensive system called ”KnowledgeButton” presented in Figure A.1. The architecture composed of multiple components that work in an integrated manner to performing the task of acquiring and appraising the research evidence from online resources.

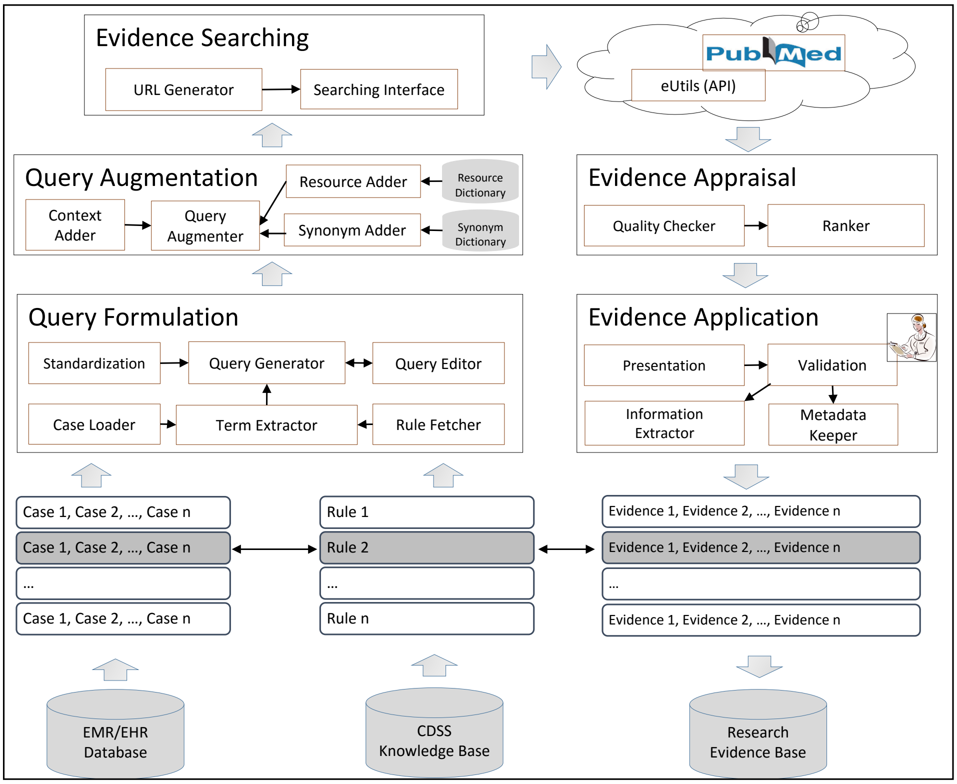


Figure A.1: KnowledgeButton Architecture

The main five components of the architecture are listed below.

- Query Formulation
- Query Augmentation
- Evidence Searching
- Evidence Appraisal
- Evidence Application

A.1 Query Formulation

Query Formulation formulates the primary query using patient information in the electronic medical record (EMR) database along with rules in the knowledge base (KB) of a clinical decision support system (CDSS). **Case Loader** loads the patient case in order to run on the rules in the KB loaded with the **Rule Loader** component. **Term Extractor** extracts the terms from the case and fired rule through the term extraction methodology discussed in Chapter 4. The primary query is generated in a standard format with **Query Generator** component. **Standardization** component standardizes the query structure and terms used in the query. For query structure, a well-known format is adapted called PICO (Patient Problem, Intervention, Comparison, Outcome) while for concept standardization, SNOMED CT vocabulary is utilized. The primary query generated is displayed to the user through the **Query Editor**.

A.2 Query Augmentation

Query Augmenter expands the terms used in the primary query using **Synonym Adder** component. The **Resource Adder** adds resources by specifying their names such as journal names, book names etc. Finally, the **Context Adder** adds the contextual terms as query filters such as treatment, diagnosis, etiology, and prognosis. The synonym terms added to the primary query terms with ”OR” logical operator, however, the resources and contexts are added at the end of teh query using ”AND” logical operator.

A.3 Evidence Searching

Evidences from online literature are searched on the basis of query, however, the query needs to be transformed into a URL acceptable by the search engine or interface. The **URL Generator** generates the URL according to the specification provided by the PubMed search engine. The PubMed URL format is given as follows.

$URL = BaseURL + DababaseName + UserQuery$

Searching Interface implements PubMed API service support called Entrez Programming Utilities (eUtils). The eUtils are a set of seven server-side programs that provide a stable interface into the Entrez query and database system at the National Center for Biotechnology Information (NCBI) [98]. Entrez currently includes 23 databases covering a variety of biomedical data. The seven programs are EInfo, EGQuery, ESearch, EFetch, ESummary, EPost and ELink. Each of these programs has its own objective to fulfill. However, most of them performed actions in a pipeline fashion. One program output is utilized by the other and so on. This work implemented three of the programs including EPost, ESearch, and EFetch with their internal methods and properties.

A.4 Evidence Appraisal

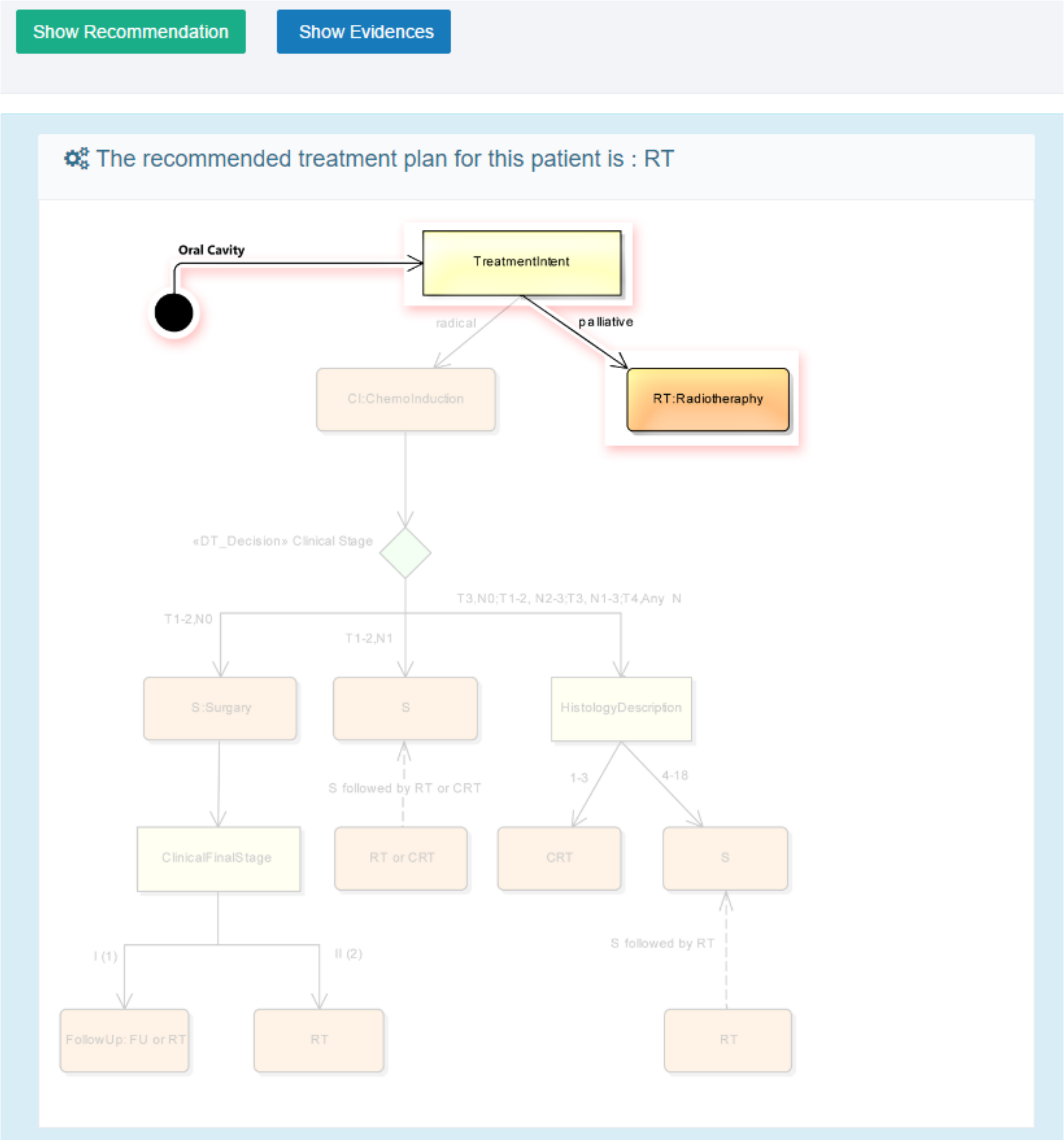
Evidence Appraisal refers to quality checking and ranking of the documents retrieved through evidence searching. **Quality Checker** checks the contents of each document and evaluate it for the rigorousness. If a document fulfills the rigorous criteria, it is included in the list of quality documents, otherwise excluded from the list. A statistical model called quality recognition model (QRM) is trained on a set of experts' team annotated data in order to evaluate the rigorousness of a document. QRM is based on support vector maching (SVM) machine learning algorithm. **Ranker** ranks the quality documents according to the rigorousness score and user context. The contextual parameters are evaluated against each document and rank the document accordingly. The contextual grading is inspired from the SORT grading system.

A.5 Evidence Application

The ranked evidentiary documents are presented to the user. The **Presenter** component displays the document in a way to provide the necessary information and associated link to the source document. Based on the intuition, experience, and knowledge, user validate the contents with **Validator** component to approve or disprove the evidence. From approved evidences, teh meta-information is collected for reusability through **Metadata Keeper**. Although not implemented yet, however, information is extracted from the approved documents with **Information Extractor** for rule mining and other analysis. The approved documents are stored in the **Research Evidence Base** by keeping links to the source documents.

A.6 Implementation Example

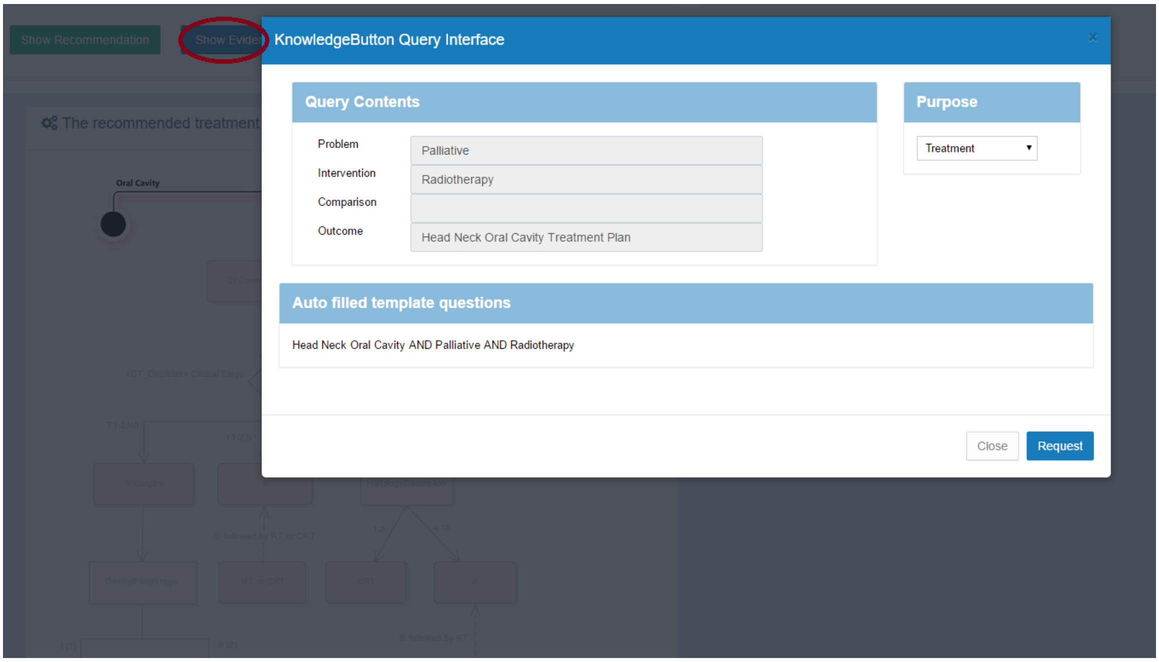
KnowledgeButton is implemented as a part of Smart CDSS [117–119]. Smart CDSS is a clinical decision support system developed for treatment recommendation of head and neck cancer. There are three major sub-systems of Smart CDSS: Knowledge Authoring Tool, Execution Engine, and KnowledgeButton. Below are some of the screen shots of the implemented Smart CDSS highlighting the KnowledgeButton sub-system. Figure A.2 shows the recommendation generated by Smart CDSS. The recommendation statement ”The recommended treatment plan for this patient is: RT” is displayed at the top left of the figure. The tree below the recommendation statement indicates the branch of rule’s tree. In other words, the involved rule are highlighted in the tree.



The recommended treatment plan for this patient is : RT

Figure A.2: Smart CDSS recommendation and rules tree

Above the recommendation statement, there is a button labeled **”Show Evidence”**. When this button is clicked. the KnowledgeButton Query interface is displaced as shown in Figure A.3. The interface is designed according to PICO format. Different elements of PICO are filled intelligently with the contents of the rules participated in the recommendation generation. Below the PICO elements, the PICO compliant query is displayed.


Show Recommendation
Show Eviden
The recommended treatment
Oral Cavity
KnowledgeButton Query Interface
×
Query Contents
Purpose
Auto filled template questions
Close
Request

Problem: Palliative

Intervention: Radiotherapy

Comparison:

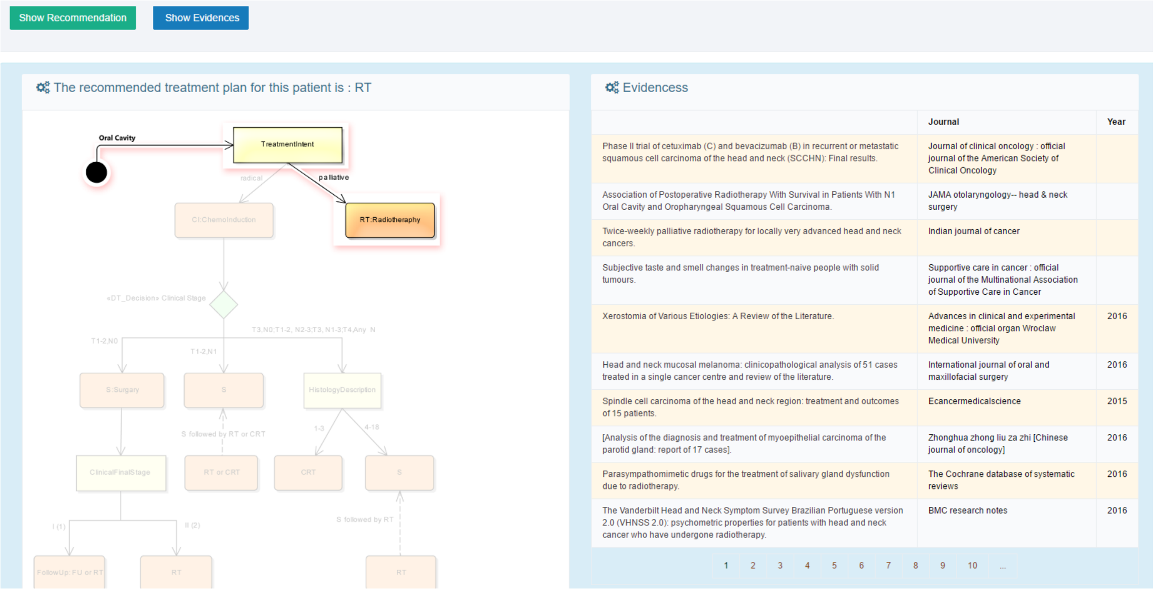
Outcome: Head Neck Oral Cavity Treatment Plan

Purpose: Treatment ▾

Head Neck Oral Cavity AND Palliative AND Radiotherapy

Figure A.3: KnowledgeButton Query Interface with PICO elements

In the bottom right, there is a button labelled **”Request”**. Upon clicking this button, the query is submitted to teh PubMed search engine. The retrieved documents are passed through teh appraisal process. Finally the appraised evidences are displayed to the user as shown in Figure A.4.



The recommended treatment plan for this patient is : RT

Evidencess

	Journal	Year
Phase II trial of cetuximab (C) and bevacizumab (B) in recurrent or metastatic squamous cell carcinoma of the head and neck (SCCHN): Final results.	Journal of clinical oncology : official journal of the American Society of Clinical Oncology	
Association of Postoperative Radiotherapy With Survival in Patients With N1 Oral Cavity and Oropharyngeal Squamous Cell Carcinoma.	JAMA otolaryngology-- head & neck surgery	
Twice-weekly palliative radiotherapy for locally very advanced head and neck cancers.	Indian journal of cancer	
Subjective taste and smell changes in treatment-naive people with solid tumours.	Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer	
Xerostomia of Various Etiologies: A Review of the Literature.	Advances in clinical and experimental medicine : official organ Wroclaw Medical University	2016
Head and neck mucosal melanoma: clinicopathological analysis of 51 cases treated in a single cancer centre and review of the literature.	International journal of oral and maxillofacial surgery	2016
Spindle cell carcinoma of the head and neck region: treatment and outcomes of 15 patients.	Ecancermedicalscience	2015
[Analysis of the diagnosis and treatment of myoepithelial carcinoma of the parotid gland: report of 17 cases].	Zhonghua zhong liu za zhi [Chinese journal of oncology]	2016
Parasympathomimetic drugs for the treatment of salivary gland dysfunction due to radiotherapy.	The Cochrane database of systematic reviews	2016
The Vanderbilt Head and Neck Symptom Survey Brazilian Portuguese version 2.0 (VHNSS 2.0): psychometric properties for patients with head and neck cancer who have undergone radiotherapy.	BMC research notes	2016

Figure A.4: List of quality evidentiary documents

Appendix B
List of Publications

Journal Publications:

[1] **Muhammad Afzal**, Maqbool Hussain, Wajahat Ali Khan, Taqdir Ali, Arif Jamshed, and Sungyoung Lee. SEAS: Smart Extraction and Analysis System for Clinical Research, Telemedicine and e-Health (SCI, IF:1.79), Accepted (August 28, 2016)

[2] **Muhammad Afzal**, Maqbool Hussain, et al. ”Knowledge-Based Query Construction Using the CDSS Knowledge Base for Efficient Evidence Retrieval.” Sensors 15.9 (SCIE, IF 2.245, 2015): 21294-21314.

[3] Oresti Banos, Muhammad Bilal Amin, Wajahat Ali Khan, **Muhammad Afzal**, Maqbool Hussain, Byeong Ho Kang and Sungyong Lee, ”The Mining Minds Digital Health and Wellness Framework”, BioMedical Engineering OnLine (SCIE, IF:1.43), DOI:10.1186/s12938-016-0179-9, 2016.

[4] Rahman Ali, **Muhammad Afzal**, Maqbool Hussain, Maqbool Ali, Muhammad Hameed Siddiqi, Sungyoung Lee, Byeong Ho Kang, ”Multimodal Hybrid Reasoning Methodology for Personalized Wellbeing Services”, Computers in Biology and Medicine (SCI, IF:1.24), Vol.69, pp.10-28, 2016

[5] Maqbool Hussain, **Muhammad Afzal**, Taqdir Ali, Rahman Ali, Wajahat Ali Khan, Arif Jamshed, Sungyoung Lee, Byeong Ho Kang, Khalid Latif; ”Data-driven knowledge acquisition, validation, and transformation into HL7 Arden Syntax”, Artificial Intelligence in Medicine (Elsevier, SCI, IF:2.03), published-online (September 15, 2015).

[6] Oresti Banos, Wajahat Ali Khan, Muha mmad Bilal Amin, , , , Maqbool Hussain, **Muhammad Afzal**, Taqdir Ali, Ł, ”   Ł ”, () Vol.32 No.11, pp.12-20, 2015

[7] Maqbool Hussain, Taqdir Ali, Wajahat Ali Khan, **Muhammad Afzal**, Sungyoung Lee, Khalid Latif,;”Recommendations service for chronic disease patient in multi-model sensors home environment” Telemedicine and e-Health (SCI, IF:1.6), 2014

[8] Maqbool Hussain, AsadMasood Khattak,Wajahat Ali Khan, Iram Fatima, Muhammad Bilal Amin, Zeeshan Pervez, Rabia Batool, Muhammad Amir Saleem, **Muhammad Afzal**, Muhammad Fahim, Muhammad Hameed Saddiqi, Sungyoung Lee, and Khalid Latif, “Cloud-based Smart CDSS for Chronic Diseases”, In Journal of Health and Technology - 3, no. 2 (2013): 153-175.

[9] Wajahat Ali Khan, M Bilal, AM Khattak, Maqbool Hussain, **M Afzal**, SY Lee and Eun Soo Kim, ”Object Oriented and Ontology Alignment Patterns based Expressive Mediation Bridge Ontology (MBO)”, Journal of Information Science, (SCIE, IF:1.08)(Accepted), 2014.

[10] Wajahat Ali Khan, AM Khattak, Maqbool Hussain, Bilal Amin, **M Afzal**, Christopher Nugent and Sungyoung Lee, ”An Adaptive Semantic based Mediation System for Data Interoperability among Health Information Systems”, Journal of Medical Systems (SCIE, IF 1.372), 38(8):28, 2014.

[11] Wajahat Ali Khan, Maqbool Hussain, **Muhammad Afzal**, Muhammad Bilal Amin, Muhammad Aamir Saleem and Sungyoung Lee, ”Personalized-Detailed Clinical Model for Data Interoperability among Clinical Standards”, Telemedicine and EHealth (SCI, IF:1.544), Vol. 19 Issue 8, pp.632-642, 2013.

[12] Wajahat Ali Khan, Maqbool Hussain, Khalid Latif, **Muhammad Afzal**, Farooq Ahmad, Sungyoung Lee, ”Process Interoperability in Healthcare Systems with Dynamic Semantic Web Services”, Computing Journal (SCI, IF:1.055), Vol. 95, Issue 9, pp 837-862, 2013.

Conference Publications:

[13] **Muhammad Afzal** and Sungyoung Lee, KAP Based PICO-Compliant Query Construction, 2016 International Symposium on Perception, Action, and Congnitive Systems, Seoul, Korea, October 27-28, 2016.

[14] **Muhammad Afzal** and Sungyoung Lee, Relevant evidence acquisition and appraisal using knowledge-intensive queries, 18th International Conference on Advanced Communication Technology (ICACT), Phoenix, Korea, Jan 31-Feb 2, 2016.

[15] **Muhammad Afzal**, Maqbool Hussain, Wajahat Ali Khan, Taqdir Ali and Sungyoung Lee, ”Towards Evidence Adaptive Clinical Decision Support System”, 12th International Conference on Ubiquitous Healthcare (u-Healthcare 2015), Osaka, Japan, Nov 30- Dec 02, 2015.

[16] **Muhammad Afzal**, Maqbool Hussain, Taqdir Ali, Wajahat Ali Khan, Sungyoung Lee, Byeong Ho Kang, ”MLM-based Automated Query Generation for CDSS Evidence Support”, Ubiquitous Computing and Ambient Intelligence (UCAmI), Belfast, Northern Ireland, December 2-4, 2014.

[17] **Muhammad Afzal**,Maqbool Hussain, Wajahat Ali Khan, Taqdir Ali, Sungyoung Lee, and Byeong Ho Kang. ”KnowledgeButton: An evidence adaptive tool for CDSS and clinical research.” In Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 2014 IEEE International Symposium on, pp. 273-280. IEEE, 2014.

[18] **Muhammad Afzal**, Maqbool Hussain, Wajahat Ali Khan, Taqdir Ali, Sungyoung Lee, and Hafiz Farooq Ahmad. ”Meaningful Integration of Online Knowledge Resources with Clinical Decision Support System.” In Inclusive Society: Health and Wellbeing in the Community, and Care at Home, pp. 280-285. Springer Berlin Heidelberg, 2013.

[19] **Muhammad Afzal**, Maqbool Hussain, Wajahat Ali Khan, Sungyoung Lee, and Hafiz Farooq Ahmad. ”Social media canonicalization in healthcare: Smart cdss as an exemplary application.” In e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on, pp. 419-422. IEEE, 2012.

[20] **Muhammad Afzal**, Maqbool Hussain, Wajahat Ali Khan, Taqdir Ali and Sungyoung Lee Sungyoung Lee, ” Role of Online Knowledge Resources in Clinical Decision Making”, The

38th Conference of the KIPS, Jeju, South Korea, Nov 2012.

[21] **Muhammad Afzal**, Maqbool Hussain, Wajahat Ali Khan, Taqdir Ali, Sungyoung Lee and Tae Choong Chung, " Knowledgebuttons in Health Systems", The 39th Conference of the KIPS, Busan, South Korea, May 10-11, 2013.

[22] **Muhammad Afzal**, Maqbool Hussain, Taqdir Ali, Wajahat Ali Khan and Sungyoung Lee, KnowledgeButton: An Evidence Support Tool for CDSS, Busan, Korea, 2013

[23] Maqbool Hussain, Taqdir Ali, **Muhammad Afzal**, Wajahat Ali Khan, Sungyoung Lee, Lets Fire the CDSS integration issue with FHIR, 14th International HL7 Interoperability Conference (IHIC 2013) and FHIR Connectathon, Sydney, Australia, 27-29 October, 2013

[24] Maqbool Hussain, **Muhammad Afzal**, Wajahat Ali Khan, and Sungyoung Lee. "Clinical decision support service for elderly people in smart home environment." In Control Automation Robotics & Vision (ICARCV), 2012 12th International Conference on, pp. 678-683. IEEE, 2012.

[25] Maqbool Hussain, Wajahat Ali Khan, **Muhammad Afzal**, and Sungyoung Lee. "Smart CDSS for smart homes." In Impact Analysis of Solutions for Chronic Disease Prevention and Management, pp. 266-269. Springer Berlin Heidelberg, 2012.

[26] Oresti Banos, Muhammad Bilal Amin,Wajahat Ali Khan,Taqdir Ali, **Muhammad Afzal** and Byeong Ho Kang, "Mining Minds: an innovative framework for personalized health and wellness support", 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2015), Trento, Italy, May 20-23, 2015.

[27] Wajahat Ali Khan, Muhammad Bilal Amin, Oresti Banos, Taqdir Ali, Maqbool Hussain, **Muhammad Afzal**, Shujaat Hussain, Jamil Hussain, Rahman Ali, Maqbool Ali, Dongwook Kang, Jaehun Bang, Tae Ho Hur, Bilal Ali, Muhammad Idris, Asif Razzaq, Sungyoung Lee and Byeong Ho Kang, "Mining Minds: Journey of Evolutionary Platform for Ubiquitous Wellness", 12th International Conference on Ubiquitous Healthcare (u-Healthcare 2015), Osaka, Japan, Nov 30- Dec 2, 2015.

[28] Wajahat Ali Khan, Muhammad Idris, Taqdir Ali, Rahman Ali, Shujaat Hussain, Maqbool Hussain, Muhammad Bilal Amin, Asad Masood Khattak, Yuan Weiwei, **Muhammad Afzal**, Sungyoung Lee and Byeong Ho Kang, ”Correlating Health and Wellness Analytics for Personalized Decision Making”, 12th International Conference on Ubiquitous Healthcare (u-Healthcare 2015), Osaka, Japan, Nov 30- Dec 02, 2015

[29] Wajahat Ali Khan, Maqbool Hussain, Bilal Amin, Asad Masood Khattak, **Muhammad Afzal**, and Sungyoung Lee. ”AdapteR Interoperability ENgine (ARIEN): An approach of Interoperable CDSS for Ubiquitous Healthcare.” In Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction, pp. 247-253. Springer International Publishing, 2013.

[30] Wajahat Ali Khan, Maqbool Hussain, **Muhammad Afzal**, Muhammad Bilal Amin, and Sungyoung Lee. ”Healthcare standards based sensory data exchange for Home Healthcare Monitoring System.” In Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, pp. 1274-1277. IEEE, 2012.

[31] Wajahat Ali Khan, **Maqbool Hussain**, Asad Masood Khattak, Muhammad Afzal, Bilal Amin, and Sungyoung Lee. ”Integration of HL7 compliant smart home healthcare system and HMIS.” In Impact Analysis of Solutions for Chronic Disease Prevention and Management, pp. 230-233. Springer Berlin Heidelberg, 2012.

[32] Wajahat Ali Khan, Maqbool Hussain, **Muhammad Afzal**, Sungyoung Lee, Interoperability Engine for Smart CDSS, AAL Summit 2012, Spain

[33] Ali Syed Imran, Muhammad Sadiq, **Muhammad Afzal** and Sungyoung Lee, ”Service Curation Framework for Context-Aware Personalzied Recommendations”, 2016 International Symposium on Perception, Action and Cognitive Systems (PACS), Seoul, Korea, Oct 27-28, 2016

[34] Jamil Hussain, Wajahat Ali Khan, **Muhammad Afzal**, Maqbool Hussain, Byeong Kang and Sungyoung Lee, ”Adaptive User Interface and User Experience based Authoring Tool

for Recommendation Systems", Ubiquitous Computing and Ambient Intelligence (UCAmI), Belfast, Northern Ireland, December 2-4, 2014.

[35] Taqdir Ali, Maqbool Hussain, Wajahat Ali Khan, **Muhammad Afzal**, and Sungyoung Lee. "Customized clinical domain ontology extraction for knowledge authoring tool." In Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication, p. 23. ACM, 2014.

[36] Taqdir Ali, Maqbool Hussain, Wajahat Ali Khan, **Muhammad Afzal**, Byeong Ho Kang, and Sungyoung Lee. "Arden syntax studio: Creating medical logic module as shareable knowledge." In Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 2014 IEEE International Symposium on, pp. 266-272. IEEE, 2014.

[37] Taqdir Ali, Maqbool Hussain, Wajahat Ali Khan, **Muhammad Afzal**, and Sungyoung Lee. "Authoring tool: Acquiring sharable knowledge for Smart CDSS." In Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, pp. 1278-1281. IEEE, 2013.

[38] Rabia Batool, Wajahat Ali Khan, Maqbool Hussain, Jahanzeb Maqbool, **Muhammad Afzal**, and Sungyoung Lee. "Towards personalized health profiling in social network." In Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in, pp. 760-765. IEEE, 2012.