**Supporting Patient Information Needs about their Diagnostic Results: A Study of Social Q&A Site**

by

Yu Lu

Submitted in partial fulfillment

of the requirements for the degree of Master of Science

in Computer Science

at

Seidenberg School of Computer Science and Information Systems

Pace University

Date

**Acknowledgment**

This thesis would not be possible without the help of numerous people. First, I would like to express my deep gratitude to my advisor, Dr. Zhan Zhang, along with Dr. Zhe He and Dr. Xiao Luo, for their continual guidance and support throughout the journey. Their mentorship empowered my endeavor in future academic training and research.

I am grateful to Dr. Christelle Scharff, of whom, along with the three mentors, for their service as the committee of my defense.

Many thanks to my friends at Pace University: Bairen Zhu, Lan Yang, Modi Zhang, Shiyun Zhang, Xuezhi Li, Daoyuan Mu, Lu Dong, Fanzhe Fu, Jingxiang Ji, Guannan Zou, Rajesh Chinni, Ameya Pingulkar, Doron Samuel and many more for making my life at Pace a memorable experience.

There are no words to express my indebtedness to my mom Yan Qi, and my dad Bin Lu, for bringing me to this world and for all of their love. My apology for constantly letting them worry. And I owe you everything.

Table of Contents

[List of Tables iv](#_Toc18093853)

[List of Figures v](#_Toc18093854)

[List of Equations vi](#_Toc18093855)

[Abstract vii](#_Toc18093856)

[Chapter 1. Introduction 1](#_Toc18093857)

[1.1 Introduction 1](#_Toc18093858)

[1.2 Research Questions and Contributions 4](#_Toc18093859)

[1.3 Thesis Organization 5](#_Toc18093860)

[Chapter 2. Related Work 6](#_Toc18093861)

[2.1 Patients’ Information Needs About Their Clinical Data 6](#_Toc18093862)

[2.2 Consumer-Health-Question Retrieval 8](#_Toc18093863)

[Chapter 3. Study I: Making Sense of Clinical Laboratory Results 12](#_Toc18093864)

[3.1 Overview 12](#_Toc18093865)

[3.2 Methods 12](#_Toc18093866)

[**3.2.1 Data Collection** 12](#_Toc18093867)

[**3.2.2 Data Analysis** 13](#_Toc18093868)

[3.3 Results 14](#_Toc18093869)

[**3.3.1 Understanding the Results of Lab Test** 15](#_Toc18093870)

[**3.3.2 Interpreting Doctor’s Diagnosis** 17](#_Toc18093871)

[**3.3.3 Learning about Lab Tests as a Source of Information** 17](#_Toc18093872)

[**3.3.4 Consulting the Next Steps** 19](#_Toc18093873)

[3.4 Discussion 20](#_Toc18093874)

[3.5 Summary 22](#_Toc18093875)

[Chapter 4. Study II: Supporting Patients’ Information Needs 24](#_Toc18093876)

[4.1 Overview 24](#_Toc18093877)

[4.2 Methods 25](#_Toc18093878)

[**4.2.1 Dataset Preparation** 25](#_Toc18093879)

[**4.2.2 Identifying Lab Test Related Questions** 26](#_Toc18093880)

[**4.2.3 Feature Extraction** 30](#_Toc18093881)

[**4.2.4 Vector Representations** 33](#_Toc18093882)

[**4.2.5 Two Similarity Calculations** 38](#_Toc18093883)

[**4.2.6 Question Ranking** 40](#_Toc18093884)

[**4.2.7 Evaluation** 41](#_Toc18093885)

[4.3 Results 44](#_Toc18093886)

[**4.3.1 Evaluation of Questions Contain Creatinine Lab Test** 44](#_Toc18093887)

[**4.3.2 Evaluation of Questions Contain Glucose Lab Test** 44](#_Toc18093888)

[**4.3.3 Evaluation of Questions Containing HbA1c Lab Test** 46](#_Toc18093889)

[**4.3.4 Evaluation of Questions Without Lab Tests** 47](#_Toc18093890)

[4.4 Discussion 48](#_Toc18093891)

[4.5 Summary 49](#_Toc18093892)

[Chapter 5. Discussions, Conclusions, and Future Work 51](#_Toc18093893)

[References 55](#_Toc18093894)

# **List of Tables**

[Table 3.1 Search terms and the number of retrieved posts. 8](#_Toc17055144)

[Table 3.2 Summary of themes. Some posts fell into multiple 10](#_Toc17055145)

[Table 4.1 Examples for the extracting the test-value statement of each lab test 17](#_Toc17055146)

[Table 4.2 Ranges for creatinine lab test 18](#_Toc17055147)

[Table 4.3 Ranges for glucose lab test 18](#_Toc17055148)

[Table 4.4 Ranges for HbA1c lab test 19](#_Toc17055149)

[Table 4.5 List of engineered features and their values of an example question 26](#_Toc17055150)

[Table 4.6 List of engineered features of an example question and their distances to query question 27](#_Toc17055151)

[Table 4.7 Weighting Scheme for questions containing glucose test results 28](#_Toc17055152)

[Table 4.8 Feature Distances for the sample candidate question 28](#_Toc17055153)

[Table 4.9 Computation of the final similarity score of between the query question and the sample candidate 28](#_Toc17055154)

[Table 4.10 Meaning of each rating score 29](#_Toc17055155)

# **List of Figures**

[Figure 3.1 Data Collection, Sampling, and Analysis Process 9](#_Toc16930185)

[Figure 4.1 Overview of the framework for retrieving similar questions. 15](#_Toc16930186)

[Figure 4.2 Distribution of questions containing lab test and values. 16](#_Toc16930187)

[Figure 4.3 Distribution of question lengths 16](#_Toc16930188)

[Figure 4.4 Distribution of sample questions containing any of the 3 targeted lab tests. 18](#_Toc16930189)

[Figure 4.5 Distribution of test ranges for glucose lab test. 19](#_Toc16930190)

[Figure 4.6 Distribution of test ranges for hemoglobin a1c lab test. 19](#_Toc16930191)

[Figure 4.7 Distribution of test ranges for creatinine lab test. 20](#_Toc16930192)

[Figure 4.8 Distribution of sentence length of the sample questions. 21](#_Toc16930193)

[Figure 4.9 Distribution of stop-word count. 22](#_Toc16930194)

[Figure 4.10 Distribution of the targeted WH question types. 22](#_Toc16930195)

[Figure 4.11 Average nDCG@[1, 10] using 5 models for creatinine query questions 30](#_Toc16930196)

[Figure 4.12 Average nDCG@[1, 10] using 5 models for glucose query questions 31](#_Toc16930197)

[Figure 4.13 Average nDCG@[1, 10] using 5 models for HbA1c query questions 32](#_Toc16930198)

[Figure 4.14 Average nDCG@[1, 10] using 5 models for query questions without lab tests 33](#_Toc16930199)

# **List of Equations**

[Equation 1 23](#_Toc16930200)

[Equation 2 23](#_Toc16930201)

[Equation 3 23](#_Toc16930202)

[Equation 4 24](#_Toc16930203)

[Equation 5 24](#_Toc16930204)

[Equation 6 25](#_Toc16930205)

[Equation 7 25](#_Toc16930206)

[Equation 8 25](#_Toc16930207)

[Equation 9 26](#_Toc16930208)

[Equation 10 27](#_Toc16930209)

[Equation 11 28](#_Toc16930210)

[Equation 12 29](#_Toc16930211)

[Equation 13 29](#_Toc16930212)

[Equation 14 29](#_Toc16930213)

# **Abstract**

Clinical data presented on patient portals are of various technical complexities (e.g. medical jargons, the lack of personal relevance and actionable knowledge), hindering patients’ ability to interpret the data and make informed decisions. As a result, patients turn to online resources such as online Q&A websites to seek fulfillment to their information needs about their clinical data. However, few studies have examined patients’ information needs in the context of understanding laboratory test results, even though understanding lab results would help patients make better decisions and achieve optimal health outcomes. Also, existing question-retrieval algorithms fail to take into account several factors such as the type of lab test and the range of test results when retrieving relevant questions and answers. To address these research gaps, this research was set to 1) assess patients’ information needs related to their lab test results, and 2) develop an algorithm to improve the accuracy and efficiency of retrieving lab tests-related questions. Through a qualitative analysis of lab tests-related questions posted on Yahoo! Answers, 15 information needs related to laboratory test results were identified and clustered under four themes: understanding the results of lab test, interpreting doctor’s diagnosis, learning about lab tests as a source of information, and consulting the next steps. This thesis also proposed a novel question retrieval algorithm that could significantly improve the accuracy of retrieving relevant questions by considering several factors such as the range of lab tests. The contributions of the thesis include: 1) providing a comprehensive, fine-grained understanding of lay people’s information needs in making sense of their laboratory test results, 2) developing a novel algorithm for more accurate retrieval of lab tests-related questions, 3) informing the design of support tools, including patient portal, to better communicate laboratory tests with personalized and actionable knowledge to lay individuals.

# **Chapter 1. Introduction**

## **1.1 Introduction**

With the wide adoption of personal health technology, such as online patient portals [1], clinical data is, therefore, becoming more accessible to patients. However, in many portals, clinical data was presented to patients in the same way as it is shown to healthcare providers, while patients may not have sufficient health literacy to process and understand the implication of the language (e.g., medical jargons) used in the clinical reports [2-4]. In this case, a communication barrier arises between patients and healthcare providers. Also, due to the lack of personalized and actionable knowledge that patient portals ought to provide [5], patients usually have challenges understanding their health status and were hence hindered from making informed decisions to achieving better health outcomes [6]. As a result, patients would turn to online resources (e.g. health forums and social media) to fill their knowledge gaps [7].

Among the many portal functionalities, access to laboratory test results is an area of high interest to patients; growing evidence suggests that patients are increasingly interested in timely and easy access to laboratory test results [8]. However, patients’ current use of test result data is significantly limited due to several reasons [9]. For example, many portals present clinical data to patients in the same way as it is shown to healthcare providers, while patients may not have sufficient health literacy to process and understand the technical nature of the language (e.g., medical jargons) used in the laboratory test reports [3, 4]. Also, patients hope to find useful information, such as actionable knowledge, in online portals, rather than just reviewing the data [5, 10]. These findings suggest that while healthcare organizations are increasing patients’ access to their clinical data via patient portals, this technology has not adequately met patients’ information needs.

Therefore, patients often turn to online resources (e.g., search engines, health forums, and social media) to fill their knowledge gaps. A recent Pew Research Center study reported that over 70% of adult Internet users in the U.S. searched online for health information [7]. As one of the most popular activities online, health information searching has been the focus of many studies over the past decades [11]. However, to date, only a few studies have started looking into patients’ online health information-seeking behaviors in the context of understanding laboratory results. For example, Reynolds et al. [4] examined the type of supports patients need related to their laboratory data by analyzing questions in an online health forum (medhelp.org). In particular, they found that patients tend to ask questions about several topics: diagnosis, management/treatment, laboratory report, test, risk, and prognosis. Their study also preliminarily assessed the feasibility of identifying and characterizing the nature of patients’ questions related to laboratory results. Building upon their work, Chapter 3 conducted an explorative study to gain further insights into patients’ general information needs concerning laboratory test results, and to inform the development of novel computational methods for supporting these needs.

Another research gap is that the accuracy and efficiency of retrieving online questions containing lab results are still not optimal, because current search engines fail to take into account the type of lab test and the range of its corresponding numeric results in questions [12-14]. Imagine such a scenario: if a user retrieved questions containing lab results that are not in the same range as the patient’s input query question, the chosen answers to these retrieved questions could potentially misinform the patient. To address this research gap, Chapter 4 developed a question-retrieval algorithm to support a major theme in patients’ information needs, *Understanding laboratory test results*, as defined in Chapter 3.

Among various online resources, social Q&A websites enable users to retrieve their targeting information through posting questions and receiving answers from others (e.g., peer patients, health professionals) who are willing to share their knowledge and opinions. Comparing to other tools (e.g. health search engines), users can ask full questions in natural language (as opposed to search terms) and include key contextual health information with these questions to seek personalized health information. In particular, *Yahoo! Answers* is one of the popular community-based social Q&A platforms. It devotes a wide range of topic categories to health-related Q&A contents, hence it is a rich source of text corpora concerning consumer health [15]. In this thesis, a content analysis was conducted to examine patients’ information needs about their laboratory test results. The results helped us understand the types of support that patients need and the facing challenges. In a follow-up study, a question-retrieval algorithm was developed to better support patients’ information seeking in social Q&A sites and to address the above-mentioned research gaps. The retrieved top 10 results of the algorithm were compared to the results of the existing sentence embeddings using the normalized cumulative gain (nDCG). This work focuses on one type of common chronic disease, diabetes, rather than any clinical conditions or symptoms. Diabetes is a worthwhile disease for investigating lay-people’s information needs regarding their laboratory test results for the following reasons: First, it is one of the most common disease that has raised public awareness globally [16]. Second, laboratory tests vary across different types of conditions and diseases, thus, focusing on one condition allowed us to generate comprehensive search terms for data collection. Third, focusing on diabetes allowed for analyzing not only general laboratory test questions but also questions on condition-specific laboratory tests.

## **1.2 Research Questions and Contributions**

This thesis focuses on the following research questions that are inspired by the previous literature:

*RQ 1:* **What are the lay individual’s information needs related to laboratory test results?**

It is crucial to understanding lay individual’s information needs related to lab test results. An answer to this question addresses the gap between patient knowledge and limited contextual information presented on their lab reports and would provide insights into the design of support tools, such as patient portals, to fulfill patient needs in understanding lab results.

*RQ 2:* **How to implement novel computational methods to better support these information needs?**

Identifying the existing computational methods for supporting these information needs and the limitations of these methods help formulate an answer to this question. Subtending questions associated with this research question are as follows:

1. What are the methods that have been implemented to fulfill these information needs?
2. What are the limitations of these methods?
3. How can we implement a method that could address these limitations while helping patients make sense of their laboratory test results?

Through studies conducted in this work, the following contributions were made:

1. understanding lay people’s common information needs in making sense of their laboratory test results;
2. Implementing and evaluating a Question Retrieval algorithm that improves the accuracy and efficiency of search results related to lab tests;
3. Informing future patient portal design for better communication of laboratory tests with personalized and actionable knowledge;

## **1.3 Thesis Organization**

The goal of the thesis is to support patients’ information needs about their clinical laboratory test results. The remaining of the thesis is organized as the following:

Chapter 2 presents an overview of previous work focused on patients’ information needs about their clinical data and information retrieval of healthcare-related questions in online forums and social Q&A sites.

Chapter 3 presents a qualitative analysis of health-related posts on Yahoo! Answers, which uncovers 15 information needs related to laboratory test results and categorizes these information needs into different themes. This paper has been published in the MedInfor 2019, “Understanding Patient Information Needs about their Clinical Laboratory Results: A Study of Social Q&A Site,” written with co-authors, Zhan Zhang, Zhe He, and Caleb Wilson [17].

Chapter 4 demonstrates the development of a question retrieval algorithm that leverages similarity measure in three vector-space models along with engineered features to retrieve previously answered questions that are similar to users’ input query questions.

Chapter 5 presents a discussion of the content analysis in Chapter 3, and the feasibility of the algorithm presented in Chapter 4; summarizes the contributions of these two chapters and lastly draws the conclusion.

# **Chapter 2. Related Work**

## **2.1 Patients’ Information Needs About Their Clinical Data**

The major cause of patients’ online information-seeking behavior was patients’ difficulty in understanding their laboratory data. Alpert et al. [5] assessed the understanding of the portal’s use from both clinicians’ and patients’ perspectives. They found that patients often experience difficulties in interpreting diagnostic laboratory data presented in portals. Overall, such difficulties can be attributed to insufficient explanation from physicians. As Vieder et al. [18] suggested, physicians’ interpersonal skills need to be more prominent to bridge the existing communication gaps between physicians and patients. Additionally, Graham et al. [2] suggested that patients’ limited health literacy also causes a communication barrier between the doctors and the patients, which in turn poses risk to patients’ safety. As a result, patients increasingly go online for supplemental information to make sense of their lab test results [7].

Existing literature has assessed patients’ online information-seeking behavior concerning their clinical data presented in patient portals. For example, Zhang [11] conducted a thematic analysis to understand the contextual factors of consumer health information-seeking in questions posted on social Q&A websites, including *linguistic features* of the questions that users formulated, *users’ motivations* for asking the questions, the *time* when the questions were asked, and *users’ cognitive representations* of the problem space. Overall, consumers’ questioning behavior was driven by a knowledge gap, disturbing feelings, or the lack of social resources [11]. Additionally, time plays an important role in patients’ information seeking. People tended to have different concerns at different stages of their health and illness [11]. In addition, Eastin and Guinsler [19] conducted an online survey about how people used the Internet for different reasons, of which suggest that individuals with even moderate levels of anxiety seek higher amounts of online health information. For those mildly or merely not worried about their health, engaging in frequent online health-information-seeking decreases the number of doctor visits they make based on their findings. Longo et al. [20] conducted 9 focus groups with 46 adults with diabetes. Their analysis of transcripts and notes from these focus groups showed that the participation of healthcare professionals helps patients to understand and manage medical information and that health literacy makes a difference, as participants across groups consistently expressed a preference for information that is easy to understand (e.g. being interpreted or written in lay-language).

However, only a few studies examined patients’ information-seeking behavior in the context of understanding laboratory test results. For example, Reynolds et al. [4] examined patients’ question posts containing laboratory results in an online health forum (medhelp.org) to reveal a spectrum of patients’ confusion and a knowledge gap between the patients and the healthcare providers. Patients’ question posts were categorized into various topics (e.g., Diagnosis/Cause, Management/Treatment, Laboratory Report, Test/Diagnostic, Risk, Prognosis) and types of requests (e.g., Opinion, Advice, Decision Support, Information (generic), Emotional Support, and Personal Experience). The results indicated that providing information and knowledge that are relevant to patients’ laboratory results would reduce time and effort burdens seeking medical information while ensuring that patients make decisions based on information curated by reliable sources. Although they highlighted patients’ confusion about the laboratory report, the study did not reveal the specific aspects of laboratory reports that patients have difficulties with, such as the meaning of lab values, medical terminologies, and the causes and effects of abnormal lab results.

In summary, despite the effort to exhaustively evaluate questions concerning laboratory results, all work listed above failed to address the various aspects of laboratory reports that patients had difficulty with. A more comprehensive, fine-grained analysis of patients’ question posts in online social Q&A sites is needed to examine patients’ information needs regarding their laboratory report and to fill the knowledge gap between patients and the healthcare providers. In contribution to this line of work, Chapter 3 presents a qualitative analysis of patients’ question posts on Yahoo! Answers, and uncovered various types patients’ information needs regarding their laboratory results under different themes.

## **2.2 Consumer-Health-Question Retrieval**

Question Answering (QA) has been a widely explored field in neural information retrieval. The mechanism of QA systems varies, with some focusing on retrieved previously-answered similar questions, while others focus on automatically answering the existing questions. For example, Luo et al. developed SimQ [12] to retrieve similar consumer-health questions. After collected question-answer (QA) pairs written by domain experts, SimQ employed semantic annotation to recognized the health-related entities from free-text questions and mapped the biomedical terms to the Unified Medical Language System (UMLS) concepts [12]. To extract syntactic features, they leveraged the AQUA parser to assign the Part-of-Speech (POS) tagging to the text corpora. Vector-space models were also used to calculate the inter-question similarity scores.

Based on SimQ, Wongchaisuwat et al. applied a supervised learning approach to identify candidate answers to a query question from historical question-answer pairs [14]. Different from Luo’s approach, Wongchaisuwat et al. employed the Dynamic Time Warping (DTW) measure instead. The data source used was strictly limited to QA pairs from Yahoo! Answers, which implies that the corpora are much noisier and that the syntactic features proposed by Luo et al. are not as useful in this model.

Shtok et al. developed an algorithm to reduce the rate of unanswered questions [13]. The input of the system is a triplet consisting of a past question and its answer, along with a new question. Lexical-syntactic features, such as sentence length, number of question marks, and stop-word count were then extracted. Then an intra-question similarity was generated to capture the degree of coherence between the titles of each two new questions and every two past questions. A triplet of question-answer similarity was computed as the following: similarity between the title of a new question and a candidate answer, similarity between the title and content of a new question and a candidate answer, similarity between the title of a past question and a candidate answer, and similarity between the title and content of a past question and a candidate answer [13].

In the domain of question retrieval, different learning models were also employed for various considerations. For example, traditional Information Retrieval (IR) models (e.g. BM25, TF-IDF) can hardly determine the similarity between a query question and a retrieved question that has a similar meaning but shares only a few words in common. e.g. example was shown in Zhang et al. [21]: a user query “how do I get knots out of my cats fur”, there are good answers under an existing question “how can I remove a tangle in my cat’s fur” in Yahoo! Answers. The user query used different words but expressed similar meaning, which reveals a lexical gap that prevents traditional IR models from matching the query with semantically similar questions.

To minimize such lexical gap, Zhou et al. [22] developed a word-embedding approach that treats a question as a Bag-of-Embedded-Words (BoEW) by employing the *skip-gram* model [23]. The BoWE is then transformed into fixed-length vectors by using the Fisher kernel (FK) framework [24]. In the next step, fisher vectors (FVs) were generate by aggregating the BoEWs for all the questions. Question retrieval was achieved by calculating the similarity between the FVs of a queried question and an existing question in the archive. Lei et al. [25] adopted gated (non-consecutive) convolutions semi-supervise question retrieval tasks. In the pre-training phase, an encoder-decoder network that generates the title of a question given title, body or merged title-body of the question, is trained. For comparison, three alternative benchmark encoders (LSTMs, GRUs, and CNNs) were also trained to map questions to vector representations. The models were evaluated using the Stack Exchange AskUbuntu dataset. For the test set, similar pairs were annotated manually as the user-marked pairs and were found to be noisy. For evaluation purposes, a set of questions retrieved using BM25 were ranked by their similarity scores computed using the learned model. The results showed that the model yielded substantial gains over a standard IR baseline and various neural network architectures.

Instead of representing query questions and candidates as bags of words, Brokos et al. represented them as pre-computed centroids of their embeddings and retrieved candidates whose centroids are closer to the centroid of the query question [26]. The method was tested on biomedical questions from the BIOASQ competition [27]. The retrieved candidates were then reranked using relaxation of the Word Mover’s Distance (WMD) proposed by Kusner et al. [28], which measures the distance between the embeddings of query question and the candidate question. Despite this method outperformed PubMed, one of the most widely-used biomedical search engines, it did not take into account the type of lab test and its corresponding numeric values in those biomedical questions.

Overall, word embeddings are models learned using a loss function defined on word pairs. And these embeddings have been widely used in the domain of question retrieval. In contrast, sentence embeddings are learned using a loss function defined on sentence pairs. This type of embedding considers the relationship among words in the sentence, e.g. the context information, and is, therefore, more suitable for tasks such as computing semantic similarities between texts. For example, Palangi et al. proposed a sentence embedding model that uses the Long Short-Term Memory (LSTM) and the Recursive Neural Network (RNN) that sequentially extracts the information contained in each word and embeds it into a semantic vector. Their model has been proven to outperform several existing state-of-the-art doc2vec models on a web search task [29]. Cer et al. proposed the Universal Sentence Encoder, which is another model for encoding sentences into embedding vectors [30, 31]. The transfer learning using sentence embedding was proven to outperform word-level transfer. However, the application of such a paradigm has not been seen in the domain of question retrieval.

In summary, existing literature in the field has made remarkable progress in developing different embedding models and applying them to health question-retrieval tasks. However, few studies assessed health question retrieval in the context of understanding clinical laboratory results. To fill such a knowledge gap, there is a need to develop a question-retrieval algorithm that takes into accounts the type of lab test and the range of its corresponding numeric results during reranking.

# **Chapter 3. Study I: Making Sense of Clinical Laboratory Results**

## **3.1 Overview**

In this study, we began our inquiry by asking: What are lay people’s information needs in making sense of their laboratory test results? We identified 15 information needs related to laboratory test results, and clustered them under four themes: understanding the results of lab test, interpreting doctor’s diagnosis, learning about lab tests as a source of information, and consulting the next steps. This study highlights the need to address the gap between patient knowledge and limited contextual information presented on their lab reports.

## **3.2 Methods**

### **3.2.1 Data Collection**

Using the application program interface (API) of Yahoo! Answers, we collected a total of 58,422 questions in the diabetes category between 2009 and 2014. The questions were downloaded in a csv format to a MySQL database. We then extracted 8655 posts using keywords suggested by the guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes, such as HbA1c, glucose, and creatinine [32]. The complete search terms and the number of posts retrieved by each term are listed in Table 3.1. The terms “glucose” OR “blood sugar” yielded the most posts (87.1%). The study was approved by the institutional review boards at Pace University and Florida State University.

Table 3.1 Search terms and the number of retrieved posts.

|  |  |
| --- | --- |
| Search Terms | Number |
| “lab” OR “laboratory” | 243 |
| “A1c” OR “HbA1c” OR “hemoglobin A1c” | 427 |
| “glucose” OR “blood sugar” | 7,536 |
| “blood pressure” OR “systolic” OR “diastolic” | 338 |
| “creatinine” | 111 |

### **3.2.2 Data Analysis**

We generated a random sample of 1,619 posts of the potentially relevant question posts (8655 posts containing keywords). Then two researchers independently reviewed posts for relevance. Duplicate or irrelevant posts were discarded. The posts were determined to be irrelevant if they did not contain any laboratory results or questions related to laboratory tests. This screening resulted in 967 posts eligible for further analysis. The relevant posts were then transferred into NVivo, a program for organizing, storing, and manipulating qualitative data. The research team performed a content analysis on these relevant posts. The analysis was performed independently by three researchers and consisted of multiple steps (Figure 3.1).

The first step was to iteratively develop a codebook using the open coding technique. Two coders, C1 and C2, independently analyzed 240 randomly sampled posts until saturation was reached. The initial list of codes was generated and then discussed in a group session to determine which codes to keep, merge, or remove. After the list of codes was set, we created a data dictionary defining each code to standardize the coding process. Our final coding scheme contained a total of 15 codes, which were clustered under four themes: understanding the results of lab test, interpreting doctor’s diagnosis, learning about lab tests as a source of information, and consulting the next step (see Table 2).

Next, a third coder (C3) coded 100 randomly sampled posts from the rest of the posts to check for exhaustiveness of the themes. Once confirming that the themes were comprehensive, C1 and C2 independently coded another set of 100 posts to check for inter-rater agreement using Cohen’s Kappa coefficient. The resulting kappa value was analyzed using the kappa interpretation scale suggested by Landis and Koch [33]. The coders presented “Almost Perfect” agreement (kappa value of 0.851). The disagreements were mainly due to the interpretive differences attributed to “Confused about doctor’s suggestions or diagnosis” and “Seeking confirmation of doctor’s diagnosis” codes; all the disagreements were resolved through discussion. Once resolving all disagreements, C1 and C2 coded the rest of the posts to conclude the analysis.

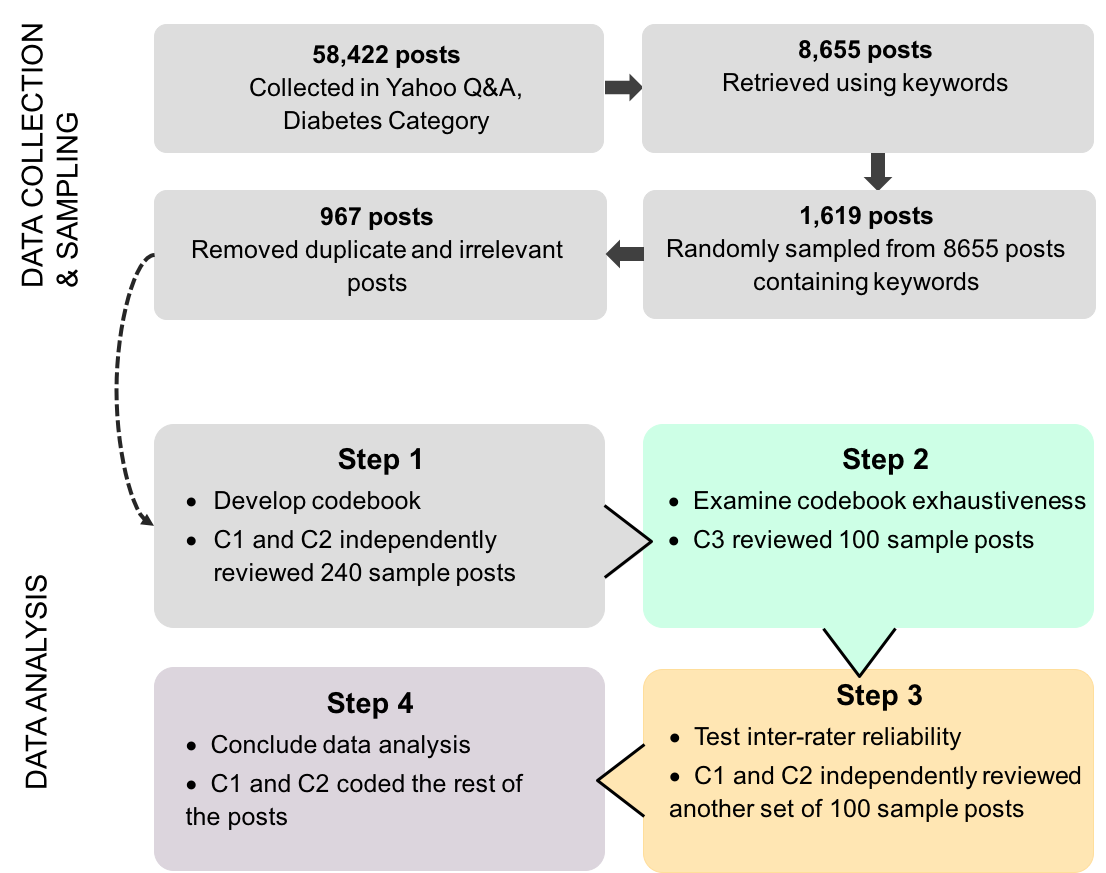


Figure 3.1 Data Collection, Sampling, and Analysis Process

## **3.3 Results**

People come to Yahoo! Q&A to request advice, suggestions, information pertaining to laboratory test results. Their information needs are multi-faceted, manifested in their multiple different but related questions. In this section, we will describe these information needs that people expressed in the questions. We will use representative quotes to illustrate salient themes.

### **3.3.1 Understanding the Results of Lab Test**

Requests for understanding laboratory test results were by far the most common in this sample (85% of the total posts). That is, posters shared parts of the report content and asked the community to explain their lab test results. As Table 2 shows, posters needed help in understanding different aspects of their lab content, including the meaning of lab value, specific terminology, and the effects and causes of abnormal and/or inconsistent results.

The most common questions in this category were related to understanding the meaning of lab values (74% of the total posts). We found that people had different needs in making sense of their lab results, which may be due to different levels of health literacy, knowledge and experience [34]. For example, a post sought a clarification whether a specific lab value falls into the normal range: “Is GFR of 73 and creatinine 1.1 normal?” In other cases, posters often asked the community for diagnosis or opinions, by providing substantial portions of their reports and relevant medical history, medication information, and symptoms. For example, a post sought opinion on what the lab results indicated:

*“I am an 18-year-old male. […] Some of the statistics from the report were as follows: high alkaline phosphatase levels, elevated T3, elevated Hbg levels. […] My AbC1 level was 6.1. What do these elevated levels seem to point to? Can anyone make sense of what might be wrong from my lab results?”*

For example, a post asked for the clarification of a specific term: *“Does anyone know what is the meaning of ‘Lymph’ on blood labs?”* This observation suggests that people have difficulty understanding medical terminologies, even though some patient portals have started implementing consumer-friendly vocabularies [35].

Finally, people wanted to know the effects and/or causes of abnormal lab results. For example, a post asked for advice on the consequences of high creatinine level: “A recent pathology test states that my creatinine is 6.28. […] What are the effects of such high levels?” In other cases, people expressed concerns about inconsistent lab results they received from different laboratories or over a period of time, as one post stated: *“My creatinine level increased from 1.0 to 1.1 with a span of 10 days’ period. What is the reason?”*

*Table 3.2 Summary of themes. Some posts fell into multiple*

*themes, so percentages add up to more than 100%.*

|  |  |
| --- | --- |
| **Theme** | **% (n)** |
| **Understanding laboratory test results** |  |
| Meaning of lab value | 63.23% (356) |
| Specific terminology | 2.66% (15) |
| The effect of abnormal/inconsistent results | 0.53% (3) |
| The cause of abnormal/inconsistent results | 5.68% (32) |
| **Interpreting doctor’s diagnosis** |  |
| Confused about doctor’s suggestions/diagnosis | 0.71% (4) |
| Seeking confirmation of doctor’s diagnosis | 2.31% (13) |
| Concerned about doctor’s misdiagnosis | 0.36% (2) |
| **Learning about lab tests as a source of information** |  |
| Inquire information about a specific lab test | 5.51% (31) |
| Ask for lab test recommendations | 0.36% (2) |
| Look for comparison among tests | 1.07% (6) |
| Concerned about lab procedure | 2.13% (12) |
| **Consulting the next steps** |  |
| Healthcare consultation | 14.92% (84) |
| Treatment options | 4.09% (23) |
| Taking medication | 3.55% (20) |
| Life-style | 5.33% (30) |

\* The percentages are calculated using the number of

posts in each category divided by the total number

of posts (N=967).

### **3.3.2 Interpreting Doctor’s Diagnosis**

Sometimes people posted questions after they discussed the results with their physicians and cited several reasons for doing this. First, people may have doubts about, disagree with, or mistrust their physician’s diagnosis, thus seeking a second opinion on their physician’s conclusions and/or interpretations (referred to as seeking confirmation of doctor’s diagnosis in Table 3.2). For example, in one post the person stated:

*“My 4-year-old [child] had all the symptoms and signs of type 1 diabetes so his doctor run test for him. What came back was Glucose, Blood 71, Insulin, Fasting 1.2, Low, C Peptide 0.4 Low. Doctor says there are a few low things, but nothing to worry about. I in my gut don't think that is right. Can someone else help me out?”*

Second, people seemed to be confused about their physician’s diagnosis or suggestions as to what to do next and whether or not the treatment is needed. Therefore, they turned to online forums to seek clarification or explanation regarding the information they received from their physician: *“Why do I need to test my creatinine level every three months as being suggested by my doctor?”*

These findings reveal a communication gap between health care providers and consumers. Misunderstanding or confusion about doctors’ diagnosis may adversely affect patients’ access to health information, resulting in poor patient understanding, trust, and satisfaction.

### **3.3.3 Learning about Lab Tests as a Source of Information**

This category concerns questions related to lab test itself. For example, lack of sufficient knowledge about lab tests led people to inquire general information about them (referred to as inquire information about a specific test in Table 2), as shown in one post: “What is creatinine cholesterol?” In other cases, people asked for some other general information about lab tests, including the relationship between lab tests and symptoms (e.g., *why are urea and creatinine levels raised with dehydration?*), how often taking a specific lab test (e.g., *how often should creatinine and eGRF levels be checked?*), and treatment options (e.g., *my creatinine is 1.6, what is the treatment for it?*).

People also inquired about the diagnostic abilities of a specific test and sought recommendations on which lab test to take (referred to as ask for lab test recommendations in Table 2). As this data sample focused on a diabetes online community, the questions, therefore, were related to lab recommendations for diabetes: *“Which laboratory test is diagnostic for diabetes?”* Similarly, people also sought comparison among different types of test (referred to as looking for comparison among tests in Table 3.2): *“Advantages and disadvantages of creatinine clearance test vs. plasma creatinine?”*

Lastly, the posters asked questions about the lab procedure. Sometimes, they posted questions while they were waiting for the tests. At this stage, posters asked questions concerning various aspects of the lab procedure, such as what they will go through during the test: *“I am going to the lab to get tested for hypoglycemia (low blood sugar) tomorrow, what exactly will they do?”* Others looked for information as to what they should do or not do to prepare for the upcoming tests: *“This is a lab test for diabetes, blood sugar, cholesterol etc. And I am wondering how long should I fast and can I drink water?”* Similarly, people also posted questions after taking their tests to inquire the turnaround time of their test results: *“How long should it take for a doctor office to call you about lab results?”* These posts tended to exhibit language indicative of distress: *“I had lab work done last Thursday and I am still waiting to hear what my A1C and all else [the doctor] had me tested for. Shouldn’t they call you with results sooner? What if something is really wrong?”*

### **3.3.4 Consulting the Next Steps**

Sometimes people also consulted the community about what they should be doing next. One reason was that people may be waiting for an appointment to discuss the results with their physician, but they wanted to obtain actionable suggestions from the online community first: *“I have lupus [and] my routine blood work shows the ck enzyme at 271 (ref range is 26-192). I have an upcoming doctor appoint. What can I do?”*

They also asked for the community’s assistance in assessing the need for a healthcare consultation or further lab test (referred to as healthcare consultation in Table 3.2). For instance, a poster expressed the lack of confidence in the accuracy of lab results and asked for advice as to if it is necessary to re-do the test or take a different test: *“High blood sugar – should I get a second opinion from a different lab? This is too important not to double check with a different lab; last reading was 6.4. This doctor was wrong before about different things.”*

Of these posts, people also asked for treatment advice (referred to as treatment options in Table 3.2). For example, one poster wrote: *“A recent pathology test states that my creatinine is 6.28. Does it require dialysis to be done? What can cure this high level?”* In such cases, people also wanted to know what medication and/or whether changing lifestyle (e.g., diet and exercise) could be of any help (referred to as taking medication and life-style in Table 2), as one post stated: *“My mother aged 45 and has only one kidney. [Her] creatinine level [is] 4.2, Urea [is] 50. What diet she should take and what medicine?”*

## **3.4 Discussion**

In the study, we characterized lay people’s general information needs related to laboratory test results, such as understanding test results, interpreting doctor’s diagnosis, learning about lab tests, and making decisions on the next steps. This study presents an early investigation for our long-term goal of guiding the design of patient portals that can provide more informative and personalized healthcare information. Building upon prior work [4], our study provides a more comprehensive, fine-grained description of lay individual’s information needs about their laboratory test results. For example, Reynolds et al. [4] highlighted that patients have confusion about the laboratory report; our study further revealed the aspects of laboratory reports that patients had difficulties with, such as the meaning of lab values, medical terminologies, and the causes and effects of abnormal lab results. While this study only examined a subset of questions in an online forum setting, our findings reveal that people need support in interpreting and acting on clinical data, as well as making personalized decisions. Below, we draw on our findings to discuss five design opportunities for supporting the understanding of laboratory results in patient portals.

***Providing consumer-friendly and credible information to assist the reading of lab results.*** Our findings suggest that the design of test results in patient portals seems to assume that patients have sufficient medical knowledge about their test results. Consequently, patients often did not receive explanatory information or result interpretation in the portal at the time they received the result, and they would search online to make sense of their results. It is, therefore, crucial to provide more useful information that patients need at the point of viewing their laboratory results in patient portals. For example, patient portals could provide links to consumer-friendly and credible information sources (e.g., entries in MedlinePlus) to help patients better understand the lab results; the portal could also suggest basic healthcare management advice, such as diet and lifestyle.

***Accommodating people with different health literacy.*** People have different levels of health literacy and numeracy as well as potential biases and personal beliefs. For patients who were recently diagnosed, they may not be literate enough to understand the terminology and the results, and thus may ask basic questions such as whether a particular lab value falls into the normal range. In contrast, some patients who have had chronic conditions may have been self-educated on relevant health knowledge (e.g., medical terminology, normal ranges of a test) and therefore need help with more comprehensive questions (e.g., how to interpret the lab results in the context of their medical history). Given such a fact, patient portals need to be designed taking into consideration of people’s health literacy differences [36].

***Considering the temporality and illness trajectory of patients.*** We also observed that patients’ information needs had a temporal dimension—the nature and extent of the needs may be different at different stages of patients’ illness trajectory [11]. For example, right before getting a medical test, patients may want to know how to prepare for the test and what they will go through during the test. Upon receiving their test results, patients may ask for interpretations of what the results mean and what they should be doing next (e.g., make an appointment with their physicians). This observation shed light on portal design with regard to the temporal organization of information materials to provide relevant health information to patients according to their illness trajectory.

***Facilitating shared decision making through personalized and contextualized information along with lab results.*** An interesting observation is that patients provided contextual information (e.g., medical history, symptom) along with their lab results in order to seek personalized advice and treatment options. This observation suggests that the same lab results may have different indications in different contexts (e.g., family history). In addition, prior work has recognized that personalized healthcare information within a shared-decision making framework leads to better patient engagement, better outcomes, and an increased level of trust between healthcare providers and patients [37]. As such, patient portals should provide more personalized content.

***Supporting the sharing of personal stories between patients who are “in the same boat”.*** Sometimes, patients sought health information due to their suspicion about a certain diagnosis made by their physician. This means that patients not only need objective explanations of terms and values in test results, but also other patients’ opinions and experiences. Such behavior constitutes reflection upon and distrust in doctors’ explanations. It seems that when authoritative explanations lost credibility in certain cases, patients were in urgent need of a second opinion, especially from patients with similar symptoms and conditions. This observation suggests that a social network in patient portals could benefit patients by connecting them with peers who have similar conditions. This also suggests that patient portals should provide a more streamlined communication channel between healthcare providers and patients in order to resolve any misunderstandings in a timely manner.

## **3.5 Summary**

This study explored lay people’s various information needs related to lab results through analyzing forum posts collected from a social Q&A site. Our results highlighted the need to address the gap between patient knowledge and limited contextual information presented on their lab reports and provide essential insights into improving the design of patient portals to fully meet patient needs in understanding the lab results. Our findings provide a foundation for our future work, including qualitative studies (e.g., interview with clinicians and patients) and analysis of medical record data to understand how to best provide personalized information and present clinical data in patient portals.

# **Chapter 4. Study II: Supporting Patients’ Information Needs**

## **4.1 Overview**

This study proposed a question-retrieval algorithm as the solution for supporting patients to make sense of their laboratory results. Vector-space weighted similarity and engineered feature distance measures were employed for assessing the similarity between two questions. The algorithm was tested using questions sampled from Yahoo! Answers. Reranking of the retrieved questions takes into accounts the type of lab test and the range of its corresponding value. At the current stage, we focused on the three clinical lab tests related to diabetes: glucose test, HbA1c test, and creatinine test. A group of questions containing no lab tests was also used for testing the similarity computation. Finally, we adopted the normalized discounted cumulative gain measurement to assess the human evaluation results of the ranked questions using our model as compared to other existing models.



Figure 4.1 Overview of the framework for retrieving similar questions.

## **4.2 Methods**

### **4.2.1 Dataset Preparation**

We collected users’ posts from the health section in *Yahoo! Answers*, and stored it as a json file. The dataset consists of 58,188 threads, each of which contains a series of properties, such as the *id* of the user, *id* of the posted question, *category* in which the post belongs to, *subject* of the question, *content* of the question, *number of the answers*, content of the *chosen answer*, and the *url* of the post. By using Valx, a system initially developed to extract numeric lab test comparison statements by Hao et al [38], we parsed the dataset and found that 31,165 out of 58,188 questions (53.6%) mentions at least one clinical laboratory results and provided a value to it, which demonstrates the importance of this project (as shown in Figure 4.2).

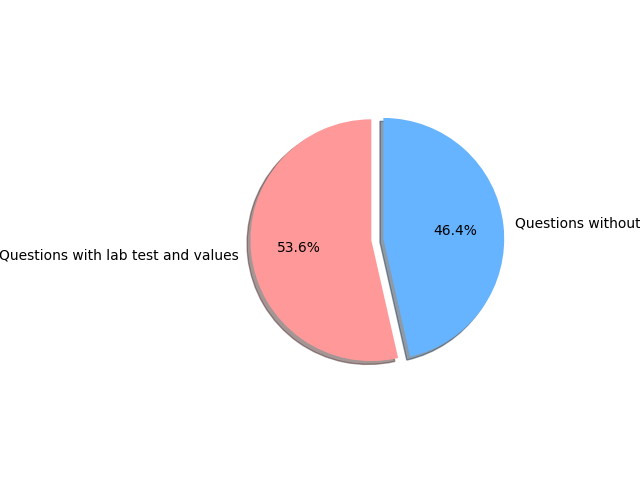


Figure 4.2 Distribution of questions containing lab test and values.

To prevent noise from impacting the accurateness of retrieval, we filtered questions by applying a length restriction after merging question titles with their corresponding contents. In this study, we consider questions with a length between 5 words and 20 words, both inclusive. After applying such length restriction, we obtained a total of 13,952 questions. The distribution of question lengths is shown in Figure 4.3.

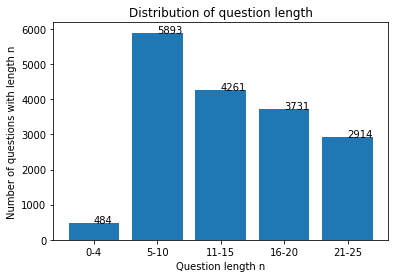


Figure 4.3 Distribution of question lengths

### **4.2.2 Identifying Lab Test Related Questions**

In this study, we focused on the three lab tests that are most relevant to diabetes diagnosis and management: creatinine, HbA1c, and blood glucose test. Both the type of lab test and the range of the corresponding numeric results were treated as two engineered features. Below is a demonstration of how Valx was used to process the sample questions and extract test-value statements.

The variable-extract algorithm of Valx takes five input parameters: a .csv input data file, a .csv domain-knowledge file, a .csv UMLS semantic knowledge file, the output directory, and a lab test variable. In short, Valx processes questions and extract test-value comparison statements by following the seven steps: 1) preprocessing of the input textual data, 2) extraction of numeric, units, and comparison operators, 3) variable identification using hybrid knowledge, 4) variable-numeric association, 5) context-based association filtering, 6) normalization of measurement units, and 7) heuristic rule-based evaluation [38].

For this study, Valx was modified to process three lab-test variables and return test-value statement for each variable. A complete test-value comparison statement consists of the following components: a lab test variable, a comparison logic operator (e.g. *“=”, “>”, “<”, “>=”, “<=”*), a threshold or a value, and a measurement unit. Because test values in many sample questions did not contain a measurement unit, the last component is often void.

Table 4.1 below presents three examples for the extracting the test-value statement of each lab test:

Table 4.1 Examples for the extracting the test-value statement of each lab test

|  |  |  |
| --- | --- | --- |
| ID | Question | Test-value statement |
| 1856600 | *“Fasting blood sugar is 130. am i diabetic? Im 30 years old female”* | ['Glucose', '=', '130'] |
| 3663888 | *“is an A1c level of 8.0 bad?”* | ['HBA1C,', '=', '8.0,'] |
| 1330863 | *“I am female. Age 55years. My S creatinine 5.1 what can I do?”* | ['Creatinine', '=', '5.1'] |

We selected 305 questions that meet the question length criteria and have at least one of the three lab tests and randomly sampled the 2,695 other questions without lab tests, which results in a pool of 3,000 questions. Among these questions, 246 questions contain glucose test results, 32 questions contain HbA1c test results, and 25 questions contain creatinine test results, comprising 8.2%, 1.07%, and 0.83% of the pool, respectively.

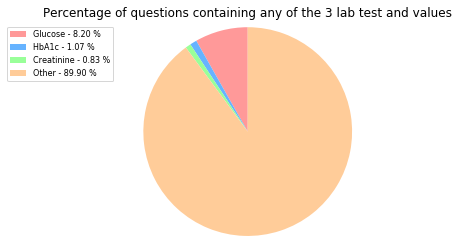


Figure 4.4 Distribution of sample questions containing any of the 3 targeted lab tests.

The three types of lab tests were treated as three binary (feature value is either 0 or 1) engineered features, respectively. e.g. if a question contains test-value statements for glucose but not the other two lab tests. The value of the glucose test feature would be 1 and the other two 0. To convert the extracted numeric lab-test values into ranges, we referenced the official websites of Mayo Clinic and National Institute of Health [39-42], and documented the three-level ranges (normal, prediabetic, diabetic for glucose and a1c test, and low, normal, high for creatinine test) for the three targeted lab tests. The ranges for each lab test are shown as Table 4.2, Table 4.3 and Table 4.4 below.

Table 4.2 Ranges for creatinine lab test

|  |  |
| --- | --- |
| Low | Below 0.84 mg/dL  (or below 74.3 mmol/L) |
| Normal | 0.84 to 1.21 mg/dL  (or 74.3 to 107 mmol/L) |
| High | Above 1.21 mg/dL  (or above 107 mmol/L) |

Table 4.3 Ranges for glucose lab test

|  |  |
| --- | --- |
| Normal | Below 100 mg/dL  (or below 5.6 mmol/L) |
| Pre-diabetic | 100 to 125 mg/dL  (or 5.6 to 6.9 mmol/L) |
| Diabetic | Above 126 mg/dL  (or above 7 mmol/L) |

Table 4.4 Ranges for HbA1c lab test

|  |  |
| --- | --- |
| Normal | Below 5.7 percent |
| Pre-diabetic | 5.7 to 6.4 percent |
| Diabetic | 6.5 percent or above |

Lab test values that fall into the level I range will be assigned a feature value of 1, values in level II range will be assigned 2, and 3 for level III range. The distributions of ranges for each of the three lab tests are as shown in figure 4.5, 4.6, and 4.7 below.

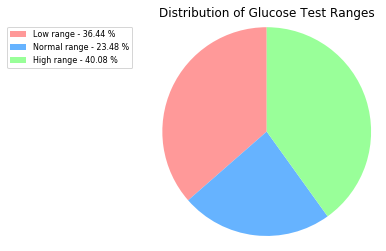


Figure 4.5 Distribution of test ranges for glucose lab test.

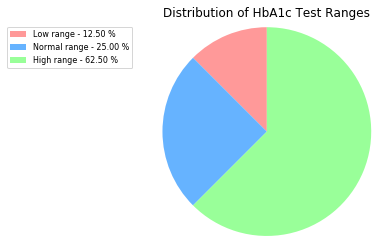


Figure 4.6 Distribution of test ranges for hemoglobin a1c lab test.

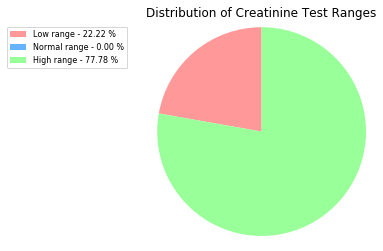


Figure 4.7 Distribution of test ranges for creatinine lab test.

### **4.2.3 Feature Extraction**

In addition to the type of lab test and the range of its corresponding value, we determined and extracted three additional engineered features after a series of brainstorming and testing. These features include: the sentence length of a question, number of stop words, WH question type. The distributions of question length and stop-word count are show in Figure 4.8 and 4.9, respectively.

**Sentence Length**

As previously mentioned in section 4.3.1, our sampled questions have a length between five and twenty words, both inclusive. As an engineered feature, sentence length helps identify the number of words contained in a sentence, which allows us to judge whether two questions are similar by comparing the difference between the number of words they contain. Distribution of sentence length of the question pool is shown as Figure 4.8, with 1,117 samples with lengths of 6 to 10 words, 958 samples with lengths of 11 to 15 words, and 925 samples containing the length of 16 to 20 words.

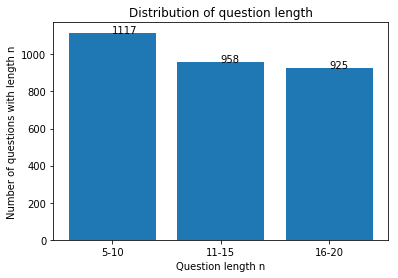


Figure 4.8 Distribution of sentence length of the sample questions.

**Stop-Word Count**

Stop words are commonly-used, short function words (e.g. “the”, “a”, “an”, “in”). These words take up space and reduce the processing time of search engines. The removal of these words in turn accelerates the processing speed and improves phrase search results. The *stop-word count* feature helps to examine question similarity by enabling us to compare the difference of the number of stop words contained in each question. For this feature, the stop word corpus of the natural language processing toolkit (NLTK) was used as a reference [43]. The distribution of the number of stop word in the question pool is shown in Figure 4.9. Overall, stop-word count of the question pool is normally distributed. Only one question contains the maximum count of stop words, which is 12. 80 questions contain no stop words at all. The most frequent count of stop words is 3 and 4, with 562 questions containing 3 stop words and 549 questions containing 4 stop words.

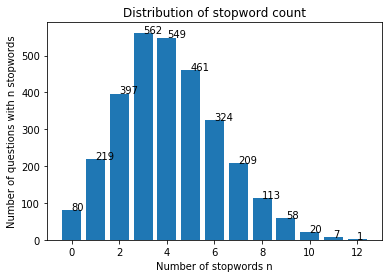


Figure 4.9 Distribution of stop-word count.

**WH Question Type**

WH questions are question words that begin with letter *wh*, except for question word *how*. This feature helps us examining question similarity by identifying whether a sample contains a WH question. In this study, the four most frequently-occurred and relevant question types were selected: *what, how, when,* and *why*. The distribution of the WH question types is shown in Figure 4.10. With *what* questions being most frequent (19.85%), *how* questions second most frequent (9.95%), *why* questions third (2.31%), and *when* questions last (2.31%).

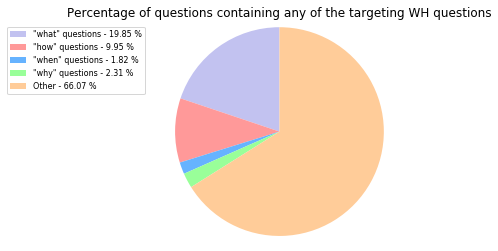
****

Figure 4.10 Distribution of the targeted WH question types.

### **4.2.4 Vector Representations**

For the sake of accuracy in calculating the vector-space weighted similarity, a series of pre-processing procedures were initiated during the process of text vectorization. First of all, we tokenized each question in our pool and removed stop words. In addition, words stemming was conducted using the *PorterStemmer* module of the Natural Language Processing Toolkit (NLTK). In information retrieval, the goal of stemming is to reduce derived or inflected words to their word stem, or root form. e.g. here are some examples of the root word like: *likes*, *liked*, *likely*, *liking,* etc.

For testing purposes, we selected 4 groups of 10 query questions for each of the 3 lab tests and a group of general questions that do not contain any lab tests. For questions mentioning any of the 3 lab tests, it must contain a corresponding numeric value. For general questions, it is ensured that no lab tests are included in the question content. Therefore, we take consideration of each engineered features: sentence, stop word count, WH question types, and lab ranges. Our goal in this phase is to ensure that each group of 10 query questions maintain diverse values for each of these features. For example, for the 10 query questions of glucose test, with all query questions containing glucose test and test ranges, the maximum sentence length was 20 words, whereas the minimum was 6 words. The number of stop words in this group of query questions ranges from 2 to 11. And half of the questions contain a WH question type (1 *what* question, 3 *how* questions, and 1 *why* question).

After the query questions are selected, the next step is to test the QA system: we take one of the query questions as the input and the algorithm returns relevant posts to it. In web information retrieval, vector space models are frequently adopted to represent text documents as vectors. And with those vectors, we were able to calculate the similarity between the query vector and the document vectors.

**TF-IDF**

First, the Term Frequency-Inverse Document Frequency (TF-IDF) was applied as our baseline model. The TF-IDF weight is a statistic measure for evaluating the degree of importance of a word to a document in the corpus. As shown in Equation 1, a typically TF-IDF weight consists of 2 components: Term Frequency (TF), and the Inverse Document Frequency (IDF). As illustrated in Equation 2, TF is the number of occurrences of a term *t* in a document *d* divided by the total number of terms in the document, which measures the frequency of a term occurring in a document. As presented in Equation 3, IDF is the logarithm of the number of documents in corpus over the number of documents with the term in it. It measures how important a term is. Since TF treats all terms in a document equally important, IDF weighs down those frequent terms that have little importance, such as “is”, “of”, “that” while scaling up the rare ones. For the Term Discrimination considerations, the product of TF and IDF is a legitimate choice in assessing term importance [44, 45].

Equation 1

Equation 2

Equation 3

As TF-IDF does not consider the semantic sensitivity of text, on top of TF-IDF, two word-embeddings, ELMo and the Universal Sentence Encoder [1], were applied for assessing the semantic and syntactic similarity between the query question and the sample questions in each category. The pre-trained encoding models are publicly available on Tensorflow Hub. Word embeddings are a set of language models that maps words in the corpus to vectors of real numbers. Unlike TF-IDF, embeddings can be implemented for semantic and syntactic parsing, which make them a crucial complement to our baseline model.

**ELMo**

ELMo stands for Embeddings from Language Models [46]. It is a set of deep contextualized vector representations derived from bi-directional language models (biLMs), which take an entire sentence as input and is pre-trained on a large corpus of texts. Peters et al. implemented ELMo to their baseline model, an advanced version of Bi-directional Attention Flow model, the resulting F1 score on the Stanford Question Answering Dataset (SQuAD) [47] improved by 4.7% from 81.1% to 85.8%, and a 24.9% relative error reduction over the baseline was achieved [46].

The limitations of the traditional word embeddings were that they could not capture the semantic and syntax of words, and can hardly represent the different meanings of a word depending on the context. ELMo solves the two limitations in that it produces embedding of a given word by modeling the context of the word to its surrounding words.

As shown in Equation (4), the biLMs consist of a forward language model and a backward language model. With a sequence of N tokens, (t1, t2, ..., tN). As Equation (5) presents, a forward LM computes the probability of the sequence by calculating the probability of token tk based on all previous tokens (t1, ..., tk−1):

Equation 4

Equation 5

As presented in Equation (6), a backward LM functions the same way as the forward LM except it scans through the sequence in reverse order to predict the previous token based on the future context-dependent tokens:

In a biLM, both the forward and the backward LM were combined to jointly maximize the log probability of the forward and backward directions.

Equation 6

ELMo combines the intermediate layer representations in the biLM. For each token , an L-layer biLM computes a set of 2L + 1 representations as Equation (7) below:

Equation 7

Where is the token layer and (output representation of forward LSTMs and backward LSTMs), for each biLSTM layer. To be included in downstream models, all layers of representations R in ELMo were collapsed into a single vector: . In general, a task-specific weighting of all biLM layers can be computed as Equation (8) below:

Equation 8

Where stask are softmax-normalized weights and parameter γtask enables the task model to scale the entire ELMo vector. γ functions to aid the optimization process.

As most supervised models share a common architecture at lower layers, ELMo could be easily implemented into this paradigm. When implementing ELMo to the supervised model, the weights of the biLM are first frozen and then the ELMo vector is concatenated with . Then pass the ELMo-enhanced representation into the task RNN.

Hence through the adoption of ELMo, we believe that the extracted semantic similarity would be an important factor in evaluating the likeness between our test questions and the candidate questions.

**Universal Sentence Encoder**

The universal sentence encoder is a sentence-based embedding model for encoding sentences into vectors. It consists of two encoders: One based on the transformer architecture [48] and affords greater model complexity and resource consumption to achieve high accuracy. The other makes use of a deep averaging network [49] and reduces accuracy for more efficient inference.

The first encoder is a general-purpose, transformer-based sentence encoding model that employs the encoding sub-graph of the transformer architecture [48]. The sub-graph was adopted to identify all other words and their ordering in order to produce a contextual representation of a word in a sentence. Then the model computes the element-wise sum of the representations at each word position to convert word representations to a fixed-length sentence encoding vector. The input of the encoder is a lowercased PTB tokenized string. And the output is a sentence embedding of a 512-dimensional vector.

The second encoding model utilizes a deep averaging network (DAN). To produce embeddings at the sentence level, DAN first averages input embeddings for words and bi-grams and then takes them as input to a feedforward deep neural network (DNN). The input and output of the DAN encoder are the same as the transformer-based encoder. Multi-task learning is employed, which enables a single DAN encoder can produce sentence embeddings for multiple downstream tasks. The DAN encoder demonstrated an advantage in that it takes linear computation time to compute the length of the input sequence.

### **4.2.5 Two Similarity Calculations**

**Engineered Features**

Table 4.5 below displays the features and their values of the example question: *is a 6.0 a1c average over a 10 year period very good control?* We noticed that except for WH Question Type and lab test mentions, the rest of the features are neither binary features nor between 0 and 1. Hence we adopted min-max normalization shown in Equation (9) to scale these features between 0 and 1.

Table 4.5 List of engineered features and their values of an example question

|  |  |  |
| --- | --- | --- |
| # | Feature | Value |
| 1 | Sentence Length | 13 |
| 2 | Stop-Word Count | 4 |
| 3 | WH Question Type | 0 |
| 4 | Lab Test (HbA1c) | 1 |
| 5 | Test Range | 1 |

Equation 9

Despite the min-max normalization strategy successfully converted feature values to a number between 0 and 1, the normalized feature value of question did not indicate its distance to the feature value of another question. Therefore, it is still hard to tell the degree of difference of these features between a query question and a candidate question. Hence we adopted the formula *1-abs(A – B)* to compute feature distances (abs stands for absolute value), where *A* is the min-max normalized feature value of a query question, and *B* is the normalized feature value of a candidate question. The feature distances between query question *“my a1c is 5.4 do i have pre-diabetes? i am anemic” and “my a1c is 5.4 do i have pre-diabetes* and candidate question *is a 6.0 a1c average over a 10 year period very good control?”* is shown in Table 4.6 below.

Table 4.6 List of engineered features of an example question and their distances to query question

|  |  |  |
| --- | --- | --- |
| # | Feature | Distance |
| 1 | Sentence Length | 0.533333333 |
| 2 | Stop-Word Count | 0.692307692 |
| 3 | WH Question Type | 0 |
| 6 | Lab Test (HbA1c) | 1 |
| 7 | Test Range | 0.666666667 |

**Vector Representations**

For measuring the level of similarity between two word-vectors, cosine similarity was used. As shown in Equation (10), for any two vectors that have the same values, their cosine similarity would be one, meaning that the two vectors are extremely similar. The cosine similarity measure is generally used in positive space, and the outcome is bounded in [0, 1]. From the three vector representations of the corpus, we measured the cosine similarity S(A, B) between the vector representation of the query question A and each candidate question B in every category.

Equation 10

### **4.2.6 Question Ranking**

We further assigned weight parameter to the similarity measures in the three vector space models as well as other engineered features for both the query question and the candidate questions. The best choice of weights was determined after extensive experiments. By summing the product of each feature value and its corresponding weight, a final value was obtained. The final value of the query question would be the highest and approximately 1, as it is comparing to itself. The candidate questions were ranked based on their final values from largest to smallest. The ranked results were then evaluated by human annotators. The formula for calculating the final similarity score would be as shown in Equation (11). The weighting scheme for questions containing glucose test results is presented in Table 4.7 below. For example, the feature distances of candidate question *“My fasting blood sugar was 157, does that mean Im diabetic? ?”* to query question *“Fasting blood sugar is 130. am i diabetic? Im 30 years old female”* is shown as Table 4.8. Therefore, the computation of the final similarity score is as presented in table 4.9.

Equation 11

Table 4.7 Weighting Scheme for questions containing glucose test results

|  |  |
| --- | --- |
| Feature | Weight |
| TF-IDF | 0.14 |
| ELMo | 0.21 |
| Universal Sentence Encoder | 0.21 |
| Sentence Length | 0.08 |
| Stop-word Count | 0.05 |
| WH Question Type | 0.05 |
| Lab Test (Glucose) | 0.1 |
| Test Range | 0.16 |

Table 4.8 Feature Distances for the sample candidate question

|  |  |
| --- | --- |
| Feature | Distance |
| TF-IDF | 0.3003 |
| ELMo | 0.7361 |
| Universal Sentence Encoder | 0.7520 |
| Sentence Length | 0.9333 |
| Stop-word Count | 1 |
| WH Question Type | 0 |
| Lab Test (Glucose) | 1 |
| Test Range | 1 |

Table 4.9 Computation of the final similarity score of between the query question and the sample candidate

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TF-IDF | ELMo | Universal Sentence Encoder | Sentence Length | Stop-word Count | WH Question Type | Lab Test (Glucose) | Test Range | Final Similarity Score |
| 0.3003\*0.14 | 0.7361\*0.21 | 0.7520\*0.21 | 0.9333\*0.08 | 1\*0.05 | 0\*0.05 | 1\*0.1 | 1\*0.16 | 0.7392 |

### **4.2.7 Evaluation**

In order to measure the ad hoc information retrieval effectiveness, a test collection was created. It consists of three components: 1. The collection of top-10 ranked results among all ranked candidate questions for 4 groups (creatinine, glucose, HbA1c, and without lab tests) using the 3 models: 1) our algorithm, 2) ELMo, and 3) TF-IDF, which includes the last two model but also augmented the Universal Sentence Encoder and engineered features; 2. A test suite of information needs, which are the query questions; and 3. A guideline of relevance judgment, which contains a rating criterion that gives each candidate question a score in the range of 0-5, both inclusive (as shown in Table 4.10). For objectivity purposes, the ranked candidate questions were independently rated by three raters, who had no knowledge about the construction of our algorithm. The results of the 3 raters’ rating were then averaged to form a gold standard.

Table 4.10 Meaning of each rating score

|  |  |
| --- | --- |
| Score | Meaning |
| 0 | Two questions consist of different words and have nothing in common |
| 1 | Two questions have similar words but convey different meanings |
| 2 | Two questions have similar words and meanings but different information needs |
| 3 | Two questions contain similar words and information needs. |
| 4 | Two questions contain the same range of lab results and information needs, but have similar sentence length and words. |
| 5 | Two questions contain the same range of lab results, convey similar information needs, and have similar sentence length and words. |

Since most of the features reflect non-binary notions of relevance, we employed the *normalize Discounted Cumulative gain* (nDCG) to the gold-standard rating to evaluate the precision of the ranked retrieval results. As Equation (12) presents, the nDCG score at position *p* is computed by having the Discounted Cumulative Gain (DCG) divided by the Ideal Discounted Cumulative Gain (IDCG). As shown in Equation (14), DCG score is calculated by having , the relevance score at index *i* divided by the logarithm of (*i* + 1) and then adding , the relevance scores of all previous indices. In our case, the relevance scores are the gold standard rating of the candidate questions at each index. As displayed in Equation (13), IDCG is computed the same as DCG, except that of IDCG represents the list of relevant questions ordered by their relevance score (highest to lowest) up to position *p*.

Equation 12

Equation 13

Equation 14

For each query question in every group, the top 10 candidate questions retrieved using each of the three models were combined. If a candidate question was ranked as top 10 by multiple models but was rated with different scores, the gold standard ratings of the question were then averaged and duplicates were removed. The remaining candidates were then ranked by their scores of each model (our algorithm, ELMo, Universal Sentence Encoder, TF-IDF). In addition, the nDCG results based on a variation of our algorithm was also added to the comparison. Three engineered features – sentence length, stop-word count, and WH question types were removed in this model.

## **4.3 Results**

### **4.3.1 Evaluation of Questions Contain Creatinine Lab Test**

As shown in Figure 4.11 below, all other models significantly outperformed the baseline TF-IDF. The nDCG results of our model outperformed the results of others only in position 1. Starting from position 2, our model without engineered features outperformed all other models. The cause of such a scenario was because our model took into accounts the creatinine test and the range of the corresponding numeric results during reranking. Overall, the universal sentence encoder was second optimal. At position 3, the results of the universal sentence encoder outperformed all models except for our model without engineered features. For the two sentence-embedding models, the universal sentence encoder outperformed ELMo at position 3 to 10.

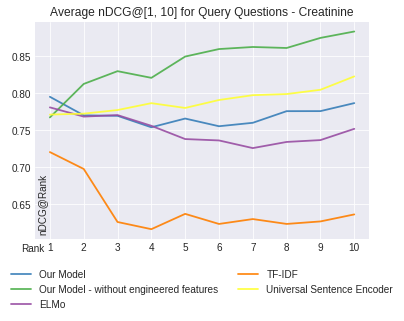


Figure 4.11 Average nDCG@[1, 10] using 5 models for creatinine query questions

### **4.3.2 Evaluation of Questions Contain Glucose Lab Test**

As shown in Figure 4.12 below, the nDCG results of universal sentence encoder at all positions outperformed the result of all other models, except it was excelled by our model without the 3 engineered features at position 1. Both our original model and the model without engineered features did not have optimal performance. Such a scenario was caused by various factors: First, Valx used diverse variable names to represent glucose test. e.g. ‘Glucose’, ‘blood glucose measurement’, ‘plasma glucose result’, ‘mean Glucose’, ‘serum Glucose level’, ‘Glucose tolerance’ etc. The extracted numeric lab results were accurate for some of the variables, and inaccurate for others. For example, for the question *what is normal blood sugar level for 15 yr old female after eating?* The test-value statement was *['blood\_glucose\_measurement,', '=', '15,']*. Valx incorrectly extracted 15 as the numeric results. Such a scenario in turn hinders the accuracy of the ranges of glucose test results, which impaired the performance of our model. However, for the question *I have FASTING blood sugar 130. am i m diagnosed with diabetes?* Valx correctly extracted test-value statement as ['Glucose,', '=', '130,']. Additionally, our model combines the weighted similarity of the other three models, hence its performance excelled the baseline TF-IDF while being outperformed by the optimal universal sentence encoder. For future investigations, the accuracy of extracting glucose test-value statements could be enhanced by improving Valx’ parsing algorithm, in particular, the mapping between glucose test variables and the corresponding numeric results.

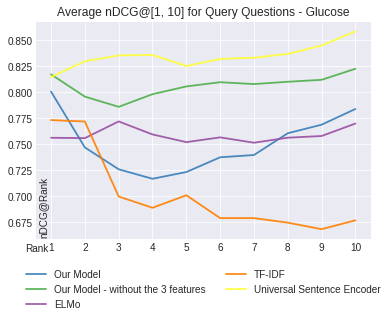


Figure 4.12 Average nDCG@[1, 10] using 5 models for glucose query questions

### **4.3.3 Evaluation of Questions Containing HbA1c Lab Test**

As shown in Figure 4.13 below, the nDCG results of our model without engineered features outperformed others at all positions except for 3 and 4, while our original model outperformed ELMo only at position 7 to 10. Nonetheless, both the original model and the model without engineered feature excelled the universal sentence encoder and significantly outperformed the baseline TF-IDF. Such result was partly due to the fact that Valx was highly accurate in extracting test-value statements of HbA1c. As for the two sentence-embedding models, the nDCG results of ELMo outperformed the results of the universal sentence encoder at all positions.

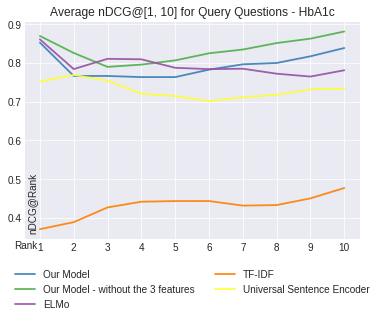


Figure 4.13 Average nDCG@[1, 10] using 5 models for HbA1c query questions

### **4.3.4 Evaluation of Questions Without Lab Tests**

As shown in Figure 4.14, the nDCG results of our model without engineered features excelled all other models at all positions except it was outperformed by TF-IDF at position 3. The results of the baseline TF-IDF was second optimal, as it outperformed the rest of the models at all positions except it was excelled by the universal sentence encoder at position 7. In contrast, our original model outperformed the universal sentence encoder at position 1 to 3, and ELMo at position 1 and 2. Such a result reveals that the manually engineered features were not optimal and could be improved by adopting neural networks to learn high-level features from data.

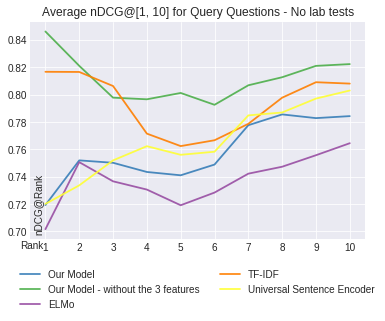


Figure 4.14 Average nDCG@[1, 10] using 5 models for query questions without lab tests

## **4.4 Discussion**

This study is an extension of our previous qualitative assessment in identifying patients’ information needs about their laboratory test results. In this study, we developed a question retrieval algorithm that takes an input question and retrieves ranked similar previously-answered questions from the question pool. Existing work in the field has not considered the inclusion of clinical lab test and test ranges during the ranking of the candidate questions. To fill this research gap, we proposed this question-retrieval algorithm. Building upon prior work, our algorithm not only retrieves similar questions; but on top of the retrieval, the algorithm takes into accounts two novel engineered features, the type of lab test and the range of the test value during the process of ranking candidate questions.

As for the reranking results, our model was outperformed when without the three engineered features. This was due to the following factors: the manually engineered features were not long to see the difference between a query question and a candidate. Because of that, the semantic similarity became more important in examining question similarity. Additionally, reranking by the USE-weighted similarity score demonstrated better nDCG results than reranking by the ELMo-weighted similarity for a variety of reasons. First, as shown in Perone et al., compared to other existing embedding models, the USE has demonstrated an optimal performance in all semantic textual similarity tasks except for semantic relatedness. Also, ELMo maintains an embedding size of 3,072 dimensions among all layers, whereas the USE only has an embedding size of 512 dimensions for the transformer-based encoders and deep averaging network (DAN), respectively. ELMo combines the 3 output layers from the more syntactic at the lowest layers to the more topical and polysemous at the highest layer, which potentially causes a lower semantic similarity between the query question and the retrieved candidates. In other words, in a high dimensional space, the same collection of words (e.g. basketball and hockey) may have a relatively lower cosine similarity as compared to a lower-dimensional space.

We originally planned to compare the retrieval results of our algorithm, the results of the baseline TF-IDF, the results of ELMo, and Universal Sentence Encoder; with the top-10 retrieved results from Yahoo! Answers. However, for most of our selected query questions, we were not able to retrieve a similar question from Yahoo! Answers. A trivial amount of our query questions was able to retrieve similar questions through Yahoo! Answers, but the number of these retrieved questions was so small, which prevented us from making a comparison to it.

## **4.5 Summary**

In this study, we developed a question-retrieval algorithm that considers the range of the lab test in the query and candidate questions during reranking. The ranked results of our algorithm were evaluated by human raters alone with the results of each model that we adopted, eg. TF-IDF (baseline), ELMo and Universal Sentence Encoder (two sentence-embedding models). The ratings were then averaged to form a gold standard, which was adopted for the evaluation of our algorithm using nDCG. A comparison between the nDCG results of each model was initiated. This study is a proof of concept that aimed to take into account the range of the lab test mentioned in question during reranking. Judging by the top 10 retrieved results of our algorithm in each group, our algorithm has succeeded in fulfilling such objective.

# **Chapter 5. Discussions, Conclusions, and Future Work**

In this thesis, we conducted an analysis to discover diabetic patients’ information needs about their lab test results and developed a question-retrieval algorithm that supports these needs through two studies:

Study 1 explored lay people’s various information needs related to lab results through analyzing forum posts collected from a social Q&A site. Our results highlighted the need to address the gap between patient knowledge and limited contextual information presented on their lab reports, and to improve the design of patient portals to fully meet patient needs in understanding the lab results. Our findings provide a foundation for our future work, including qualitative studies (e.g., interview with clinicians and patients) and analysis of medical record data to understand how to best provide personalized information and present clinical data in patient portals.

This study has several limitations. First, it only focused on one disease, namely diabetes, one type of health information, laboratory test results, and one health forum, Yahoo! Answers. While our findings pertain to the characteristics of the specific domain, the results may not be generalizable to other types of diseases and types of health information. Our future work will expand to other health conditions (e.g., cancers), other health forums (e.g., eHealth.com, healthboard.com), and include other types of health information (e.g., radiology report, physician notes, discharge summaries) to assess the generalizability of our findings. Second, we only analyzed the question posts and therefore did not discuss how those questions were answered on this Q&A site. Future work will synthesize the types of information people gain from online communities and how these answers were constructed to meet their needs. Lastly, due to various constraints, we did not collect posters’ demographic data, such as level of disease severity, gender, age, and different stages of life/illness trajectory. Future work will take these factors into consideration.

In Study 2, a question-retrieval algorithm was developed to support patients’ understanding of their lab test results. In particular, the study focused on the following three clinical lab tests that are of great concerns to the diabetes community: glucose test, HbA1c test, and creatinine test. The comparison of nDCG result between the algorithm and existing sentence embeddings demonstrated that it is feasible to implement the algorithm to extract questions containing lab test results. The system is also proven to be well-functioning for questions without any lab test mentions.

However, certain limitations still exist. Due to the explorative nature of the project, the study only focused on questions containing one of the three above-mentioned lab tests; alone with questions without clinical lab tests. For WH questions, the ranking results would be more accurate if each of the four WH question types was considered as a feature separately. Additionally, Valx’s extraction of clinical lab-value statement demonstrates high accuracy in HbA1c test, but its extraction of test-value statements for Glucose and Creatinine tests still contain certain noise. In particular, glucose test has a variety of types and its expressions are not as straight forward. Hence some non-lab numeric values were incorrectly mapped to those expressions. e.g. some non-lab-result numbers were incorrectly identified as lab values and were mapped to the lab test mentioned in the question alone with some impossible units, such as *hours, weeks, years, months* etc. The normalization of these values could lose a certain level of accuracy in the ranking. To assure the accurateness of ranking, postprocessing was conducted to remove those impossible units.

Future work aims to improve the issues listed above. More importantly, the algorithm will be more generalizable to other types of clinical data, e.g. blood pressure test, radiology reports, and would also include the evaluation of more contextual consumer-health information presented in patients’ questions posted on social Q&A sites. Instead of using the manually engineered features, neural networks will be applied for feature learning and weight assignment in order to achieve better performance in the health question retrieval tasks.

# **References**

1. *What is a patient portal?* ; Available from: <https://www.healthit.gov/faq/what-patient-portal>.

2. Graham, S., & Brookey, J, *Do Patients Understand?* Perm J, 2008. 12(3): p. 67-69.

3. Hong, M.K., et al., *Supporting Families in Reviewing and Communicating about Radiology Imaging Studies*, in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, ACM: Denver, Colorado, USA. p. 5245-5256.

4. Reynolds, T.L., et al., *Understanding Patient Questions about their Medical Records in an Online Health Forum: Opportunity for Patient Portal Design.* AMIA Annu Symp Proc, 2017. 2017: p. 1468-1477.

5. Alpert, J.M., et al., *Applying Multiple Methods to Comprehensively Evaluate a Patient Portal's Effectiveness to Convey Information to Patients.* J Med Internet Res, 2016. 18(5): p. e112.

6. Fraccaro, P., et al., *The influence of patient portals on users' decision making is insufficiently investigated: A systematic methodological review.* Int J Med Inform, 2018. 111: p. 100-111.

7. Zickuhr, K., *Generations 2010.* Pew Internet and American Life Project, 2010.

8. Pillemer, F., et al., *Direct Release of Test Results to Patients Increases Patient Engagement and Utilization of Care.* PLoS One, 2016. 11(6): p. e0154743.

9. Gee, P.M., et al., *e-Patients Perceptions of Using Personal Health Records for Self-management Support of Chronic Illness.* Comput Inform Nurs, 2015. 33(6): p. 229-37.

10. Giardina, T.D., et al., *Patient perceptions of receiving test results via online portals: a mixed-methods study.* J Am Med Inform Assoc, 2018. 25(4): p. 440-446.

11. Zhang, Y., *Contextualizing consumer health information searching: an analysis of questions in a social Q&#38;A community*, in *Proceedings of the 1st ACM International Health Informatics Symposium*. 2010, ACM: Arlington, Virginia, USA. p. 210-219.

12. Luo, J., et al., *SimQ: real-time retrieval of similar consumer health questions.* J Med Internet Res, 2015. 17(2): p. e43.

13. Shtok, A.D., G.; Maarek, Y.; Szpektor, I., *Learning from the Past Answering New Questions with Past Answers.* WWW, 2012.

14. Wongchaisuwat, P., D. Klabjan, and S.R. Jonnalagadda, *A Semi-Supervised Learning Approach to Enhance Health Care Community-Based Question Answering: A Case Study in Alcoholism.* JMIR Med Inform, 2016. 4(3): p. e24.

15. He, Z., et al., *Enriching consumer health vocabulary through mining a social Q&A site: A similarity-based approach.* J Biomed Inform, 2017. 69: p. 75-85.

16. Organization, W.H., *Global report on diabetes.* 2016.

17. Zhang, Z., et al., *Making Sense of Clinical Laboratory Results: An Analysis of Questions and Replies in a Social Q&A Community.* The 17th World Congress on Medical and Health Informatics (MEDINFO 2019), 2019.

18. Vieder, J.N., et al., *Physician-patient interaction: what do elders want?* J Am Osteopath Assoc, 2002. 102(2): p. 73-8.

19. Eastin, M.S. and N.M. Guinsler, *Worried and wired: effects of health anxiety on information-seeking and health care utilization behaviors.* Cyberpsychol Behav, 2006. 9(4): p. 494-8.

20. Longo, D.R., et al., *Health information seeking, receipt, and use in diabetes self-management.* Ann Fam Med, 2010. 8(4): p. 334-40.

21. Zhang, K., et al., *Question Retrieval with High Quality Answers in Community Question Answering*, in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 2014, ACM: Shanghai, China. p. 371-380.

22. Zhou, G., et al., *Learning Continuous Word Embedding with Metadata for Question Retrieval in Community Question Answering.* ACL, 2015. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers): p. 250-259.

23. Mikolov, T., et al., *Distributed representations of words and phrases and their compositionality*, in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. 2013, Curran Associates Inc.: Lake Tahoe, Nevada. p. 3111-3119.

24. Clinchant, S. and F. Perronnin, *Aggregating Continuous Word Embeddings for Information Retrieval.* Association for Computational Linguistics, 2013. Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality: p. 100-109.

25. Lei, T., et al., *Semi-supervised Question Retrieval with Gated Convolutions.* NAACL, 2016. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: p. 1279–1289.

26. Brokos, G.-I., P. Malakasiotis, and I. Androutsopoulos, *Using Centroids of Word Embeddings and Word Mover’s Distance for Biomedical Document Retrieval in Question Answering.* Association for Computational Linguistics, 2016. Proceedings of the 15th Workshop on Biomedical Natural Language Processing: p. 114-118.

27. Tsatsaronis, G., et al., *An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition.* BMC Bioinformatics, 2015. 16: p. 138.

28. Kusner, M.J., et al., *From word embeddings to document distances*, in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. 2015, JMLR.org: Lille, France. p. 957-966.

29. Palangi, H., et al., *Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval.* IEEE/ACM Trans. Audio, Speech and Lang. Proc., 2016. 24(4): p. 694-707.

30. Camacho-Collados, J. and M.T. Pilehvar, *From word to sense embeddings: a survey on vector representations of meaning.* J. Artif. Int. Res., 2018. 63(1): p. 743-788.

31. Cer, D., et al., *Universal Sentence Encoder.* arXiv:1803.11175, 2018.

32. Sacks, D.B., et al., *Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus.* Diabetes Care, 2011. 34(6): p. e61-99.

33. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* Biometrics, 1977. 33(1): p. 159-74.

34. Sarkar, U., et al., *The literacy divide: health literacy and the use of an internet-based patient portal in an integrated health system-results from the diabetes study of northern California (DISTANCE).* J Health Commun, 2010. 15 Suppl 2: p. 183-96.

35. Irizarry, T., A. DeVito Dabbs, and C.R. Curran, *Patient Portals and Patient Engagement: A State of the Science Review.* J Med Internet Res, 2015. 17(6): p. e148.

36. Park, M.S., et al., *Consumers' Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites.* JMIR Med Inform, 2016. 4(4): p. e41.

37. Kamel Ghalibaf, A., et al., *Design, Implementation, and Evaluation of Self-Describing Diabetes Medical Records: A Pilot Study.* JMIR Med Inform, 2017. 5(2): p. e10.

38. Hao, T., H. Liu, and C. Weng, *Valx: A System for Extracting and Structuring Numeric Lab Test Comparison Statements from Text.* Methods Inf Med, 2016. 55(3): p. 266-75.

39. *A1C test*. Available from: <https://www.mayoclinic.org/tests-procedures/a1c-test/about/pac-20384643>.

40. *Diabetes*. Available from: <https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451>.

41. *Creatinine test*. Available from: <https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646>.

42. NIH, *The A1C Test & Diabetes.*

43. *2. Accessing Text Corpora and Lexical Resources*. Available from: <https://www.nltk.org/book/ch02.html>.

44. Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval.* Information Processing & Management, 1988. Vol. 24(No. 5): p. 513-523.

45. Salton, G. and C.S. Yang, *On the Specification of Term Values in Automatic Indexing.* Journal of Documentation, 1973. 29(4): p. 351 - 372.

46. Peters, M., et al., *Deep contextualized word representations.* Association for Computational Linguistics, 2018. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers): p. 2227-2237.

47. Rajpurkar, P., et al., *SQuAD: 100,000+ Questions for Machine Comprehension of Text.* Association for Computational Linguistics, 2016. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.

48. Vaswani, A., et al., *Attention is All You Need.* Neural Information Processing Systems, 2017.

49. Iyyer, M., et al., *Deep Unordered Composition Rivals Syntactic Methods for Text Classification.* Association for Computational Linguistics, 2015. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers): p. 1681-1691.