July 2020

Computational Approaches to Assisting Patients' Medical Comprehension from Electronic Health Records

Jiaping Zheng

COMPUTATIONAL APPROACHES TO ASSISTING PATIENTS' MEDICAL COMPREHENSION FROM ELECTRONIC HEALTH RECORDS

A Dissertation Presented

by

JIAPING ZHENG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2020

College of Information and Computer Sciences

COMPUTATIONAL APPROACHES TO ASSISTING PATIENTS' MEDICAL COMPREHENSION FROM ELECTRONIC HEALTH RECORDS

A Dissertation Presented

by

JIAPING ZHENG

Approved as to style and content by:

Hong Yu, Chair

W. Bruce Croft, Member

Benjamin Marlin, Member

Feifan Liu, Member

James Allan, Chair of the Faculty
College of Information and Computer Sciences

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my advisor, Professor Hong Yu. Without her invaluable mentoring, this research would not have been possible. My thanks are also extended to Professor Bruce Croft, Ben Marlin, and Feifan Liu for their helpful input on this thesis.

I wish to thank the members in the lab that have provided support in my journey: Abhyuday, Alice, Bhanu, Elain, Emily, Frank, Jesse, Jinying, John, and Subhendu. Special thank you to Weisong who had to respond to many a request from me on technical issues. Thank you to the fantastic staff at CICS, especially Leeanne and Eileen.

Finally, I wish to thank my family for their unfailing support and encouragement throughout my study.

ABSTRACT

COMPUTATIONAL APPROACHES TO ASSISTING PATIENTS' MEDICAL COMPREHENSION FROM ELECTRONIC HEALTH RECORDS

MAY 2020

JIAPING ZHENG

B.Eng., NANKAI UNIVERSITY

M.Sc., UNIVERSITY OF MINNESOTA

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Hong Yu

Patient-centered care has been established as a fundamental approach to improve the quality of health care in a seminal report by the Institute of Medicine published at the start of the century. Improved access to health information and demand for greater transparency contributed to its move into the mainstream. Research has also demonstrated that actively involving patients in the management of their own health can lead to better outcomes, and potentially lower costs. However, despite the efforts in many areas of medicine to embrace patient-centered care, engaging patients is still considered a challenge. One of the barriers is the lack of effective tools to help patients understand their health conditions, options and their consequences.

Patient portals are now widely adopted by hospitals and other healthcare practices to provide patients with the capabilities to view their own Electronic Health Records. They are a rich resource of information for patients. However, the language in the

records are generally difficult for patients without training in medicine to understand. Furthermore, the amount of information can often be overwhelming as well. In this work, we propose computational approaches to foster patient engagement from three aspects by exploiting the rich information in the medical records.

First, we design a framework to automatically generate health literacy instruments to measure a patient's literacy levels. This framework exploits readily available large scale corpora to generate instruments in a commonly used test format. Second, we investigate methods that can determine the readability of complex documents such as health records. We propose to rank document readability, instead of assigning a grade level or a pre-defined difficulty category. Lastly, we examine the problem of finding targeted educational materials to facilitate patient comprehension of medical notes. We study methods to formulate effective queries from specialized and long clinical narratives. In addition, we propose a neural network based method to identify medical concepts that are important to patients.

The three aspects of this work address the issues of the overabundance and technical complexity of medical language in health records. We demonstrate that our approaches are effective with various experiments and evaluation metric.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Motivation

Patient-centered care has been established as a fundamental approach to improve the quality of health care in a seminal report by the Institute of Medicine published at the start of the century [132]. The report defined patient-centered care as "care that is respectful of and responsive to individual patient preferences, needs, and values" and that ensures "that patient values guide all clinical decisions". Improved access to health information and demand for greater transparency, among other factors, contributed to its move into the mainstream [36]. The patient-physician relationship has thus shifted from a paternalistic style, where the physicians would decide what was best for their patients, to a shared decision making model, in which the physician would engage the patient to participate in determining the optimal course of action. Research has also demonstrated that actively involving patients in the management of their own health can lead to better outcomes, and potentially lower costs [63, 13]. However, despite the efforts in many areas of medicine to embrace patient-centered care [103], engaging patients is still considered a challenge [10]. One of the barriers is the lack of effective tools to help patients understand their health conditions, options and their consequences.

Patient-centered care and shared decision making require more participation from the patients. Underlying these demands are assumptions of their literacy skills. A measurement of health literacy would facilitate patient engagement by tailoring communications to the capacity appropriate for the individual patients. In the clinical

practice setting, clinicians commonly overestimate the health literacy of their patients [11, 116]. Assessing the health literacy of a sample of patients can provide the clinician with information about his or her patients' average reading level, which then can be used as a guide in the selection and development of patient education materials.

Related to the health literacy demands on the patients, readability of the information provided to patients also plays an important role in enabling effective patient engagement. It is estimated in the National Assessment of Adult Literacy that the average American has a reading level between the 7th and 8th grade [101]. It is also reported in the same assessment that about 36% of the US population or 75 million Americans have basic or below basic health literacy. A recent study estimated that among uninsured adults in the US, only 8.6% possess the numeracy skills to make complex, informed health decisions, such as managing chronic diseases [138]. In fact, materials beyond the patients' reading abilities are widely reported in the literature [2, 73, 43]. There is a great need in presenting concise and easy-to-understand materials for more patients to benefit from these resources. Text readability has often been studied in the context of grade school education, where children's reading abilities need to match reading materials. Measuring the readability of complex documents in the health care domain is less explored.

Furthermore, patients now have unprecedented access to their own health information. According to the data from the Office of the National Coordinator for Health Information Technology, the percent of hospitals that enable patients to electronically view, download, and transmit their health information grew almost 7-fold between 2013 and 2015 [72]. In the same survey, 95% of hospitals in 2015 provided their patients with the ability to view their information. Access to such information can, besides enhancing patients' medical understanding and facilitating shared decision

making with physicians, provide clinically relevant benefits [174], including increased medication adherence [44].

The patients are also increasingly seeking assistance online for health related information [37, 46]. Well-known Information Retrieval (IR) techniques may work well if the patients can create effective queries from EHR notes. However, it can be overwhelming for patients without medical training to parse the opaque language in the notes, which in turn hinders their efforts to identify the key information and formulate effective queries. In these cases, they would benefit from a system that automatically identifies content important to patients, thereby reducing their information load.

1.2 Measuring Targeted Health Literacy

The development of validated health literacy tests is a laborious process, mostly relying on expert's selection of words or passages in an ad-hoc fashion. Current methods in use such as Rapid Estimate of Adult Literacy in Medicine (REALM) [41] and Test of Functional Health Literacy in Adults (TOFHLA) [136] require manual creation of test questions.

These methods are often based on a broad set of basic knowledge, without emphasis on any particular disease or condition. In order to better manage their health, patients may benefit from specific knowledge about their own conditions, especially chronic diseases. For instance, an understanding of A1c may be useful for a patient with diabetes, but less valuable for a patient with hypertension.

The existing instruments are also static in their test content, and often do not measure literacy beyond the lowest level required to function in the health care setting. As a patient becomes more knowledgeable, the tests can exhibit a ceiling effect, where literacy levels above a certain limit cannot be distinguished. A better testing method should be able to focus on a patient's specific literacy needs based on his or her health and can evolve with the patients' understanding of the relevant knowledge.

We propose to build a computationally inexpensive test framework that can be instantiated from a large corpus, specific to users' literacy needs. The test format itself is also easy and fast to administer to the end users, minimizing clinical workflow interruption if it is given to patients in a clinic.

1.3 Assessing Readability of Medical Documents

In document readability metrics, the widely used methods apply readability formulas that only take into account a limited number of factors such as sentence length and word length [55]. Moreover, they are usually developed for texts aimed for school education and not validated in the health domain. In this work, we first demonstrate that these grade level measures on complex health documents do not align with readers' perceived text difficulty. We collect anonymous reader's perceptions of text difficulty on English text for the general public on health topics and de-identified EHR notes. Using a variety of measures, we show that in the samples we collected, user perceptions of readability significantly differ from existing methods' estimations. We further propose a ranking based approach that can better predict document readability than the widely adopted methods. User ratings on difficulty of documents are collected to learn a ranking based model, employing features that do not rely on complex syntactic or semantic processing on the text. We show that this model performs comparably to human annotators and outperforms existing formulas significantly.

1.4 Facilitating Comprehension of Electronic Medical Record Notes

To facilitate patient comprehension of their own EHR notes, we devise methods to generate queries from EHR notes to retrieve education materials. Work has been done in domain-specific information retrieval to generate Boolean queries that are preferred by users [95]. Others assumed the user can select passages of interest from a long

document [106]. In biomedical information retrieval, WRAPIN requires indexing by a domain ontology [58].

One characteristic of the search behavior in consumer health that is distinct from other document-based information retrieval is the user's expertise level. In those scenarios, the users are typically professionals or experts in the field or are at least knowledgeable. For example, in patent retrieval, legal search, and academic literature search, the users are typically practitioners in the respective field, and likely to have professional knowledge. In our case, however, the patients in general do not have medical training.

In our work, we focus on generating queries for patients without the expertise in the domain. We explore topic modeling, filtering, and learning-based query generation methods. Our experiments show that identifying key concepts from EHR notes is crucial to effectively retrieve education materials. We collect expert annotations on relevant education articles from Medline Plus on de-identified medical notes. We show that machine learning models that are adapted using out of domain general English text on medical topics achieve highest performance.

Furthermore, we propose to identify medical concepts that are important to patients. EHR notes generally incorporate a comprehensive longitudinal description of patients' medical courses yet patients may care more about their immediate concerns. In patient support applications, providing comprehension assistance for all the concepts are likely to overwhelm them and may be unnecessary in the first place. Our aim is to develop an automatic system that can identify a small number of important medical concepts specific to a patient. These medical concepts can then be used to provide tailored interventions to improve EHR comprehension and disease management. We show that our identification method can outperform competitive baselines.

1.5 Summary

Patient-centered care, one of the foundations of high quality care, requires active engagement of patients in the health care process. EHRs are a rich resource for developing applications to engage patients and foster patient activation, thus holding a strong potential to enhance patient-centered care. Given the potential in better outcome and lower cost in engaging patients through EHR, we propose to develop innovative methods to foster patient activation by evaluating health literacy, assessing document readability, and retrieving education materials tailored to the patients.

1.6 Contributions

The contributions of this work include the following.

- A flexible framework that can dynamically generate targeted health literacy instruments for a specific domain.
- Empirical evidence that current readability measurement tools are inadequate at measuring users' perceived text difficulty.
- Method to measure complex document readability.
- Method to identify medical concepts that are important and tailored for patients.
- Linking targeted educational materials for patients based on their medical records.
- Improving patient EHR comprehension by incorporating tailored medical concepts.

1.7 Organization

The remaining chapters of this thesis is organized as follows. Chapter 2 proposes a health literacy framework to generate customized instruments. Chapter 3 describes

the our empirical evaluations of widely readability formulas, and proposes a ranking-based method to measure complex document readability. Chapter 4 proposes query generation approaches to retrieve educational materials for EHR narrative notes, and a method to identify medical concepts important to patients. Chapter 5 concludes this thesis by summarizing our approaches and findings, and discussing future work.

CHAPTER 2

MEASURING TARGETED HEALTH LITERACY

2.1 Introduction

The past few decades have seen a proliferation of health literacy instruments. Recent reviews have identified dozens of tools [48, 133, 3, 82, 71], ranging from general measurements to disease-, content- or population-specific ones. These instruments aim to measure a variety of skills necessary to function in the health care system. For example, one study [71] categorized 51 instruments based on 11 dimensions, including the ability to perform basic reading tasks, to communicate on health matters, and to derive meaning from sources of information, etc. The ability to understand information is one of the four skills of health literacy identified in a systematic review [156]. It is also one of the most measured skills in the instruments. Those that measure this skill are widely used in research.

Designing an instrument measuring reading ability, or print literacy, is a time- and effort-intensive process. It usually starts with experts curating passages of text or word lists, followed by psychometric validation and revision based on test results obtained from a sample population. Once validated, the instruments stay static. When a new scenario arises, the process has to be repeated to create a new instrument.

There are a few potential drawbacks of reusing instruments designed long in the past. First, language use patterns evolve over time. Health literacy, reading ability in particular, needs to adapt to these changes. Instruments that were designed from early text sources may be out of date when employed decades later. Although we are not aware of reports of this nature in the health literacy literature, researchers

working on general vocabulary estimation tools have seen the need to update old tests [122].

Moreover, the public's reading abilities may also change because of increased exposure to print material. Statistics of educational attainment show that the population is receiving more education. Degrees conferred at various post-secondary levels all rose more than 30% over the decade between 2004-05 and 2014-15 according to a recent US national report [155]. More exposure to advanced text material at or above college level may improve one's reading ability. Older instruments that tend to use low-grade-level text may struggle to distinguish readers proficient above the very basic level that's required to function in the health care system. This ceiling effect, many test takers obtaining perfect scores [163], can be more pronounced when such tests are administered to groups in the general population, reflecting that many were developed with convenience samples of patients in a health care setting. Therefore, they function well as screening tools to detect low health literacy, but may fail to properly separate advanced readers.

In this work, we propose a new task of measuring targeted health literacy, in which a patient's knowledge of a specific area is examined, as opposed to the general health literacy that tests patients' ability in broad settings without a focus on their individual needs. We also aim to develop a test framework that can be customized to a need on demand, and can measure skills beyond the basic level.

2.2 Related Work

We highlight a few instruments in this section that measure the individual skills and abilities of understanding written text. For a more complete review of instruments that measure both reading and other skills, we refer the reader to a recent review [71].

Numerous instruments have been developed to test health literacy since the 1990s. There are two frequently used instruments: the Rapid Estimate of Adult Literacy

in Medicine (REALM) [40] and the Test of Functional Health Literacy in Adults (TOFHLA) [136], with its shortened form S-TOFHLA [8].

REALM is a tool based on word pronunciation. A list of 66 common medical terms is organized into three columns according to the number of syllables and pronunciation difficulty. The administrator records the number of terms correctly pronounced by the test taker, and the raw count can be converted to one of four grade levels: 0–3, 4–6, 7–8, and 9 and above. Criterion validity of REALM is established with Wide Range Achievement Test-Revised (WRAT-R) and other tests in the general domain. Estimate of administration time is under three minutes, making it easy to fit in a busy clinical workflow.

TOFHLA is designed to measure patients' ability to read and understand what they commonly encounter in the health care setting. It consists of 17 numeracy items and 3 prose passages. The passages are drawn from actual materials a patient may need to read, including instructions for preparation for an upper gastrointestinal series, the patient "Rights and Responsibilities" section of a Medicaid application, and a standard informed consent form. They are converted to a Cloze test with 50 items. Total scores are divided into three levels: inadequate, marginal, and adequate. Those who score in the adequate range do well on these tasks, but may have some difficulty comprehending the more difficult tasks like determining financial eligibility and the informed consent document. TOFHLA's correlations with WRAT-R and REALM were tested to establish validity. TOFHLA takes up to 22 minutes to administer.

Aiming to reduce the administration time, TOFHLA was abridged to an abbreviated version, S-TOFHLA, which takes a maximum of 12 minutes [8]. Two passages with 36 items were selected from the full version. S-TOFHLA's validity is compared to the long version of the TOFHLA and the REALM.

Since the publication of REALM and TOFHLA, many new instruments were derived from them, for different use cases. They were often used as the reference to test

for criterion validity. The development process remains largely the same, requiring expert curation and time-consuming validation. For instance, Literacy Assessment for Diabetes (LAD) [131], Rapid Estimate of Adult Literacy in Vascular Surgery (REAL_VS) [170] and Arthritis-Adapted REALM (A-REALM) [159] were examples in the REALM family. (Oral Health Literacy Instrument (OHLI) [150] and Nutritional Literacy Scale (NLS) [45] followed the design of TOFHLA.

New instruments are constantly developed for particular use scenarios. Examples of specific disease or condition included tests on asthma [175], hypertension [59], diabetes [176], colon cancer [137], and heart failure [7]. Tools for a specific population such as adolescents [42, 181] were also developed. In different health domains, Rapid Estimate of Adult Literacy in Dentistry (REALD)-30 [107], REALD-99 [148], Test of Functional Health Literacy in Dentistry (TOFHLiD) [61], Health Literacy in Dentistry (HeLD) [79], and short-form HeLD-14 [80] targeted dentistry, Rapid Estimate of Adult Literacy in Genetics (REAL-G) [52] measured literacy in genetics.

Another line of research used self reported comprehension assistance seeking-behavior, as opposed to testing an underlying reading ability, to identify patients with inadequate health literacy. One such study presented three questions that can each screen for low literacy [32]. An instrument with a single item was evaluated in a primary care setting to rule out patients with limited health literacy [129].

Among the menagerie of instruments, Medical Term Recognition Test (METER) [146] bears the most similarity to our framework. It included 40 actual medical words and 40 nonwords, and required the participant to mark the actual words. This format is generally known as a Yes/No test in the language testing research community. It was proposed in the 1980s as a simple alternative to the traditional multiple-choice method of testing vocabulary knowledge [123]. Scoring of the METER test suffers from a problem that is common to this type of tests: ambiguity in unmarked items. It is not clear whether the participant was uncertain about the item or genuinely did

not know it. Our work addressed this problem by explicitly giving various degrees of familiarity with an item as answer options. A second drawback of this tool is that it reused many of the REALM words, rendering the test somewhat redundant.

2.3 Methods

2.3.1 Targeted Health Literacy

The existing literacy instruments that were reviewed in the previous section do not focus on an individual patient's needs. For example, REALM uses pre-selected medical terms that a patient may not encounter in his or her disease management activities. Failing to understand those irrelevant terms may not pose significant challenge for the patient. TOFHLA included texts that a patient is likely to encounter. However, they are generic, thus do not reflect what the patient needs to comprehend to make better decisions in his or her particular circumstance. The other instruments do incorporate a focus of a certain specialty (such as dentistry in REALD) or condition (such as REALM, asthma knowledge), but not to a level that is specific for the patient. Many patients have more than one condition that may require careful management. In these cases, measuring knowledge in one condition is not sufficient, and testing knowledge of multiple conditions using multiple instruments is cumbersome.

Our work addresses this problem by designing a generic framework that can be customized to target a specific condition or a combination of conditions. It also has the flexibility to produce tests that are condition agnostic and applicable to a large patient population.

2.3.2 Instrument Framework

We modeled our test framework after the Yes/No vocabulary test. Vocabulary is critical to text comprehension [119]. A meta-analysis showed that vocabulary knowledge most likely played a causal role in comprehension [157]. Another work

showed that self-reported comprehension scores improved after lay definitions were provided for medical jargon [143].

In psycholinguistic research, the Yes/No test for vocabulary knowledge usually comprises words at different frequency levels, and pseudowords to calibrate for random guessing. Pseudowords are strings of letters that follow the phonotactic and morphological rules of a language, but are generally not actual words. The participants are asked to indicate whether they know each of the items.

Although this test format seems simple, creating them is not. Our framework generalized this format by relaxing the need to curate a new set of words and pseudoword items each time a new test is required. Moreover, it can account for uncertainty in the participant's familiarity with a word. Our framework can also be customized to a particular domain of interest such as dentistry or hypertension.

There are two parts to generating a test set under our framework. We start from a vocabulary with their associated occurrence frequencies in a large corpus. The vocabulary is first divided into 10 equally sized tiers based on their frequency. A total of five words are then randomly selected from each tier. Next, two pseudowords are generated from two random words in each tier. The 50 words and 20 pseudowords constitute a complete instantiation of the framework. The options a test taker has for each item are a 4-level Likert scale:

1. I have never seen this word and do not know its meaning.

2. I have seen this word but do not know its meaning.

3. I think I know the word's meaning, but I am not sure.

4. I am sure I know the word's meaning.

2.3.3 Scoring Method

To calculate a score, we measure the agreement between a user and a master. A master reader can perfectly answer all the true words with the most confident value and all pseudowords with lowest value on the Likert scale. We generalized Cohen's κ as a measure of agreement, which calculates the observed and chance disagreement.

$$\kappa = 1 - \frac{q_o}{q_e} \tag{2.1}$$

where q_o is the observed disagreement proportion and q_e is the expected disagreement by chance. In an ordinal scale like ours, the proportions can be weighted to account for varying degrees of disagreement [38].

When all the items are considered equal, as in weighted κ, the ratings from the two raters can be summarized in a $K \times K$ contingency table, where K is the number of categories into which a test item can be assigned. The disagreement proportions can be found from this table by multiplying the different degrees of disagreement v_{ij}, where v_{ij} is the weight indicating the disagreement when one rater assigned i whereas the other assigned j to an item.

We generalized this agreement by allowing the test items to carry different weights, thus accounting for their prevalence in a corpus, and a person's likelihood of knowing them. We calculate the observed disagreement proportion by summing the individual item's disagreement, weighted by an item weight. Let $\mathbf{u} = [u_1, u_2, \ldots, u_N]$ denote the item weights for N test items. Note that the weights are normalized such that $0 \leq u_i \leq 1$ and $\sum_{i=1}^{N} u_i = 1$. Let $\mathbf{k} = [k_1, k_2, \ldots, k_N]$ and $\mathbf{l} = [l_1, l_2, \ldots, l_N]$ denote the category assignments given to the test items by the two raters respectively. Finally, let $v(i, j)$ denote a function that returns the disagreement weight between categories i and j. The observed disagreement can be found in Equation 2.2.

$$q_o = \mathbf{u}^\top v(\mathbf{k}, \mathbf{l}) \tag{2.2}$$

The chance disagreement follows from weighted κ, with the distribution of category assignments for each rater weighted by $\mathbf{u}$.

$$q_e = \sum_{i=1}^{K} \sum_{j \neq i} v(i,j) P_k(i) P_l(j) \tag{2.3}$$

$$P_k(x) = \mathbf{u}^\top \mathbb{1}_{[k=x]} \tag{2.4}$$

$$P_l(x) = \mathbf{u}^\top \mathbb{1}_{[l=x]} \tag{2.5}$$

Our generalized κ can be found by substituting the two disagreement proportions in Equation 2.1 with Equations 2.2 and 2.3. The score still has a value range between 0 and 1.

2.3.4 Assessment of Reliability

2.3.4.1 Test Format

In total, two parallel instantiations of our framework were created using the same corpus, and scores were calculated using the same disagreement weight and scheme. The two tests were shown back to back to participants without demarcations. The test takers were not informed that they were taking two equivalent tests.

2.3.4.2 Test Administration

We administered the two parallel instantiations of our framework to 100 Amazon Mechanical Turk (AMT) users. They were screened to be from the United States and had an approval rate of at least 90%. We then eliminated answers from users that were not native speakers of English.

Several quality control items were randomly embedded in the test. They were simple and unambiguous questions with only one clear and correct answer. They served to identify users that attempted to game our test.

2.3.5 Assessment of Validity

2.3.5.1 Test Format

We designed a four-part online questionnaire to validate our health literacy framework. The test consisted of S-TOFHLA, Short Assessment of Health Literacy-English (SAHL-E) [109], our QuikLitE framework, the short form ComprehENotes test [104], and 5 self-reported document difficulty questions.

To generate a set of test items from our framework, we used the Google Books Ngram Corpus [115] as our starting vocabulary. This corpus is a large multilingual collection of digitized books, which were automatically annotated with syntactic information. The English corpus contains approximately 4.5 million volumes and close to half a trillion words. Since the earliest volumes date from the 1800s, we selected a subset of books from 2000 and onward to ensure the vocabulary frequencies reflect current language usage patterns. Because of digitization errors, there were non-English words and non-letter symbols in the resulting vocabulary list. We filtered this list to only keep those that appear in WordNet [128]. The required pseudowords were generated by Wuggy [91]. Wuggy's algorithm operates by building a chain of subsyllabic elements from a large lexicon, and then iterates through this chain to search for possible pseudowords. Given a template word, Wuggy can generate pseudowords that match the template's subsyllabic structure and transition frequency between them.

The ComprehENotes test is an instrument to assess electronic health record (EHR) notes comprehension. It includes 55 snippets of EHR notes from six common diseases or conditions (heart failure, diabetes, cancer, hypertension, chronic obstructive pulmonary disease, and liver failure), and questions generated using Sentence Verification Technique from these notes. In our online set up, we employed the 14-item short-form test.

Texts in the document difficulty questions were randomly selected from Wikipedia articles in the Medicine category. As the writing quality and style vary among

Wikipedia articles, we limited our article selection to those that were marked as feature articles. These featured articles, according to Wikipedia editors, are "professional, outstanding, and thorough", and are "a definitive source for encyclopedic information". Furthermore, only articles designated with top or high importance were considered in order to eliminate obscure topics. These designations signify "extremely important" or "clearly notable" articles, and there are "strong interests from non-professionals around the world" or "many average readers". Finally, to control for document length, the first few paragraphs of the selected articles were used, and all documents were approximately 300 words long. For each document, the users were asked to rate its difficulty from 1 (easiest to understand) to 10 (most difficulty to understand).

Similar to the parallel form reliability test, quality control items that were designed to resemble real test questions were also randomly inserted to filter out cheating test takers.

2.3.5.2 Test Administration

We recruited AMT users to take three versions of our online test. The tests differed in the instantiation of our framework, and the document difficulty self-assessment. We generated two sets of word items from our framework. Two sets of Wikipedia article excerpts were selected for the document difficulty questions. The three versions of the test included different combinations of the vocabulary test and document difficulty test.

A power analysis projected a sample size of 158 to achieve a power of 0.8 with a medium effect size. Published instruments such as S-TOFHLA and SAHL-E, with which we compared in this study, used data from approximately 200 users for validation. We therefore recruited 200 users for each of our test versions. They were screened in the same fashion as in the reliability assessment.

When scoring our literacy test, we adopted a linear disagreement weight, i.e., $v(i, j) = |i-j|$. Item weights for true words were based on their transformed frequency in the Google Books Ngram Corpus. Specifically, the word frequencies were converted to logarithmic scale, and standardized. These transformed frequencies were then passed through a logistic function to obtain the item weights. This item weight scheme emphasizes words with high frequencies, and applies minimum weight on the rare words. We expect high frequency words to be know by most native speakers, and unfamiliarity indicates lower language ability and literacy. At the other end of the frequency spectrum, rare words may pose a challenge for most people, holding little power to distinguish the test takers' vocabulary knowledge. Pseudowords were each assigned a weight equal to the average weight of the true words.

2.4 Results

2.4.1 Score Distribution

We first present a distribution of health literacy scores as assessed by our framework in Figure 2.1. Mean scores among users in the 3 groups were 0.514 (SD=0.114), 0.498 (SD=0.154), and 0.528 (SD=0.101).

2.4.2 Reliability

Of the 100 users that participated in the parallel form test, 90 responses were legitimate. Demographic information of the users is shown in Table 2.1. The correlation between scores of the two equivalent forms was 0.78 (95% CI 0.69–0.85 $P < 0.001$), suggesting a high level of reliability.

2.4.3 Validity

Demographic information of the AMT users is shown in Table 2.2.

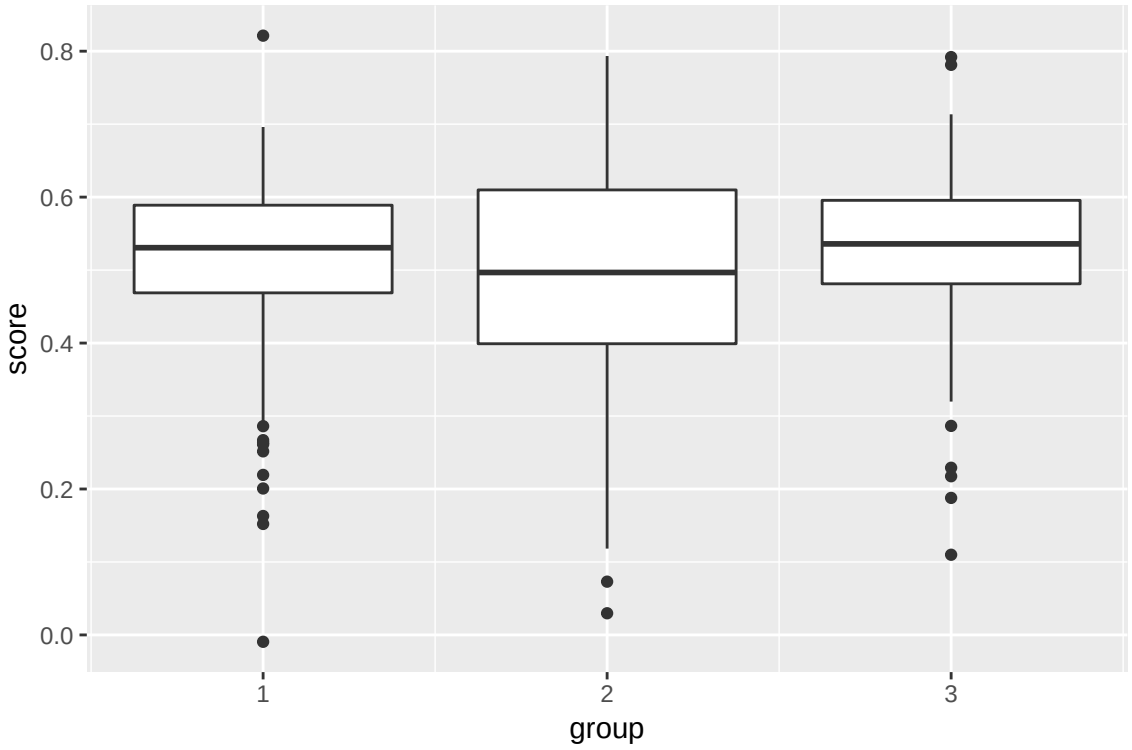

Figure 2.1. Box plot of AMT users' health literacy score according to our framework.

Table 2.1. Demographic information of AMT users in reliability assessment.

Characteristic		Percent (%) (N=91)
Sex	Female	54.95
	Male	45.05
Race	White	81.32
	Black	8.79
	Hispanic	5.49
	Asian	3.30
	IAHPI	1.10
Age	18–24	6.59
	25–34	48.35
	35–44	26.37
	45–54	9.89
	55–64	8.79
Education	High school diploma	29.67
	Associate	29.67
	Bachelor	32.97
	Master or higher	7.69

IAHPI = American Indian/Alaska Native/Native Hawaiian/Other Pacific Islander

Correlation measured between user score and ComprehENotes on the 3 groups of users were moderate to decent, shown in Table 2.3. The correlation coefficients were 0.61 (95% CI 0.51–0.69), 0.49 (95% CI 0.38–0.59), and 0.47 (95% CI 0.35–0.57).

We also measured polyserial correlation between our score and the self-reported document difficulty. The document difficulty scores were reverse coded in the analysis, and treated as an ordinal variable. The correlations of the 3 groups were 0.30 (95% CI 0.17–0.43), 0.21 (95% CI 0.07–0.34), 0.29 (95% CI 0.15–0.41). The weak correlations may be partially explained by the fact that despite given a range of 1 to 10, the AMT users on average rated the document difficulty at 3.8 with a standard deviation of 2.0. Since the Wikipedia document excerpts were taken from well-written articles for a wide readership, and over 70% of the users had at least an associate degree, the actual document difficulty ratings concentrated in a narrow range.

Table 2.2. Demographic information of AMT users in validity assessment.

Characteristic		Group 1 (%) (N=192)	Group 2 (%) (N=196)	Group 3 (%) (N=193)
Sex	Female	46.35	53.06	56.48
	Male	53.65	46.94	43.52
Race	White	70.83	78.57	73.06
	Black	10.94	9.69	10.36
	Hispanic	6.25	3.57	8.29
	Asian	7.81	6.63	7.25
	IAHPI	2.08	0.51	-
	Other	2.08	1.02	1.04
Age	18–24	14.06	11.22	12.44
	25–34	43.23	36.73	38.86
	35–44	23.44	29.59	30.05
	45–54	8.33	13.27	11.40
	55–64	4.69	6.12	4.66
	65+	6.25	3.06	2.59
Education	Less than high school	-	2.04	-
	High school diploma	27.60	31.63	25.91
	Associate	25.00	15.31	18.65
	Bachelor	39.06	37.76	41.45
	Master or higher	8.33	13.27	13.99

IAHPI = American Indian/Alaska Native/Native Hawaiian/Other Pacific Islander

Our framework achieved higher correlation with both ComprehENotes, and self-reported document difficulty than the two existing instruments.

2.4.4 Subpopulation Differences

We compared the score differences between the subpopulations in our validation data. We divided the data based on gender, race, and age to test differences in the subpopulations. Analysis of variance (ANOVA) showed that there was no significant difference between males and females ($F(1, 579) = 2.895 \quad P = 0.089$). Older users tended to score higher ($F(1, 579) = 21.182 \quad P < 0.001$). White users achieved better scores than non-white users ($F(1, 579) = 15.462 \quad P < 0.001$).

Table 2.3. Validity measured by correlation with ComprehENotes and self-reported document difficulty.

	ComprehENotes			Document Difficulty		
	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
QuikLitE	0.61	0.49	0.47	0.30	0.21	0.29
SAHL-E	0.42	0.38	0.43	0.23	0.10	0.11
S-TOFHLA	0.34	0.46	0.40	0.23	0.14	0.11

2.4.5 Ceiling Effect

Existing health literacy instruments may exhibit a ceiling effect, as shown in our data. A total of 52.8% of the users received the full score in SAHL-E, and 55.1% in S-TOFHLA, whereas 32.4% scored perfectly in both tests. Furthermore, an overwhelming majority (94.3%) of the users made at most one error in either one of the tests. This phenomenon was also reported in other studies [116, 31]. In contrast, our framework can accommodate a large variation of user health literacy levels. Among different educational attainment levels (high school or less, college, graduate), ANOVA analysis showed that scores under our framework were significantly different ($F(2, 578) = 5.605 P < 0.01$).

2.4.6 Administration Time

The median time the AMT users finished our test is reported in Table 2.4. The majority (90.36%) of users completed the test in less than 5 minutes. On average, they finished the test 1.5 minutes faster than S-TOFHLA (paired t-test $P < 0.001$). Compared to SAHL-E, users took an additional 1.5 minutes (paired t-test $P < 0.001$). Among the 30 health literacy instruments with a reported administration time from a catalog [71], our test time is shorter than or equal to 23 measures.

Table 2.4. Median administration time in seconds.

	Group 1	Group 2	Group 3
QuikLitE	173.5	180.5	189
SAHL-E	64	63	64
S-TOFHLA	194.5	199.5	192
ComprehENotes	376	432.5	376

2.5 Discussions

2.5.1 Administration

Unlike REALM and its derivatives that rely on word pronunciation checks, our framework can be used in a waiting room without the presence of an administrator, or even at home, where the test taker may experience less anxiety. In a clinic, a test can be administered by a nurse with minimal interference to the clinical work flow since it takes less than 5 minutes. For patients uncomfortable with an electronic device, a paper format can be used, either in a clinic or at home.

Our test can be useful for patients who have seen the material in other instruments. For patients with more exposure to written material, our test can still measure their literacy level. Moreover, if an instrument does not exist for a particular domain of interest, a test can be prepared using our framework.

2.5.2 Flexibilities

Many aspects of our framework can be easily adjusted to a test designer's focus. This has several advantages over existing instruments that are static. First, our framework allows for easy instantiation to suit the test designer's emphasis on a particular subject matter or health care domain. The test may be customized to a particular health care domain, or personalized for a specific patient's need. For example, the education material given to a diabetes patient is different from that for a cancer patient. Separate vocabularies can be compiled from source texts of various subjects, and subject-specific tests can be created to target patients' particular needs.

Furthermore, administering the same test over time to monitor a patient's health literacy can be problematic because repeated testing may result in memorization of the test items, making the measurement unreliable. Creating a unique test on demand with our framework can reduce item repetition, while maintaining similar coverage of word knowledge in a vocabulary.

Additionally, there is no inherent limit to the number of items that can be included in a test under our framework. The only limit is a test taker and the administrator's available time. Therefore, to get a broader coverage test of health literacy, a health practitioner can use more words and pseudowords. The same scoring method can be applied without modification.

Finally, our framework can be adapted to other languages. This is especially helpful in languages such as Spanish that REALM style pronunciation tests are difficult to develop because of the phonemic orthographic rules.

2.5.3 Corpus Size

The sampling process in QuikLitE to create a test depends on the frequencies of words in a corpus. Therefore, it is essential to provide accurate frequency estimation in the instantiation process. Previous studies in psycholinguistics [22] have shown that a corpus of 1 million words can be used to estimate frequencies of high-frequency words (more than 20 occurrences per million). Low-frequency words (less than 10 occurrences per million) require 16–30 million words to estimate reliably.

Corpora of this size in the health domain are not difficult to assemble. In our study to link educational materials for patients (Section 4.2), the Medline Plus corpus contains about 7 million words, which is more than adequate to reliably estimate high-frequency words. To test advanced readers using more rare words, a larger corpus may be obtained from MEDLINE abstracts. A study reported that 5 years of MEDLINE abstracts contain approximately 46 million words [186].

2.5.4 Limitations

As large corpora are readily available, it is straightforward to create a test set with our framework electronically. However, scoring our test manually is challenging. This may limit its utility when a test is administered in a paper format, and a score is needed immediately.

As a test can be generated dynamically, there may be discrepancies with each administration if a new set is created, making comparison difficult. Nevertheless, in our reliability assessment, the median score difference between the two equivalent forms is only 0.06. This difference may have little impact on the overall health literacy assessment of a test taker.

In our data set, the samples were biased toward educated white users. More tests may be needed to assess reliability and validity on underrepresented population in future studies.

Lastly, our framework focuses on print literacy. Numeracy and other skills are also recognized as important for managing one's health. Reading and understanding health-related text is, however, still a critical component to successful engagement with the health care system.

2.6 Summary

Over the decades, a plethora of health literacy instruments were published. Designing such instruments are often time-consuming. When a new need arises, such as a new health context, a specific disease or condition, the laborious development process has to be repeated. We, therefore, proposed QuikLitE, a novel framework that can dynamically generate and score a word recognition-based health literacy instrument. Test results with online Amazon Mechanical Turk users showed high parallel form reliability. Validity as assessed by correlation with ComprehENotes, an EHR comprehension instrument, was higher than two existing health literacy instruments.

Our framework also displayed higher correlation with AMT users' self-reported document difficulty than S-TOFHLA and SAHL-E. Furthermore, QuikLitE is among the easiest to administer and does not exhibit a ceiling effect.

CHAPTER 3

ASSESSING READABILITY OF MEDICAL DOCUMENTS

3.1 Introduction

The research community has relied on readability formulas to assess a variety of information materials for patients. Numerous readability metrics have been developed to assess the grade level or the number of years of education needed for a person to understand the content. One of the most widely used in the health domain is Flesch-Kincaid Grade Level [55] (FKGL), which predicts a grade level using the average sentence length and the average word length.

This metric and many others rely on the assumption that the longer the words and the sentences are, the more difficult the text is. However this assumption may not hold for EHR narratives as sentences are usually short and abbreviations are common.

Accurate measurement of the readability of the Electronic Health Records notes is one important step toward making the notes accessible to the patients. Many studies [21, 73, 64] have evaluated the difficulty of health information intended for patient consumption using readability formulas. They conclude that the materials are often written at a grade level higher than common recommendations [6]. However, the trust in these formulas to measure difficulty may be overextended. Grade-level readability formulas were originally developed to try to ensure that a school textbook for a particular grade was appropriate for children at that grade level [147]. Their capabilities in measuring documents of a highly technical nature such as health care are not thoroughly validated.

In this chapter, we first empirically explore the relationship between these readability formulas and the *perceived* difficulty on general health information and EHR notes in Section 3.2. In Section 3.3, we then propose to build a new model based on consumer perceptions of text difficulty.

3.2 Perceptions of Text Difficulty and Readability Formulas

3.2.1 Overview

As patients express interests in reading their own EHR data [174], health care institutions have also begun to open up access to the EHR records [161]. However, EHRs are written by physicians to communicate with other health care professionals [130]. Merely providing patients with their own EHR records, therefore, does not necessarily help the patients better understand their own conditions. Measuring the readability of the EHR notes is one important step towards making the notes accessible to the patients.

3.2.2 Related Work

Numerous readability metrics have been used for the purposes of preparing texts for school children, language learners, and ensuring smooth written communication. These metrics assess the grade level of text using a linear regression type prediction. Here we briefly introduce three of the metrics. For more discussions on these traditional readability formulas, we refer the reader to the review in [98].

Flesch-Kincaid Grade Level (FKGL) [55] (Equation 3.1) predicts a grade level using the average sentence length and the average word length.

$$0.39 \times \left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8 \times \left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59 \tag{3.1}$$

Simple Measure of Gobbledygook (SMOG) [120] (Equation 3.2) predicts readability based on the number of polysyllabic words (words with more than three syllables),

and the number of sentences.

$$1.0430 \times \sqrt{\frac{\text{polysyllabic words}}{\text{sentences}} \times 30} + 3.1291 \tag{3.2}$$

Similarly, Gunning-Fog Index (GFI) [66] (Equation 3.3) employs sentence length and the proportion of polysyllabic words.

$$0.4 \times \left(\frac{\text{words}}{\text{sentences}} + 100 \times \frac{\text{polysyllabic words}}{\text{words}} \right) \tag{3.3}$$

These metrics, although not developed in the health care domain, are used extensively to measure the readability of various health documents, including patient handouts [21, 177, 178, 180], online health information for patients [73, 30, 51], medication inserts [92, 171], informed consent forms [64, 162, 134], clinical trial information [182], and Wikipedia medical entries [164, 5]. FKGL, in particular, is used in more than half of readability studies compared in one review [172].

In general, these aforementioned metrics rely on the assumption that the longer the words and the sentences are, the more difficult the text is. However this assumption may not hold true for EHR narratives, which contain lists of clinical events (e.g., medication list), abbreviations, and incomplete and short sentences, unduly lowering the readability score.

Less research has been conducted on whether the readability grade levels predicted by these formulas or computational models agree with actual users' perceptions of text difficulty. In other research fields, objective characteristics are shown to not always align with user perceptions. In one study, user perceptions of computer manufacturers' Web sites are different from content analysis tools [108]. In this section, we examine the relationship between users' perceptions of text difficulty and the readability formulas' output.

We evaluate FKGL and other widely used traditional readability metrics. These metrics usually hinge on a few textual characteristics and do not take into account the domain of the text. We also explore the effectiveness of the existing readability formulas on predicting the users' perceptions of difficulty. We show that the perceived readability of technical documents on complex topics is dependent on the domain of the text, not an absolute measure of the difficulty of a piece of text.

3.2.3 Methods

We evaluate existing metrics for assessing EHR readability and investigate their utility in EHR notes. In the following we first describe the data we use for evaluation, followed by an analysis of this corpus.

3.2.3.1 Data

We collected documents about diabetes from two different resources: English Wikipedia (denoted as wiki) and de-identified EHR notes (denoted as med). In wiki documents, we traversed from the Diabetes category. The EHR notes are selected using the International Classification of Diseases, Ninth Revision (ICD-9) code range 250.00 to 250.93. The two sources provide a contrast between texts aimed at the general audience and those written with health care professionals in mind. The statistics of this collection is shown in Table 3.1 under the columns labeled "all".

Diabetes is a common disease that we can expect a large body of readers to be aware of and can provide reasonable judgments on readability. This is especially important in the EHR collection, because randomly selected EHR notes may contain information about rare conditions, which can confuse the readers. The common theme of the content in the two sources also helps address the problem of variations of a user's knowledge in different areas. By constraining to a single condition, we can limit the confounding effect of a user's different levels of familiarity in different areas.

Table 3.1. Document collection statistics. Columns labeled "all" include all documents. Columns labeled "paired" include only documents where another one with a similar length and FKGL score is also available.

Genre	Documents		Sentences		Tokens		FKGL	
	all	paired	all	paired	all	paired	all	paired
wiki	140	58	5703	1084	142,106	23,185	7.33–21.85	7.33–17.82
med	242	133	8715	4232	120,315	57,655	6.48–15.76	6.99–15.76

3.2.3.2 Amazon Mechanical Turk Annotators

To validate one of the most frequently used readability formulas, FKGL, we paired analogous documents in our collection to ask Amazon Mechanical Turk (AMT) users to compare them. Specifically, documents are paired so that they have similar lengths (within 50 token difference) and comparable readability levels according to FKGL (within 0.5 grade level). The statistics on documents that were paired are shown in Table 3.1 under the columns labeled "paired".

We recruited 15 AMT subjects to read and rate pairs of documents. The readers are screened to have English as their native language and be AMT master workers. Three readers had a high school diploma, seven had an associate degree, four had a Bachelor's degree, and one did not report education level. Each reader is presented with 20 randomly selected pairs of documents side by side on the computer screen. The 20 document pairs consist of 5 pairs of wiki documents, 5 pairs of med documents, and 10 pairs of mixed genre documents. The readers are requested to rate the readability of the documents on a scale from 1 (easiest to understand) to 10 (most difficult to understand). Figure 3.1 is a screenshot of the interface with a mixed genre pair.

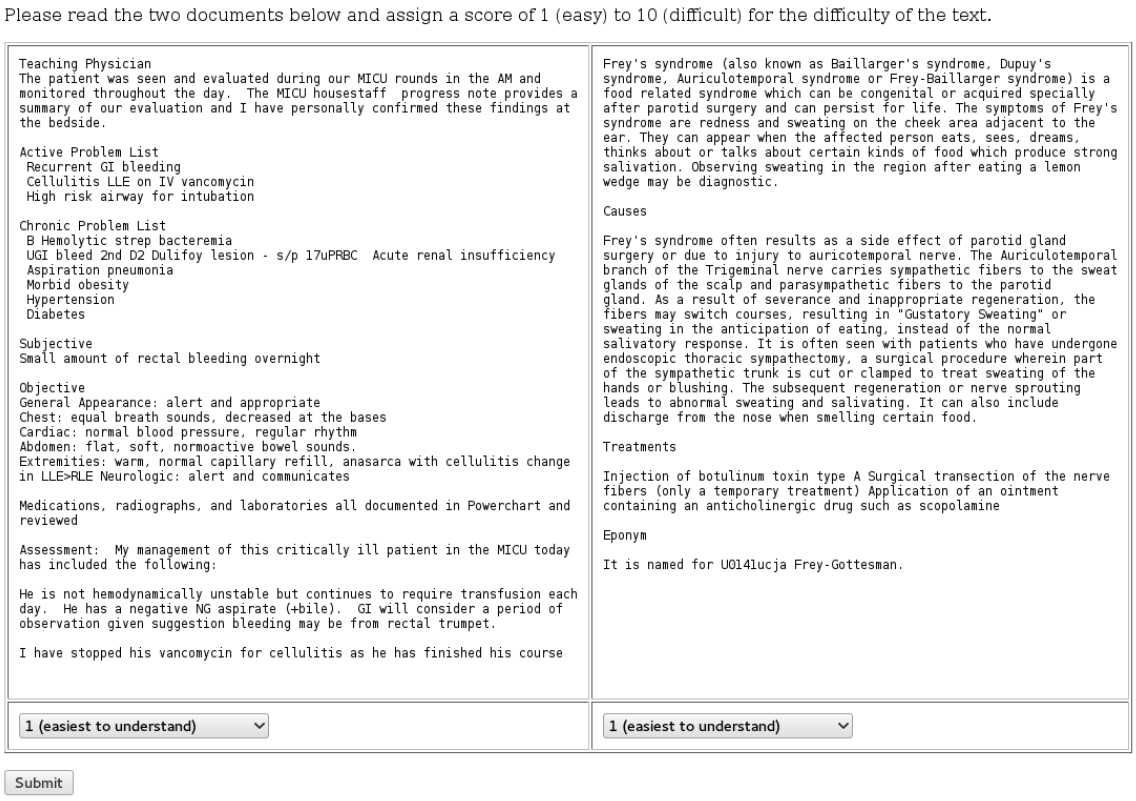

Figure 3.1. Screenshot of the interface for the AMT users.

3.2.3.3 Corpus Analysis

3.2.3.3.1 Readability and User Rating Distributions We first analyze the empirical distribution of AMT users' ratings on the text difficulty, and compare it to the empirical distribution of the readability formulas' scores.

3.2.3.3.2 Correlation between AMT Users We next measure correlations between different AMT users. For each user, all the documents that he or she provided a rating were collected. Since the document pairs were randomly assigned, in general no two users worked on an identical set of documents. Only a subset of the documents were rated by any two users. On average, a document was rated by 2.3 users. Between two users, 8.6 documents were on average rated by both.

We calculated correlations for a user's and any other user's ratings on the documents that were rated by both. The average for each user was obtained by first transforming the correlations by Fisher's z transformation, and then back-transformed [153]. Document genres were not separated in the calculation; otherwise, it would result in too few instances.

3.2.3.3.3 Correlation between AMT User and Readability Formulas To evaluate traditional readability formulas' applicability in technical documents, correlations between each AMT user's ratings and the three readability formulas are measured separately for the wiki and med genres. The average over each user's correlations are also obtained by Fisher's z transformation.

3.2.3.3.4 Differences in Users' Perceived Difficulty To validate the generalizability of FKGL to different genres of text, we tested whether users perceive a difference when the readability scores are similar. The AMT users in our experiments are presented with documents of comparable difficulty (within a difference of 0.5) according to FKGL and of similar length (within 50 tokens difference). We tested the statistical significance of the difference between the difficulty values assigned by the users to two similar documents, separately for wiki, med, and mixed pairs. Two statistical tests are employed—Wilcoxon signed rank test and Kolmogorov-Smirnov test.

We also tested the generalizability of two other formulas using the same procedure. Among all of the document pairs, we selected the subset of documents pairs in which the SMOG scores are within 0.5 between each pair. The same process was repeated using GFI scores.

Furthermore, we explored the disparity in users' perceived difficulty when a readability formula reports a difference between two documents. For each user, we generated pairs of documents from all of the documents he or she rated, then removed the

pairs that were presented during the AMT work session. These document pairs are separated into three types based on the genres of the documents, as in the previous experiments.

3.2.3.3.5 Correlation between Readability Formulas Since FKGL, SMOG, and GFI all involve similar variables (sentence length in words or polysyllabic words, word length), we examined the correlations between different readability formulas on the two genres of text in our data set. Many studies adopt more than one of the traditional formulas to ascertain readability grade level on documents intended for patient consumption [70, 69, 168, 167, 160, 140]. Analyzing the formulas' correlations would inform us of this approach's utility.

3.2.3.3.6 Word Usage We compare the word usage patterns in the two genres of text by examining the common words. First, words in both med and wiki sources are ordered by the frequency they appear in their respective genre. Then, the common words that are in both genres of text in the top frequently used words are counted. The shared vocabulary size may reveal a difference in word usage in different text genres.

3.2.3.3.7 Impact of Medical Concepts Medical jargon is one of the barriers for the patient to understand health information. The eligibility criteria in clinical trials are found to be too difficult for the average American population, mainly due to the frequent use of technical jargon [86]. One study has shown that linking medical terms in EHR notes to Wikipedia pages can improve patient's comprehension [145]. Moreover, many methods have been proposed to identify important or potentially unfamiliar medical terms [50, 189].

We explore the effects of the medical concepts by measuring the correlation between users' ratings and the number of concepts. Medical concepts are identified by running MetaMap [4], a system that identifies biomedical concepts and their semantic

types. In this experiment, we exclude concepts from the following semantic groups and types: Activities & Behaviors, Concepts & Ideas, Geographic Areas, Objects, Occupations, Organizations, Age Group, Animal, Family Group, Group, Human, Patient or Disabled Group, Population Group, Professional or Occupational Group, Educational Activity, Health Care Activity, Research Activity. These semantic groups and types usually do not contain technical medical jargon, and are uncommon in EHR notes. We also excluded Anatomical Structure because in our dataset almost all terms in this category are "body", with the rest being such common body parts as "head" that would not pose difficulty for an average reader.

3.2.4 Results

3.2.4.1 Readability and User Rating Distributions

Empirical distributions of the FKGL readability scores and users' ratings are shown in Figures 3.2 and 3.3. The FKGL histograms (Figure 3.2) on the two genres have clear distinctions. However, contrary to the general belief that EHR notes are more difficult to read, the histogram on the med data peaks to the left of the wiki data histogram. The users' ratings (Figure 3.3), although to a smaller degree, show a higher difficulty level for the med than the wiki data.

Table 3.2 shows the average score of each readability formula and the AMT users' ratings. All the three readability scores suggest the technical EHR notes are significantly easier than lay language wiki articles, whereas the AMT users rated the opposite—wiki articles are 21.31% harder EHR notes.

These results suggest that although FKGL might distinguish the readability of different genres, its counterintuitive predictions could lead to underestimation of difficulty levels on highly complex documents.

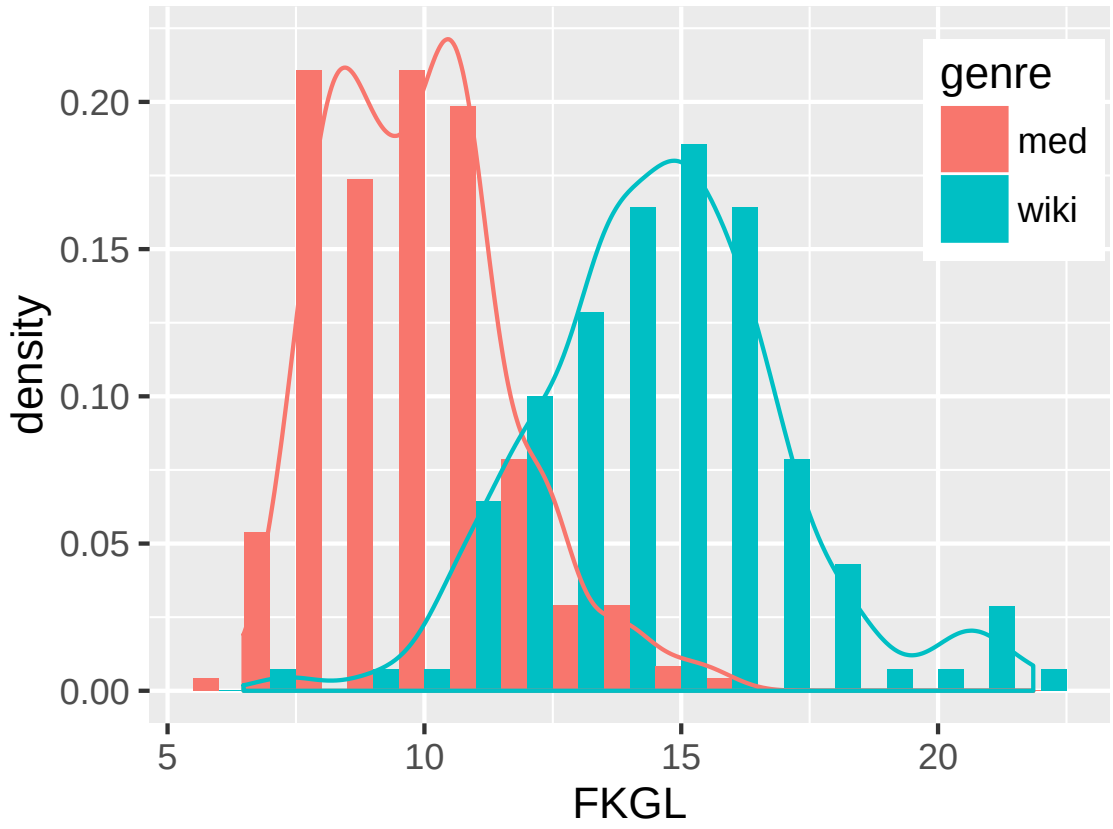

Figure 3.2. Histogram of Flesch-Kincaid Grade Level.

Table 3.2. Average readability score and users' ratings. All differences in scores between the wiki and med genres are statistically significant at level $p = 0.01$ (Mann-Whitney U test). The second to last row shows the percentage med score is higher than wiki.

Genre	Average Score or Rating			
	FKGL	SMOG	GFI	AMT user rating
wiki	14.75	11.07	12.33	4.48
med	9.87	8.74	8.16	5.41
Diff. (%)	−33.09	−21.03	−33.76	20.76
p-value	$< .001$	$< .001$	$< .001$	$< .001$

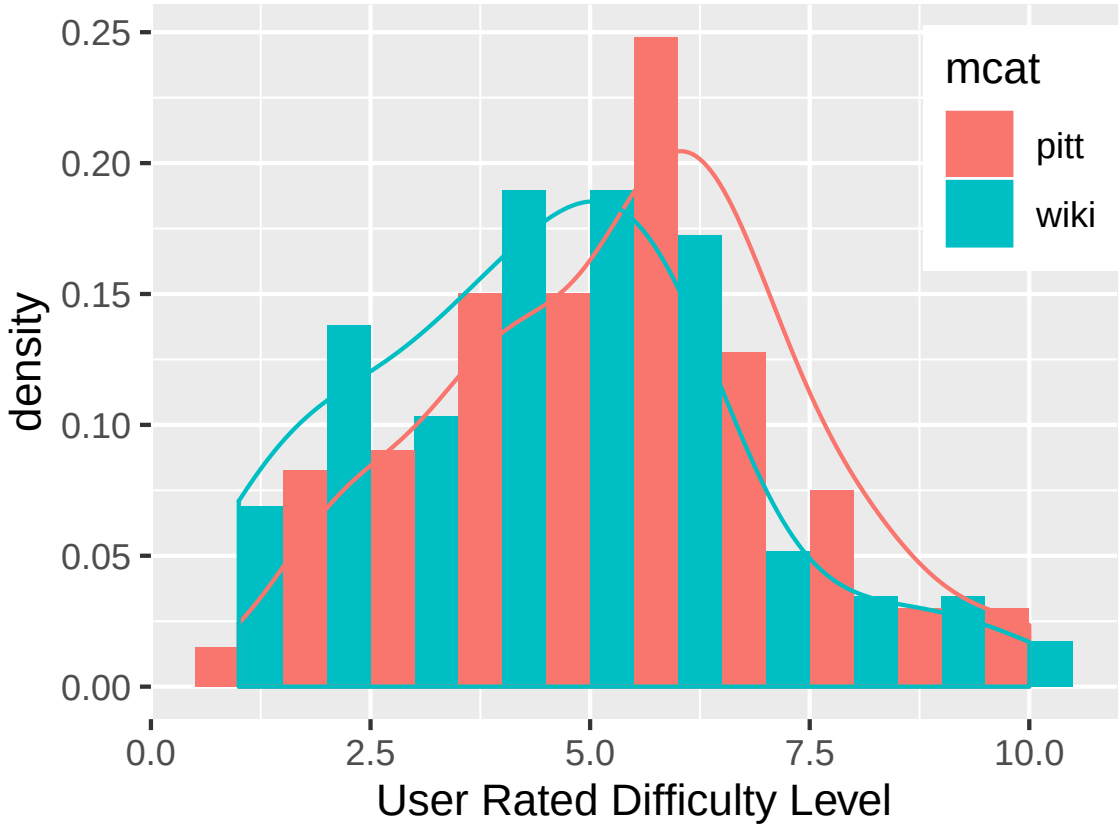

Figure 3.3. Histogram of AMT users' ratings.

Table 3.3. Average correlations between a user and everyone else.

Avg Cor (ρ)	# Users
0.4–0.6	22
> 0.6	78

Table 3.4. Average correlation between users' ratings and readability formulas.

	wiki	med
FKGL	0.171	0.297
SMOG	0.415	0.104
GFI	0.274	0.129

3.2.4.2 Correlation between AMT Users

Table 3.3 summarizes the correlations between two users' ratings. Most users show strong correlation with other users, suggesting that the AMT users' perceptions of difficulty are congruous among themselves.

3.2.4.3 Correlation between AMT User and Readability Formulas

Table 3.4 shows the average correlation coefficients between an AMT user's ratings and the three readability formulas' output. All the correlations are very low, especially in the med category. The SMOG and user rating correlation on wiki data, although slightly higher than FKGL and GFI, is barely moderate. The low correlations suggest that users' perceived difficulty levels are inconsistent with the readability formulas' predictions. For example, one user consistently assigned low difficulty levels to documents with FK scores 12 to 16. However, another user's scores for documents with FK levels approximately 13.5 vary considerably. In contrast, the difficulty perceptions among different users are highly consistent (Table 3.3).

This pattern of inconsistency highlights the inadequacy of these formulas' utility in measuring EHR readability. It also highlights their weakness in testing readability of documents of complex topics such as medicine, as they were developed to help users in the education community to gauge text difficulty below 12th grade. All three formu-

Table 3.5. Statistical significance of difference in AMT users' perceived difficulty between documents of similar Flesch-Kincaid Grade Level.

Genre of Pair	p-value	
	Wilcoxon signed-rank test	Kolmogorov-Smirnov test
wiki	.406	.515
med	.112	.147
mixed	.006	$< .001$

las rely on word counts and sentence counts to estimate text readability. The implicit assumption that longer words are more difficult, however, can often be violated. For instance, abbreviations that are not normally used outside the medical domain such as "CHF" (Congestive Heart Failure) and "EKG" (electrocardiogram) are prevalent in EHR notes, without full definitions. Because these short abbreviations are often comprised of very few, if any, syllables, they would have exactly the same impact on the readability score as do the common stop words such as "the". However, the abbreviations are obviously one of the barriers for a patient to understanding an EHR note. Furthermore, many abbreviations are ambiguous. For example, "MI" can be the shorthand for both "myocardial infarction" and "myocardial ischemia", two different clinical conditions. In fact, disambiguating these abbreviations has been actively studied [184, 96]. Finally, SMOG and GFI's use of polysyllabic words can also exacerbate the problems with abbreviations. For example, "COPD" may be considered a one syllable word in calculating FKGL, but it would make no contribution to the calculation of SMOG or GFI.

3.2.4.4 Differences in Users' Perceived Difficulty

When two documents of similar length and FKGL score are shown together, the ratings assigned by the AMT users exhibit different patterns depending on the genres of the two documents. Using a Wilcoxon signed rank test, the p-values are displayed in Table 3.5 under "Wilcoxon signed-rank test".

Table 3.6. Statistical significance of difference in AMT users' perceived difficulty between documents of similar SMOG or GFI levels.

Genre of Pair	SMOG		GFI	
	Sgn-Rank test	K-S test	Sgn-Rank test	K-S test
wiki	1	1	1	1
med	1	.999	.821	1
mixed	$< .001$	$< .001$	$< .001$	.003

The p-values for a pair of same genre documents show that the users' assignments are not significantly different, consistent with the traditional formula's assessment. However, the p-value for a pair of documents from different genres indicates that despite being assessed at similar difficulty, actual users perceive them as significantly different in terms of readability. Kolmogorov-Smirnov test (Table 3.5) also shows the same trend.

The same tests, when repeated on a subset of document pairs whose SMOG or GFI score difference is within 0.5, confirm that they are not generalizable to different text domains either. Significance test results are displayed in Table 3.6. The AMT users again show significant perceived difference in a document pair of mixed genres, whereas documents in the same genre do not exhibit significant difference.

AMT users' perceptions of difficulty vary depending on the genre of text, even though a readability formula shows no difference. We then explore users' perceived difficulty disparity when a readability formula reports a difference between two documents. Figure 3.4 shows the average difference in users' ratings on a pair of documents with varying differences in FKGL scores.

For a pair of EHR notes, as the difference in FKGL scores widens, AMT users' rating in fact decreases, before increasing when the FKGL score difference is large. A similar trend is present in a mixed pair of documents: AMT users' rating difference decreases initially. For a pair of Wikipedia documents, AMT users' rating difference gradually increases when the FKGL difference is small. However, in all the genres,

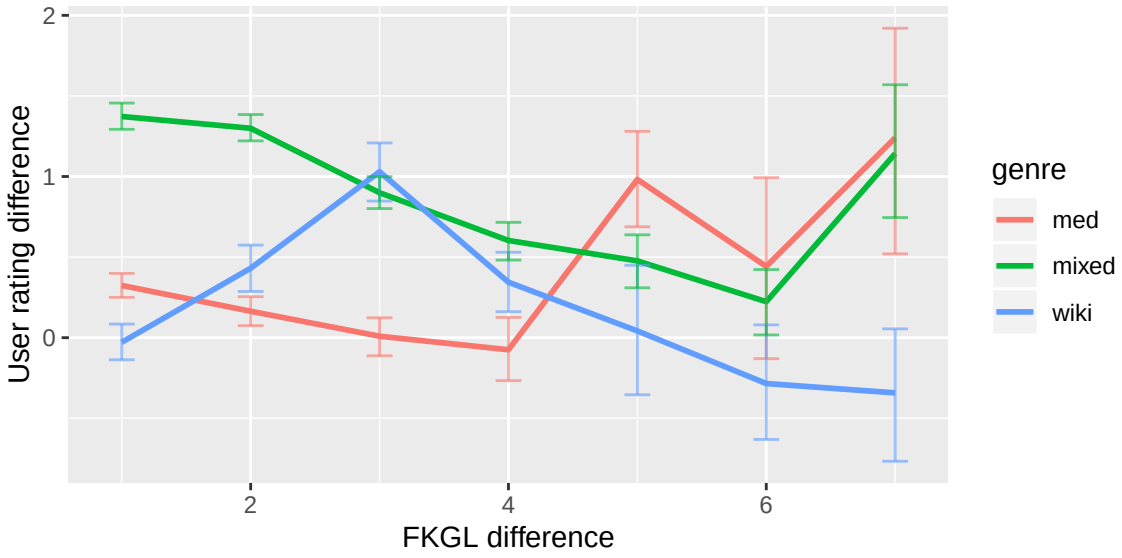

Figure 3.4. Average user rating difference on two documents of different FKGL scores. Error bars are bootstrapped 95% confidence interval.

the users' ratings are limited to no more than 2 levels of difference even for large FKGL differences. These patterns suggest that in a wide range of FKGL scores, users' ratings do not agree with FKGL.

3.2.4.5 Correlation between Readability Formulas

Figures 3.5 and 3.6 show the correlations of SMOG and GFI measured separately against FKGL. Both plots show a positive linear trend between FKGL and the other formulas. The correlation coefficients are shown in Table 3.7.

The correlation coefficients between different formulas confirmed that all three formulas were strongly correlated on our data set regardless of text genre, consistent with the findings from previous studies [158, 166]. The substantial correlation implied that there was limited utility in employing multiple formulas, especially those relying on word and sentence lengths, to reduce potential bias of the individual ones when assessing text readability, as is often done in research studies [70, 69, 168, 167, 6].

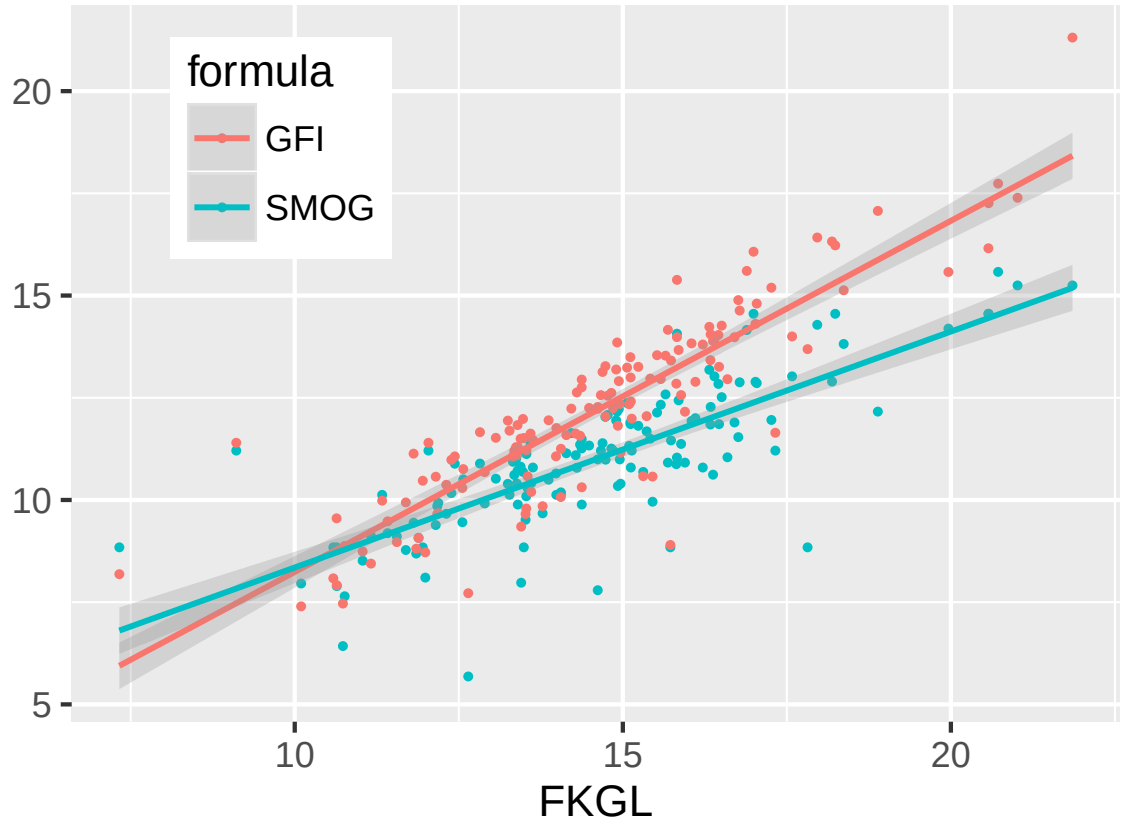

Figure 3.5. Scatter plot of SMOG and GFI scores against FKGL on wiki genre text.

Table 3.7. Correlation coefficients between readability formulas. All correlations are significant ($p < .001$).

	wiki	med
FKGL-SMOG	0.8124	0.8428
FKGL-GFI	0.9191	0.8784
SMOG-GFI	0.8952	0.9696

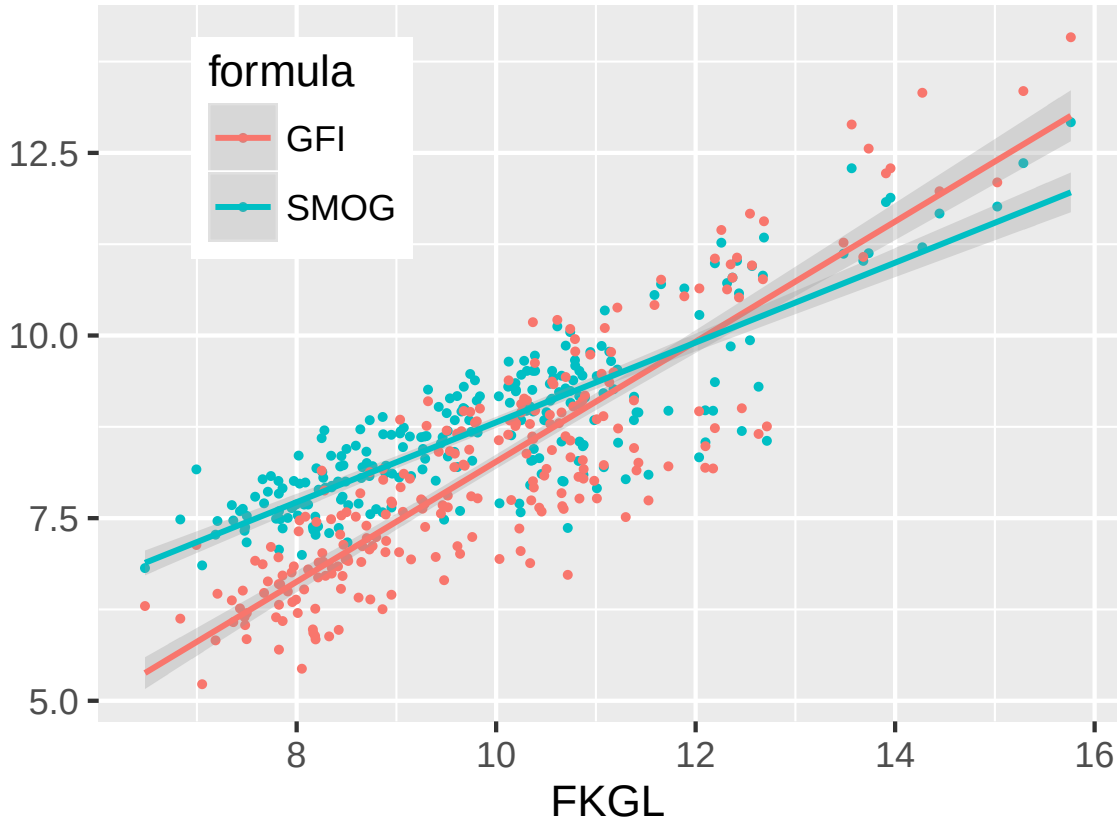

Figure 3.6. Scatter plot of SMOG and GFI scores against FKGL on med genre text.

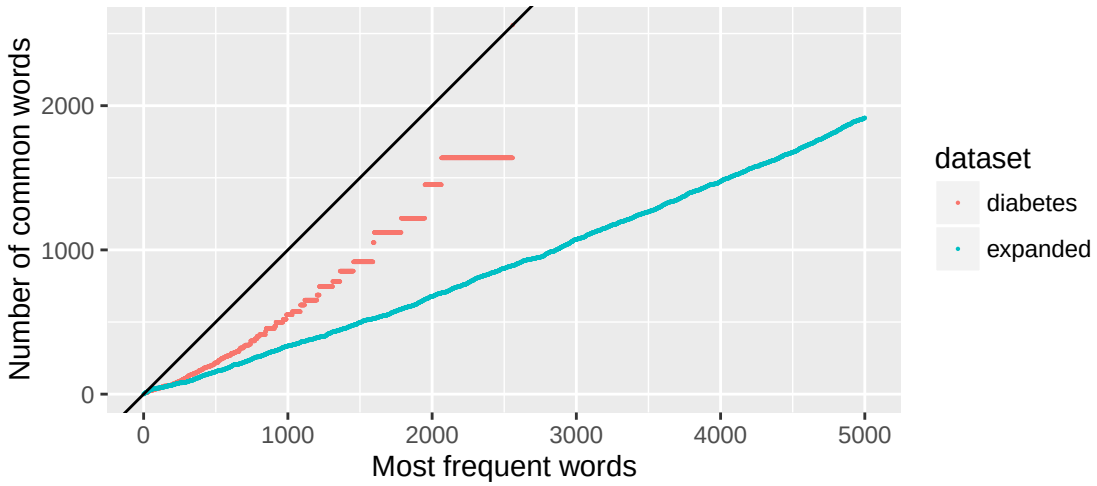

Figure 3.7. Common words in the med and wiki genre texts.

3.2.4.6 Word Usage

In two similar corpora, the N most frequent words from each corpus would be similar. Therefore, the number of common words would increase at approximately the same rate as more frequent words are examined from the two corpora. Significant deviations from this pattern are indications of different word usage patterns. As shown in Figure 3.7, in our set of diabetes documents, the rate of increase in common words between wiki and med documents is significantly smaller (at the level $p < .001$) than one (shown as the solid line in the figure). This suggests that the word usage patterns in the technical (med) and lay language (wiki) documents on the same topic are different.

Expanding to more topics, we built the same word frequency statistic in all Wikipedia articles and about 100,000 EHR notes. Shown in Figure 3.7 as the "expanded" collection, the slope of common word count is also significantly smaller than one (at the level $p < .001$).

Table 3.8. Average correlations between users' ratings and number of medical concepts.

	wiki	med
Number of all concepts	0.4434	0.3987
Number of unique concepts	0.5041	0.4329

3.2.4.7 Impact of Medical Concepts

The correlation coefficients between the number of medical concepts and user difficulty ratings, shown in Table 3.8, are measured for each user and averaged. The average correlation is again obtained by Fisher's z transformation. Correlations with unique concepts are slightly higher than correlations with all concepts in both med and wiki texts. More unique medical concepts are likely to result in more cognitive load for a user to comprehend. It is also worth noting that EHR notes show a lower level of correlation than the Wikipedia documents. This could be in part attributed to the multitude of complexities of EHR notes not limited to the abundance of technical jargon. Writing style such as choice of words and textual cohesion may also account for some of the variances in perceptions of EHR notes' difficulty. In spite of the differences, these correlations suggest that medical jargon is a substantial contributor to readers' perceived difficulty of both genres of text.

3.3 Predicting Readability

3.3.1 Overview

As discussed in Section 3.2.2, most readability metrics in use today in the health care domain are based on formulas developed for the general English text. We have demonstrated that these formulas' prediction did not align with perceived difficulty in health documents from users. In this section, we describe a ranking approach to compare document difficulty.

3.3.2 Related Work

There are efforts in the general domain to build machine learning models to predict text readability. They are usually designed around classification, which are often limited to a few pre-defined labels or require corpora labeled at distinct levels.

One measurement that tailors to the medical domain was proposed in [93]. This method compares surface text, syntactic, and semantic differences to predefined easy and difficult documents, and reports normalized scores instead of grade levels. Another method for health text based on a naive Bayes classifier was developed [113]. The authors collected training documents from online blogs, patient education documents, and medical journal articles. Vocabularies in these documents are used as features for the classifier. Both of the methods rely on manually curated reference documents. Therefore, different choices in constructing the document sets may result in variation in the scores or classification results. Moreover, the classifier is limited, as it assigns only three categories—easy, intermediate, and difficult, and does not assign a grade-level scale.

In this work, we view measuring readability as a ranking task, where the relative difficulty of two documents are compared. This may be a more natural task. In many cases, a grade level is only used to compare to a user's reading ability. A patient-facing EHR system may instead learn from its users' reactions to infer their reading ability and present appropriate education materials. Such a system can be personalized for an individual user. A user with limited literacy will only see straightforward materials and quality materials that require higher literacy levels can be presented to an advanced user.

Table 3.9. Document statistics

		# Documents	# Sentences	# Tokens
wiki	cancer	215	2510	46,349
	diabetes	74	1352	33,402
	hypertension	85	2007	45,440
med	cancer	127	2067	37,830
	diabetes	195	6335	81,085
	hypertension	231	6594	90,784
	total	927	20,865	334,890

3.3.3 Materials and Methods

3.3.3.1 Data

Following a similar procedure in Section 3.2.3.1, we collected difficulty levels on health related documents from AMT users.

Three common diseases were selected as topics from the document sources: cancer, diabetes, and hypertension. Wikipedia documents were randomly selected by recursively traversing up to 3 levels from the respective disease category page, excluding pages about lists, people, and countries. EHR notes were selected using ICD-9 codes (140–195 for cancer, 250.00–250.93 for diabetes, and 401.0–401.9 for hypertension). The document statistics are shown in Table 3.9. For each disease topic, 30 AMT users' data were collected.

3.3.3.2 Learning to rank

We developed a supervised learning system for EHR readability. Traditionally, readability is measured at grade levels. Formulas that see wide adoption in the health care domain include FKGL, Simple Measure of Gobbledygook, Gunning Fog Index, Coleman-Liau Index, and New Dale-Chall. They all use a limited number of factors, mostly word and sentence lengths, to estimate a document's grade level. These simple features, however, are not able to fully capture the complexity of medical documents when used alone as in the formulas. For instance, EHR narratives often contain

abbreviations and lists, which are treated as short words and sentences, thus lowering the estimated grade level. However, the abbreviations present a great challenge to a lay person's understanding.[89, 144]

In the machine learning community, many systems were developed to classify documents into a pre-defined set of readability levels. Such systems can include a multitude of features, including lexical, syntactic, discourse features. These methods are nevertheless constrained in the granularity that they can estimate since the pre-defined difficulty levels are often limited.

In our work, we approached readability as a ranking problem, in which the difficulty levels between documents are compared. This approach overcomes the problems in both the traditional formulas and the classification methods: We are not solely reliant on word and sentence lengths as in the formulas, and our approach can produce an order on readability levels for a set of documents.

A support vector machine (SVM) model was learned from the pairwise comparisons of AMT users' assigned document difficulty levels using the SVM^{rank} package [77]. SVM models normally optimize a hinge loss function based on a binary label for every training example. In the pairwise scenario, the objective is to minimize the number of discordant pairs, i.e. pairs that are ordered incorrectly with respect to the true order. More formally, given a set of training examples $\{(\mathbf{x}_i, y_i)\}$, the primal form of the problem is

$$\begin{aligned} \min \quad & \mathbf{w}^2 + C \sum \xi_{i,j} \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{x}_i \geq \mathbf{w}^T \mathbf{x}_j + 1 - \xi_{i,j}, \forall y_i > y_j \\ & \xi_{i,j} \geq 0, \forall i, j, \end{aligned}$$

where $\mathbf{w}$ is the weight vector, C parameterizes the trade-off between training error and margin size, and ξ are slack variables. Rearranging the first constraint,

$$\mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j) > 1 - \xi_{i,j}.$$

This is equivalent to a classic SVM problem on the modified input vectors $\mathbf{x}' = \mathbf{x}_i - \mathbf{x}_j$. Therefore, a binary classification SVM optimizer can be used to solve the problem.

In our dataset, we generated pairwise difference vectors $\mathbf{x}'$ from each AMT user's ratings. The difference vectors were not generated from different users because ratings across users may not form a consistent ranking as those from a single user do.

3.3.3.3 Features

We employed several types of features, including those from traditional readability formulas. We included average words per sentence, average syllables per word from the FKGL formula, the proportion of polysyllabic words (words with more than three syllables) from the GFI formula, and the percentage of difficult words from the new Dale-Chall formula. Although these formulas do not correlate well with human perceptions of difficulty [190], these word length based features are useful at capturing some longer medical jargon (e.g., Huntington's disease). There is also evidence that perceived difficulty of a word is correlated with its length [112]. We also included word frequency based histogram features. The frequencies were obtained from English Wikipedia and de-identified EHR notes, since common words are found to be likely perceived easier to understand [112]. The frequencies are grouped into 10 bins and the percentage of words in each bin were used as features. Additional features included document length measured in words and sentences. Long documents require more cognitive processing to comprehend, which might translate to higher perceived difficulty. Lastly, we captured language patterns using two word embeddings learned separately from a snapshot of February 2017 English Wikipedia documents and de-identified EHR notes from University of Pittsburgh Medical Center. The mean of all the words in a document under the two embeddings models were included as two sep-

arate sets of features. Word2vec [127] was used to learn a 200-dimensional skip-gram embeddings with a context window size of 5, trained using negative sampling.

3.3.4 Results

3.3.4.1 System performance

We split the annotated data three ways into training (60%), development (20%), and test (20%) sets. The three disease topics were stratified in the split. Hyperparameters were optimized on the development set. Final test results were obtained from a model trained using the optimized parameters.

We evaluated our system using Kendall's coefficient of concordance W [88], a statistic that measures the agreement between rankings from multiple raters. The coefficient aggregates the ranks assigned to each item from all raters, and measures the variance. The variance is then normalized to be between 0 and 1.

Specifically, let $r_{i,j}$ denote the rank given to item i by rater j, where $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. The total rank R_i for item i is

$$R_i = \sum_{j=1}^{m} r_{i,j}.$$

Let $\bar{R}$ denote the mean of the total ranks R_i, and S their variance.

$$\begin{aligned} \bar{R} &= \frac{1}{n} \sum_{i=1}^{n} R_i \\ S &= \frac{1}{n} \sum_{i=1}^{n} (R_i - \bar{R}). \end{aligned}$$

Kendall's W is defined as

$$W = \frac{12S}{m^2(n^2 - 1)}.$$

Higher value represents a high level of concordance. Since the documents in our dataset were randomly assigned to AMT users, common documents among more

Table 3.10. System performance. Bold values indicate a significant increase over the FKGL baseline at 0.05 level using a Wilcoxon signed rank test. Numbers in parentheses are percentage improvements over FKGL.

	cancer	diabetes	hypertension	all
FKGL	0.541	0.490	0.561	0.531
our system				
(all)	0.656 (21.3)	**0.790** (61.3)	**0.715** (27.5)	**0.734** (38.3)
(-eccentric users)	**0.694** (28.3)	**0.762** (55.5)	**0.727** (29.6)	**0.722** (36.0)
(-controversial docs)	0.650 (20.1)	**0.790** (61.3)	**0.759** (35.2)	**0.737** (39.0)

than two users were rare. Therefore, we report the average of the W coefficients between all pairs of users.

The system performance was shown in Table 3.10. As a baseline, we evaluated the performance of the widely used FKGL readability formula. The average agreement between this formula and the AMT annotators was 0.531. Our system achieved an agreement of 0.734 with the AMT annotators, outperforming the FKGL baseline by 38.3%. The increase is statistically significant using a Wilcoxon signed rank test at $p = 0.05$ level.

We also trained and tested separate models for each of the disease topics following the same process. Our system showed consistent improvement over the baseline across all disease categories. The diabetes and hypertension categories saw significant increase in agreement over the baseline FKGL metric. Although not significant, the cancer category still showed substantial improvement over the baseline. These results suggested that our method is robust across different topics.

3.3.4.2 User behavior

A variety of factors may influence a reader's reading comprehension, which in turn determines his or her judgment on a document's difficulty. We examined the differences in the AMT users' difficulty ratings using the same Kendall's W coeffi-

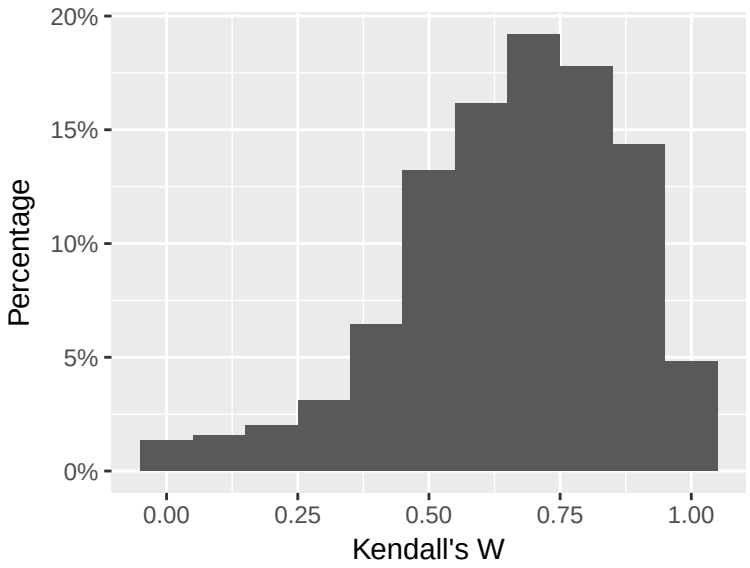

Figure 3.8. Histogram of Kendall's W between two AMT users.

cient. The average concordance between any two users is 0.658. Figure 3.8 shows the distribution of concordance between any two users in our dataset.

While there are pairs of users whose concordance were low, the majority (66%) have a concordance greater than 0.6. When examined on an individual level, the low concordance can often be attributed to a few users that appeared to disagree with many others. There are 9 users that had less than 0.5 concordance with more than 10 other users. Furthermore, five of these users' mean concordance with other users were less than 0.5.

To measure a user's "conformity" in relation to others, we calculated the mean Kendall's W between a user and all his or her peers. The distribution is shown in Figure 3.9.

Approximately one third of the users were highly conforming (mean W at least 0.7) with others, whereas 7% were eccentric (mean W less than 0.4). This result suggests that despite the individual differences in the background knowledge about the subject matter, AMT users still exhibited a consensus on a document's difficulty level. We also noted that our system was able to predict readability orders similar to

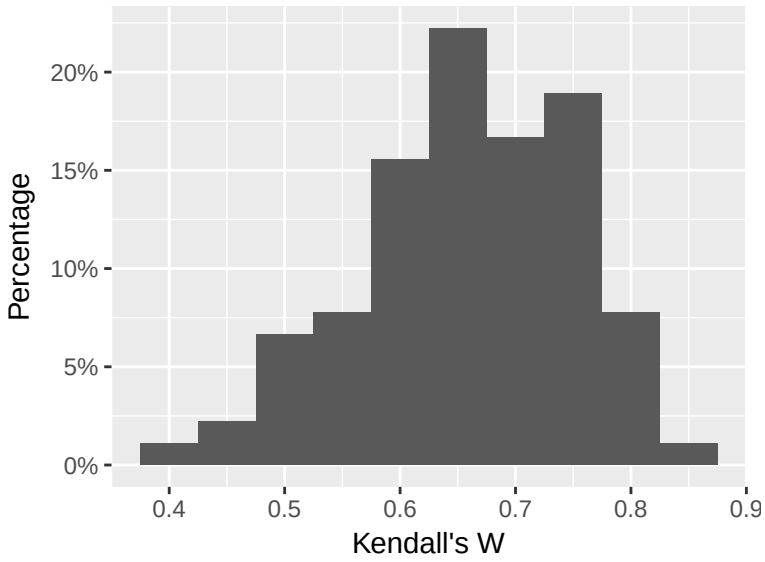

Figure 3.9. Histogram of AMT user's "conformity" (measured by the mean of Kendall's W against peers).

that of a "regular" user. Its mean W was highly correlated with a user's conformity ($\rho = 0.85$). In contrast, the FKGL formula's predicted grade levels did not show strong correlation ($\rho = -0.13$) with conformity.

Table 3.10 (row "-eccentric users") shows performance of models trained from data excluding eccentric users. All disease topics performed significantly better than FKGL. Performance on the entire data set, also significantly higher than FKGL, was slightly lower than the system using full data. This could be due to the large amount of samples removed from training even when a small number of users were excluded, because the difference vectors were generated from all possible pairwise comparisons. On the individual disease topic level, however, the cancer and hypertension models outperformed our system learned from the full training data.

3.3.4.3 Controversial documents

In addition to annotator differences, another factor that contributes to inconsistent annotations is the nature of the documents. We postulate that some documents

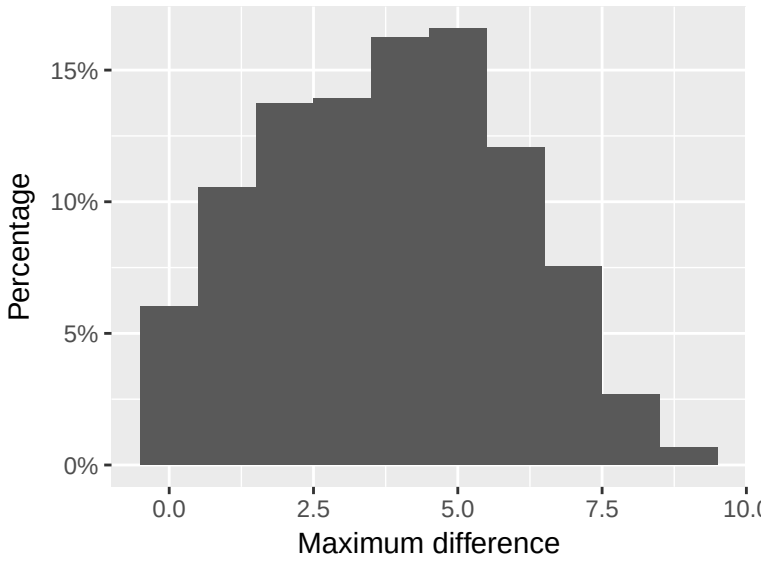

Figure 3.10. Histogram of maximum difference in AMT users' ratings on a document.

may be challenging for the AMT users. For example, certain types of domain-specific writing may appear easy to understand to some but not all users, leading to inconsistent user ratings. These "controversial documents" would also confuse our system, which attempts to learn from the conflicting human annotation. To highlight the range of difficulty perceptions by AMT users, Figure 3.10 shows the maximum difference in ratings assigned by AMT users on documents that were rated by at least two users.

The mean difference is 3.8, suggesting that there were considerable variations in the perceptions of difficulty among users. The two genres of documents (Wikipedia and EHR notes) contained approximately the same number of controversial documents (maximum difference greater than 5), and the cancer topic had more such documents than the other two topics.

We further trained new models after removing controversial documents from the dataset. The performance of these models are shown in Table 3.10 in the last row ("-controversial docs"). Performance of two categories, cancer and diabetes, remained

Table 3.11. Feature ablation. "full" is the system with all proposed features, "-X" indicates a system that excluded feature X.

feature set	cancer	diabetes	hypertension	all
full (from Tab.3.10)	0.656	0.790	0.715	0.734
-frequency	0.652	0.792	0.710	0.733
-formula	0.648	0.789	0.709	0.728
-length	0.636	0.785	0.702	0.716
-embedding	0.677	0.784	0.703	0.714

similar to the models trained from the full data. The hypertension set received an appreciable increase.

3.3.4.4 Feature ablation

We compared the contribution of the different types of features included in our system. Separated models were trained without the readability formula features, word frequency based features, length-based features and word embedding based features. Performance of these models are shown in Table 3.11.

Excluding word embeddings showed the most decrease in performance. The only exception is in the category in which removing this feature resulted in a slight increase in performance. The reason could be that cancer is a very broad topic encompassing many different subtypes, unlike diabetes or hypertension, and our embeddings learned from a different source that may be less representative of our test documents. The word frequency based features did not appear to contribute much to the overall performance. Removing these features only resulted in a 0.1% performance decrease. This could be due to the nature of the word frequency corpus (a general English corpus without any particular emphasis on any domain) we used to calculate these features. The surface text characteristics captured by the formulas shows moderate contribution, although not reliable indicators when used standalone. With the exception of one case, the contributions of the features were consistent across differ-

Table 3.12. Performance of a regression approach on readability assessment, compared to the readability formulas and our ranking approach.

	cancer	diabetes	hypertension	all
FKGL	0.541	0.490	0.561	0.531
regression	0.489	0.624	0.560	0.580
ranking (from Tab. 3.10)	0.656	0.790	0.715	0.734

ent disease topics—word embedding and length-based features the highest and word frequency lowest.

3.3.5 Discussion

We formulated readability assessment as a ranking task. An alternative is to apply a regression model to predict the difficulty levels assigned by the AMT users. We used the same training, development, and test data set as in our ranking approach to learn a SVM-based regression model using $\text{SVM}^{\text{light}}$ [76]. The same set of features were adopted in this model as well. We compared the two approaches with Kendall's W.

Table 3.12 shows that the ranking approach we proposed outperforms the regression approach. One reason is the subjectivity of the difficulty levels by human annotators. Two users may assign very different difficulty values if their background and knowledge are different. Therefore, learning to predict an absolute value can be challenging. However, our AMT users were largely consistent in the relative order of their annotations. Thus the ranking model can more reliably learn from the ordering, even though the absolute values may differ by a large amount.

A limitation of a ranking approach is the lack of an easy-to-interpret grade or difficulty value that can be attached to a document. Such a value can be useful in information retrieval systems that index not only the content but also document metadata. With these values, a user can easily request documents at certain levels using a familiar scale.

3.4 Summary

Patient's access to EHR notes has increased dramatically according to national statistics. Studies have shown that such access to their own EHR notes may lead to improved health care outcomes. Measuring the readability of the EHR notes is one important step towards making the highly complex and technical narratives accessible to the patients. Despite being widely used in the health care domain, existing readability formulas are not thoroughly validated for its appositeness in this domain. In this study, we evaluated several such formulas' abilities in predicting *perceptions* of difficulty in health-related text from Wikipedia and EHR notes. We found that the readability formulas' predictions do not align with perceived difficulty in either text genre.

Better readability assessment of EHR notes and other complex documents is imperative to designing patient support systems that provide accessible information. Toward this end, we developed a new machine learning based method to assess EHR readability from relative orders of text difficulty. We trained a learning to rank system to predict relative difficulty levels of given documents, instead of the traditional classification approach in which documents are assigned levels from a limited pre-defined set of values. Our experiments showed that this method significantly outperformed the widely used Flesch-Kincaid Grade Level formula, and the improvement was consistent across different topics. Our system's concordance with a human user's ratings was higher than the concordance between different human annotators.

CHAPTER 4

FACILITATING COMPREHENSION OF ELECTRONIC HEALTH RECORD NOTES

4.1 Introduction

Decisions involving one's health often need to be made outside a face-to-face visit with a health care provider. For example, choosing a health insurance plan and following directions on medication labels are scenarios where a patient needs to make informed decisions using health knowledge. Growing prevalence of chronic diseases [169, 173, 65] places more burden on the patients to make decisions about their health conditions and treatment options. Demands on the consumers for greater involvement in health-care choices, driven by the increase in "consumer-directed" health plans, may also contribute to higher demands on their literacy skills [49].

Health-related content is one of the most searched-for topics on the internet. Decision support systems such as UpToDate that provide physicians with clinical information at point of care are already deployed in many major academic medical centers in the US. Such information is of interest not only to clinical practitioners but also to patients and their families. However, as they are targeted to the physicians, patients without professional training in medicine will have to resort to other consumer-oriented sources.

Furthermore, the physicians, with their solid understanding of the basic health science, are likely to be able to isolate from a complex scenario individual questions they need answers to. On the other hand, the patients, when faced with a similarly complex EHR note, may struggle to locate the key information that are crucial to their comprehension and management of their health conditions.

In this chapter, we study methods to extract and rank key content and retrieve targeted educational materials to facilitate better comprehension of EHR notes for patients.

4.2 Linking Educational Materials

4.2.1 Overview

EHR notes, as we have discussed, are full of medical jargon, abbreviations, and other domain-specific usages and expressions that are ill-suited for the lay people (patients). One study showed that nearly two thirds of the surveyed patients consider physicians' notes difficult to understand, and radiology reports and nurses' notes [89] are also perceived as difficult. Another study recruited healthy volunteers to read and retell medical documents [90]. Common retelling errors included misunderstanding clinical concepts and physician's findings during a patient's visit. A recent patient survey on web-based access to laboratory results concluded that test result comprehension still needs improvement [118]. Findings from an assessment of lay understanding of medical terms suggest that a substantial proportion of the lay public do not understand phrases often used in cancer consultations and that knowledge of basic anatomy cannot be assumed [28]. This section describes our approaches to generate effective queries from long EHR notes and retrieve educational materials to help patient better understand their conditions.

4.2.2 Related Work

There is a wealth of work in improving patient understanding of medical text. The medical jargon, which is prevalent in the EHR notes, is one evident difficulty in patients' understanding [89]. [187] and [154] created mappings between the medical and consumer terminologies. Unsupervised methods are employed to identify difficult terms and definitions are retrieved using commercial search engines [50]. [85]

developed tools to simplify difficult terms. [62] used morphological analysis and text mining to collect paraphrases for medical terms. Providing definitions of medical jargon is also shown to improve EHR notes' readability. For instance, the NoteAid system [143] identifies medical concepts and fetches definitions from Unified Medical Language System (UMLS), Medline Plus, and Wikipedia and evaluation has shown significant improvement in self-reported comprehension.

High quality information obtained through education materials can potentially lead to better outcomes [19]. The Patient Clinical Information System [35] provides patients with online information resources and educational aides, and evaluations by patients have been positive. However, no automated systems have been reported. The Infobutton Manager Project [34, 33] links EHR notes to other information resources (e.g., drug databases, Google, PubMed, AskHERMES [24]). However, Infobuttons were developed mainly to assist physicians, and were not designed for patients. PERSIVAL is another physician-centric system that accepts user provided queries to retrieve personalized results from a patient care library [121]. EHR notes are used to build topics from consumer health texts. Probabilistic topic modeling is also utilized to recommend education materials to patients with diabetes [84]. Education materials are ranked according to frequencies of terms and topics in a given EHR note. The authors show that the top two recommended documents are significantly more relevant than a randomly selected document from the same domain. [47] use structured data from EHR to retrieve health-related information. An early system was designed to provide personalized health information from a knowledge base by filling manually created templates [27].

Research in domain-specific Information Retrieval is closely related to our work as well. In these searches, it is common to use a document as the base for queries. Patent retrieval [56], as an example, has been widely studied. Patent documents are generally long and complex, necessitating methods to generate shorter queries. For example,

words in the summary section of a patent document can be ranked by tf-idf scores and extracted to form a query [185]. Sentences that are similar to pseudo-relevant documents according to a language model are also used to reduce query length [57]. Other similarity measures such as Kullback-Leibler divergence are used to extract terms, which are expanded to generate queries in the patent retrieval domain [117]. However, the patent retrieval domain is recall-driven, while in our scenario, patients are generally not expected to read relevant education documents exhaustively.

Other than retrieving patents, various methods have been proposed to retrieve documents relevant to passages of text or web documents. A model extended from Conditional Random Fields (CRF) is proposed to identify noun phrases and named entities from a user-selected passage as queries [106]. Similarly, noun phrases in a verbose query are also used as candidates for key concepts [14]. Other related work that reduces long queries includes ranking all subsets of the original query [100]. However, the passages and verbose queries in these systems are shorter than typical EHR notes, which makes the graphical model and other learning based models less efficient. Moreover, parsers and Named Entity Recognizers for the medical domain are less effective than the general domain. Pseudo-relevant documents are exploited to identify concepts for query generation [94].

Information Retrieval in the biomedical domain is also related to this work. WRAPIN is a system that analyzes web pages and retrieves related health documents [58]. The system is limited by the design that the health document sources are only indexed by Medical Subject Headings (MeSH) terms and their synonyms, a controlled vocabulary thesaurus for indexing biomedical publications. Our system does not require indexing the document collection with ontology sources, thus eliminating the computationally expensive extraction of the MeSH terms. More IR systems in the biomedical domain are developed to help physicians and researchers. In a review article by [142], it states that one third of searches may have a positive impact on

physicians. A full text index of EHR notes and query-based IR allowed healthcare providers to perform tasks such as medical management of patients, medical research, and improving the traceability of medical care in medical records [16]. A life science IR system LAILAPS utilizes query expansion and suggestion to improve retrieval results [54]. Another study also found query expansion helpful in retrieving biomedical documents from a subset of MEDLINE [149]. Query expansion using a large, in-domain clinical corpus is reported to be useful for patient cohort identification [191]. The CLEF eHealth [87] challenge includes a task to retrieve information to address questions patients may have when reading clinical reports. This task provides participants with expert-formulated concise queries for one central disorder in discharge summaries [60]. In our study, we aim to generate queries from long EHR notes without the help of experts. TREC Clinical Decision Support Track is another information retrieval challenge involving EHR notes. A number of participants extracted terms from the query descriptions exhaustively using external knowledge bases, and expanded them with synonyms defined in medical ontologies. Relevance feedback is also a popular technique among the participating systems. Unlike our method that filters the pseudo-relevant documents, some systems use manual judgments or the top documents. The task is designed to address the physicians' information needs of diagnosing the condition, further testing, and treating the patients, rather than the patients' needs of education materials. Case reports are provided as query descriptions, which can be shorter and more focused than an EHR note.

4.2.3 Generating Queries from EHR Notes

We have explored several methods to automatically generate queries from EHR narratives. In our queries, sequential dependence model [125] was used to capture the dependencies in a multi-word query term. In this model, given a query, documents are ranked based on features of documents containing a single query term, two query

Table 4.1. Statistics of MedlinePlus Collection

Document Type	Documents (Tokens)	Average Tokens (StdDev)
Health Topics	956 (141,185)	147.7 (37.6)
Medical Encyclopedia	7078 (5,126,101)	724.2 (363.7)
Drugs, Supplements, and Herbal Information	1332 (1,726,570)	1296.2 (992.8)
Total	9366 (6,993,856)	749.1 (565.9)

terms sequentially appearing in the query, and two query terms in any order. This model has been shown to be effective in many applications [9, 26, 15].

4.2.3.1 Data

MedlinePlus[1] provides current and reliable information about over 900 health topics pages and 1000 medication pages to users in consumer-oriented lay language. Additionally, the medical encyclopedia section includes over 7000 articles about diseases, tests, symptoms, injuries, and surgeries. We include in this study the textual narratives in the "health topics", "drugs, supplements, and herbal information", and "medical encyclopedia" sections of the MedlinePlus as the collection of educational materials. There are a total of approximately 9400 articles in this collection, which we designate as *MedlinePlus*. Table 4.1 summarizes the characteristics of the collection.

We index the *MedlinePlus* documents with Galago, an advanced open source search engine. Galago implements the inference network retrieval model [165]. This model calculates the probability of the user's information needs being satisfied given a document in a directed acyclic graph. There are four types of binary nodes in this graph: documents nodes corresponding to the event of a document being observed; representation nodes corresponding to document features; query nodes that combine evidence from representation nodes and other query nodes; and an information need

[1] `http://www.nlm.nih.gov/medlineplus/`

Table 4.2. Example EHR Note and its relevant documents

Summary of EHR Note
Patient remains in ICU with the following problems: respiratory failure, hemodynamics, renal failure, status post liver transplant, atrial fib, infectious disease, nutrition.
Select Relevant Documents
Respiratory Failure Deep Vein Thrombosis Aspiration pneumonia Pulmonary Hypertension Kidney Failure Atrial Fibrillation or Flutter Liver Transplantation Dialysis - Hemodialysis

node that combines all of the evidence from the other query nodes. This framework is a formal and robust model to allow efficient combination of not only word-based evidence but also structure, metadata, and other types of evidence. It has been applied in many information retrieval tasks, and shown to be successful [126].

Twenty progress notes are randomly selected from a corpus of de-identified EHR notes as the EHR document collection. Each note contains on average 261 tokens, with a standard deviation of 133. A physician read each note, and manually identified relevant education materials from the *MedlinePlus* documents. For example, a note about various conditions and symptoms of liver disease is linked to an education document on alcoholic disease to discourage the patient from drinking alcohol. Each EHR note is linked to 22 education material documents on average. For example, Table 4.2.3.1 shows the summary of one EHR note and some of its relevant *MedlinePlus* documents. There are approximately 30 sentences or 360 tokens in the actual document.

To evaluate the IR systems, we use the Mean Average Precision (MAP) metric, a common standard in the IR community to evaluate ranked retrieval results. Another metric that we use to evaluate our system performance is precision at 10. It is useful as patients are less likely to read more than a few related documents.

Precision at position k ($P@k$) is defined as the proportion of k retrieved results that are relevant.

$$P(R, D_k) = \frac{|R \cap D_k|}{k}$$

where R is the ground truth set of relevant documents to a query, and D_k is the top k retrieved documents.

MAP is defined on the Average Precision (AP) on retrieval results.

$$AP(R, D_k) = \frac{\sum_{d_i \in D_k \cap R} P(R, D_i)}{|R|}$$

For a set of queries Q, MAP is defined as

$$MAP(Q) = \frac{\sum_{q \in Q} AP\left(R^{(q)}, D_k^{(q)}\right)}{|Q|}$$

where $R^{(q)}$ is the ground truth relevant documents for query q, and $D_k^{(q)}$ is the top k retrieved documents for query q.

4.2.3.2 Baseline

Our baseline approach uses a full EHR note as the query to the *MedlinePlus* document index and retrieves top 500 relevant documents. Although queries are not generally as long as EHR notes, an average patient without adequate medical knowledge may have difficulties constructing effective queries. Thus, this baseline can be considered as a proxy to how a patient actually conducts his or her own search in the real word.

Since EHR text is not patient-oriented, to investigate the gap between medical language and lay language, we substituted the medical concepts from the EHR notes with their consumer-oriented counterparts created by the Consumer Health Vocabulary (CHV) Initiative [187]. The EHR notes were first processed by MetaMap [4] to recognize medical concepts. Those recognized concepts that have a corresponding layman term in CHV were subsequently replaced. In order to limit concepts to domain-specific medical terms, we filtered the MetaMap recognized concepts to the following semantic types, as defined in Unified Medical Language System (UMLS) [20]: acquired abnormality, antibiotic, cell or molecular dysfunction, clinical attribute, diagnostic procedure, disease or syndrome, experimental model of disease, finding, laboratory procedure, laboratory or test result, organ or tissue function, pathologic function, physiologic function, pharmacologic substance, sign or symptom and therapeutic or preventive procedure.

MetaMap is a highly configurable system that maps biomedical text to the UMLS Metathesaurus. It employs several steps to process input text, including lexical and syntactic analysis (such as tokenization, sentence boundary detection, part-of-speech tagging, lexical lookup in a lexicon, shallow parsing), variant generation, candidate identification, mapping construction, and word sense disambiguation. UMLS is a suite of knowledge sources in biomedicine and health, and their associated tools produced by the US National Library of Medicine. Metathesaurus, a core component of UMLS, is a multi-lingual vocabulary database that contains various ontologies of biomedical and health related concepts and the relationships among them. It is organized by concept, and groups alternative names and views of the same concept from different source ontologies together. Each of the concepts is assigned a Concept Unique Identifier (CUI) and at least one semantic type. CHV was developed to map the most frequently occurring concepts observed among MedlinePlus queries to consumer-friendly display names.

4.2.3.3 Topic Models

Full EHR notes typically discuss several aspects of the patient's conditions, including diagnoses, medication, procedures, etc. We trained Latent Dirichlet Allocation (LDA) topic models [18] from over 6000 de-identified EHR notes to infer topics from the test notes. Three models were learned with 20, 50, and 100 topics, of which the one with 100 topics shows the highest performance.

Traditional LDA models extract distributions over individual word tokens for each topic. However, medical concepts often contain more than one token. We employed turbo topics [17] to find phrases from these topics. This method builds significant n-grams based on a language model of arbitrary length expressions from normal LDA posterior distributions. It first assigns each word a topic using the posterior distribution of the topic variable $z_{d,i}$ for the ith word in the dth document. A log likelihood ratio is then calculated for a bigram under two different language models, one incorporating the bigram as a multi-word expression (the expanded model), the other assuming independence (the unexpanded model). Turbo topics use a back-off language model in which only a sparse set of words are dependent on their history. The conditional distribution of a word v following u under this model is

$$P(v|u) = \begin{cases} \pi_{v|u} & \text{if } v \text{ is dependent on } u \\ \gamma_u P(v) & \text{otherwise} \end{cases}$$

where γ_u is a normalization factor to ensure the distribution sums to one. Permutation test determines the significance of this ratio. Higher order n-gram are built in the same fashion recursively. In our experiments, we set the significance level to 0.001. To translate the topics into queries, we first performed inference on the test notes to find the topic mixture, and then took the top 5 phrases from the most likely topics whose combined probability is over 80%.

To concentrate on medical terms, we trained another LDA model solely from the medical concepts contained in the EHR notes. The same de-identified EHR notes used to train LDA models in the approach above were first processed to find medical terms, in the same way as described in Section 4.2.3.2. The notes were then converted to collections of the UMLS Concept Unique Identifiers (CUIs), corresponding to the medical terms recognized by MetaMap, disregarding the textual content. These converted notes were training documents for the new LDA model. Topics were inferred on the test EHR notes, after being processed similarly. The top 5 CUIs from the most likely topics (with combined probability of over 80 %) are mapped back to phrases in UMLS, and are generated as queries.

4.2.3.4 IDF-filtered Concepts

We also more directly focus on the medical concepts by selecting the top concepts based on their inverse document frequency (IDF) from the EHR note corpus we used to learn LDA models. In a large corpus in general, concepts that occur in a small number of documents are more unique to the document being analyzed. In an EHR note, these concepts are presumably more important for the patient. Therefore, we selected 10 concepts that have the lowest IDF from each note to construct a query, using the following definition:

$$IDF(c) = \log \frac{N}{1 + n_c}$$

where c is a concept, N is the total number of notes, and n_c is the number of EHR notes that contain the concept c.

4.2.3.5 Key Concept Identification

We developed learning-based key concept identification to build queries from EHR notes. We employed linear-chain Conditional Random Fields (CRF) model [102] to

identify key concepts, which are most in need of explanation by external education materials. These key concepts can be considered in a broad sense topics, as they also capture various aspects of the EHR note content. We explored lexical, morphological, UMLS semantic type, and word embeddings as features. A 200-dimensional word embeddings model was induced from a combination of Wikipedia articles in the Medicine category and de-identified EHR progress notes, using the skip-gram architecture trained with negative sampling. We adopted the BIO scheme for the single concept type label: KEY_CONCEPT.

To address the issue of sparse training data, we applied domain adaptation strategies. Wikipedia articles were selected as the out-of-domain data. These articles, especially those in the Medicine category, are an appealing resource as they share similarities to our task of key concept identification. According to Wikipedia's manual of style, internal links in a page are curated by editors to other articles that "help readers understand the article more fully", with "relevant information", or "explain words of technical terms, jargon or slang expressions/phrases". Therefore, the links are usually concepts that are important to topic of the page containing them. The article leads, the sections before the table of contents and the first heading, are used in our dataset, as the manual of style encourages providing internal links in these sections. We treat anchor texts in these paragraphs as key concepts.

We compared three different methods of domain adaptation to identify the key concepts—instance weighting, instance pruning, and feature augmentation. In accordance with the common terminology, we refer to the larger Wikipedia data as source domain, and the smaller EHR notes the target domain data.

Instance weighting [75] merges the data from both corpora with different weights during training. The weights are usually inversely proportional to the size of the corpus. A model is then trained using this weighted training dataset. In our experiments, we used leave-one-out cross validation on the target domain data. In each

fold, the training data is a weighted combination of the Wikipedia data and EHR notes. The test data is the left out EHR note.

Instance pruning [75] removes misleading training instances from the source domain by first applying a model learned from the target domain. For example, if an instance is assigned different labels in the source and target domain corpora, it is removed to prevent the algorithm from learning from this confusing data. We first trained a model on the target domain data, and then predicted the labels on the source domain data. Instances in the source domain that were incorrectly labeled were pruned from the source training set. Finally, a new model was trained using this pruned source domain dataset.

Feature augmentation [39] adds additional features to the training instances to identify which corpus they come from. For each original feature in a training example, a new indicator feature is included to indicate the origin domain of the feature, so the learning algorithm can distinguish features important to each domain. A model is then trained on the combined dataset. In our experiments, we applied cross validation on the target domain in a similar fashion to the instance weighting experiments. In each fold, a feature-augmented corpus was built from all the Wikipedia data and EHR notes, and the test data consisted of one EHR note.

As a baseline system, we used leave-one-out cross validation on the EHR notes. The features in the model include lexical, capitalization, prefix, suffix, word shape, and UMLS semantic type. The semantic types are provided by MetaMap, and added as a feature to each token of the MetaMap-recognized terms.

4.2.3.6 Query Expansion

We explored query expansion by incorporating relevance feedback from pseudo-relevant documents. The initial queries are generated using methods described previously. Among the top 20 retrieval results, those with a title that matches one of the

Table 4.3. System Performance of retrieving educational materials for EHR notes.

	System	P@10	MAP	Increase
1	Baseline	0%	0.0091	-
2	CHV	5%	0.0240	2.6
3	LDA	10%	0.0489	5.4
4	LDA on concepts	10%	0.0410	4.3
5	IDF-Filtered Concepts	14.5%	0.0681	7.5

identified key concepts are considered pseudo-relevant documents. This additional requirement is to ensure that the expanded concepts do not drift from the main topic of the medical notes. From these documents, medical concepts are extracted using MetaMap. These concepts, with their synonyms provided by the UMLS Metathesaurus, are used as expansions.

4.2.4 Experiment Results

4.2.4.1 Baseline Approaches

Performance of the baseline approach of using full EHR notes is shown in Table 4.3, row 1. The result using full text with CHV as shown in Table 4.3, row 2 more than doubled. The gap between medical language and lay language highlights the issue that patients may have difficulty finding relevant health information without assistance.

4.2.4.2 Topic Models

100 topics are learned from the de-identified EHR note collection. This level of topic granularity shows the best performance in our experiments. The retrieval result is shown in Table 4.3, row 3. The improvement over the baseline is 5.4 folds, and is statistically significant using a paired Student's t-test ($p < 0.05$). Performance of LDA on concepts is also statistically significant over the baseline system, using the same test, as shown in Table 4.3, row 4.

Table 4.4. Key concept identification results using domain adaptation strategies.

	Precision	Recall	F1
No augmentation	45.77%	26.51%	31.76%
Instance Weighting	47.59%	34.41%	38.32%
Instance Pruning	40.00%	6.02%	10.23%
Feature Augmentation	46.60%	28.86%	34.08%

Table 4.5. System performance of retrieving educational materials for EHR notes, using augmented data.

System	P@10	MAP	Increase over baseline
No augmentation	16.5%	0.0921	10.1
Instance Weighting	18.3%	0.1111	12.2
Instance Pruning	7.8%	0.0316	3.5
Feature Augmentation	14.5%	0.0684	7.5

4.2.4.3 IDF-filtered Concepts

The system performance using IDF-filtered concepts is shown in Table 4.3, row 5. Compared to the baseline and the topic model based methods, this experiment shows that medical concepts are effective query terms.

4.2.4.4 Key Concept Identification

The key concept identification performance of the three CRF models is shown in Table 4.4. Retrieval performance of these models are shown in Table 4.3, rows 4 to 6. All systems showed a statistically significant improvement over the baseline. The last model's improvement is also statistically significant over the LDA approach. Query expansion methods further improved system performance, as shown in Table 4.6.

4.2.5 Discussions

We found that the top 10 retrieved results of the baseline system for each of the EHR notes are nearly identical, with minimal order variations. We also found that none of the top 10 retrievals is a true relevant document according to our gold

Table 4.6. System performance with pseudo-relevance feedback.

System	P@10	MAP
LDA	11.5%	0.0513
LDA on concepts	9.5%	0.0389
IDF-filtered concepts	12%	0.0662
Key concept (no augmentation)	20.5%	0.1114
Key concept (Instance Weighting)	22.5%	0.1424
Key concept (Instance Pruning)	18.5%	0.1002
Key concept (Feature Augmentation)	17%	0.1081

standard. The results are not surprising. EHR notes are written by physicians, containing domain-specific medical jargons. In contrast, consumer-oriented education materials are written in lay language, a different text genre. In addition, the full text of an EHR note may contain noise to the extent that distinguishing content is difficult to locate. For example, an EHR sentence "I am glad to see Ms. Smith today" provides little information other than the gender of the patient, which may still be identified from other parts of the note. Search engines are not optimized to process queries as long as over 500 tokens, and cannot automatically filter out the noise without significant adaptations. The unique language and style in these medical notes makes the filtering all the more difficult.

From the LDA model, Table 4.2.5 shows the top 10 n-grams from 7 topics trained on the medical text. It is clear that while topics like the first one capture medical concepts, others like the second one do not. The LDA results also highlight the noisy nature of the EHR notes. Queries formed by including the generic or noisy terms such as "continue on" will not benefit retrieval results. Examining the retrieval results, we found that when the prominent topics include medical concepts, the top 10 results usually contain at least one relevant document. When only generic topics are identified, relevant documents are absent in the top 10 results.

Table 4.7. Top 10 n-grams from 7 topics using the LDA model

	Phrases with the highest probability
1	dialysis, hemodialysis, catheter, renal failure, renal, coumadin, line, picc line, dialysis catheter, failure
2	job id, today, point, continue on, reasonable, try to, continue, yesterday, left, right
3	continue, patient, job id, pain, patient has, normal, patient s, white count, secondary to, culture
4	liver, ascites, normal, tenderness, fluid, stable, elevated, today, edema, chest
5	preliminary, patient, patient s, time, blood, mmoll, patient has, routine, high, vial
6	diarrhea, abdominal, flagyl, stool, abdominal pain, colitis, abdomen, difficile, fluid, distended
7	bipap, pneumonia, year old, respiratory failure, failure, minutes, requiring, encephalopathy, ards, patient

In the domain adaptation experiments, the precision of the three approaches were relatively close to the baseline of not using augmented data. However, the recall scores vary greatly. In the instance weighting experiment, the model was able to identify many abbreviations that are rare in the target domain. For example, "EGD" and "DVT" were successfully identified as key concepts despite their occurring only once and three times in the target domain corpus. On the other hand, the instance pruning approach removed over half of the training instances from the source domain data, resulting in a lower performance. The Wikipedia Manual of Style states that only the first occurrence of a term should be linked, and generally a link should only appear once. This resulted in many valid instances being removed because of multiple occurrences. For example, repeated mentions of "glucose" in Wikipedia articles were predicted as key concepts by the target domain model. However, most were removed because only one of them in each article was linked to the glucose article. The reduced training size lowered the recall of this model.

In the IR experiments, the instance weighting approach outperformed the baseline of no augmentation in both the single query and query expansion designs. This can be attributed to the higher recall of this approach in the CRF model. Due to its low recall in key concept identification, instance pruning failed to retrieve many relevant documents. For example, in six of the EHR notes, only one phrase was labeled as key concept, and one of them was incorrect. Despite feature augmentation's improvement in the key concept identification experiments over the baseline, queries generated from this approach did not improve over the baseline query result. The identified key concepts by this method included abbreviations such as "CHF" and general symptoms such as "nausea", which can be associated with a multitude of diseases.

4.3 Ranking Important Medical Concepts for Patients

4.3.1 Overview

We have demonstrated that employing key clinical concepts achieved top performance in retrieving relevant education materials for patients. These concepts themselves are a source of confusion for patients without medical training. Many studies have highlighted that patients have difficulty in comprehending medical jargon [144, 89, 28, 110, 81]. To support patient EHR comprehension, we focus on identifying medical terms that matter the most to individual patients in this section.

EHR notes generally incorporate a comprehensive longitudinal description of patients' medical courses. However, patients may care more about their immediate concerns. In patient support applications, providing explanations or educational materials for all the concepts are likely to overwhelm them and may be unnecessary in the first place. Our aim is to develop an automatic system that can identify a small number of important medical concepts specific to a patient. These medical concepts can then be used to provide tailored interventions to improve EHR comprehension

Mr. X is a X-year-old gentleman with history of right-sided heart failure with preserved LVEF, *COPD*, *chronic kidney disease*, coronary artery disease, status post CABG x3, *hypertension*, paroxysmal AFib/flutter who I have been following closely for the last several months. I last saw him on X, at which time, he has significant lower extremity edema, though otherwise did not appear to be in *acute heart failure*. At that time, I did increase his *diuretics* slightly, and recommended that he keep his legs elevated regularly. During that visit, he has stage III kidney disease with a *BUN* of 00, *creatinine* 0; he also remained in *sinus rhythm*. Today in clinic, Mr. X is accompanied by his daughter, X. X, he has no specific complaints related to *heart failure*. He denies *shortness of breath* at rest or with low level activity. He states that he does very little activity throughout the day; he states he is sitting down in the chair 80% of the day, and the other 20%, he is in bed sleeping. He is not participating in *physical therapy* and feels that he has become very weak and deconditioned. He states he has normal balance, and has had no recent *falls*. He denies *orthopnea* or *PND*. He does note his lower extremity *swelling* has improved. He denies any *chest pain* or *pressure*. He does have a *chronic cough*. Denies *lightheadedness*, *dizziness* or *presyncope*. He does not have an *ICD* or *pacemaker* and denies *palpitations*.

Figure 4.1. An example medical record narrative with important medical concepts underlined, and all other concepts italicized.

and disease management. In this work, we designed a neural network based ranking system to automatically order the medical concepts in an EHR note.

Figure 4.3.1 is an excerpt of an EHR note with concepts that are deemed important underlined, and other concepts italicized. In this example, medical concepts appear in almost every sentence. However, only a small number of them are important for patients to understand according to human expert annotations.

4.3.2 Related Work

Research on designing tools to help patients understand health information has focused on substituting difficult terms with easier synonyms or other closely related terms. For example, [188] developed a system to extract medical concepts, and replace them with consumer-friendly terms in CHV, a lexical resource with mappings between medical concepts and consumer vocabulary. If the difficult medical concept was not found in CHV, a term with a broader or narrower sense in a medical ontol-

ogy was searched to find its consumer-friendly counterpart. They reported that on 9 EHR notes, a majority of the terms were translated correctly and helpful. [85] extended this work to use a larger set of relationships to generate explanation phrases of difficult terms. Cloze test score by reviewers on clinical records improved from 35.8% to 43.6%. It also incorporated a module to simplify compound sentences. In [111], the authors created a semi-automated system for writers to choose alternatives of unfamiliar words in medical text. They used word frequencies from the Google Web Corpus as a proxy for term familiarity. Words that occur less than a pre-defined threshold were considered difficult. The system generated synonyms or hypernyms as candidates for these difficult words from WordNet, UMLS Metathesaurus, simple English Wikipedia, and regular English Wikipedia. The candidates were ordered by type and term familiarity. Evaluation by Amazon Mechanical Turk users showed a significant effect of simplification on perceived difficulty and slightly improved understanding with better question-answering for simplified documents. In Swedish medical text, [1] adapted word frequency based difficulty measure by incorporating word substring frequencies, to account for the compounding nature of the Swedish language. In a corpus of medical journal text, all terms having a MeSH synonym that was assessed to be easier were replaced with the easier alternative. Evaluation by two readability measures differed on the difficulty of the replaced text. However, a reader study showed improved readability after replacement.

These work all target difficult terms as a method to simplify medical text. Our study instead focuses on identifying terms that are important for the patients. This approach complements the text simplification methods. Patients reading complex medical records face two challenges: First, the specialized language in the records that deviates from what they normally read and use everyday. Second, the abundance of medical concepts that overwhelm them. The existing works addressed the first challenge by providing simpler alternatives, whereas we tackle the second challenge

by reducing the cognitive load of processing large amount of unfamiliar concepts. We note that all difficult terms are not necessarily important. As shown in Figure 4.3.1, many medical concepts that do not occur frequently in daily usage, such as "ICD" and "presyncope", are not considered important for the patient by physician annotators that read this report. Conversely, important terms may appear to be easy and familiar in everyday English. "Heart failure" and "kidney disease" are examples from the excerpt. However, these are highlighted by physicians because of their significance to the patient's health.

Our approach also differs from the aforementioned work in that our method is patient centered. The importance of a medical concepts depends on the patient's conditions. For example, in the previous excerpt, "COPD" is an important concept. However, in a different report, it is not highlighted as important for the patient. Our work aims to rank the importance of the concepts tailored to the patient's needs.

Our work is also related to systems that extract key phrases from a document. These systems identify topical terms or phrases that are important to the documents, which can be used to index them for later retrieval. Binary classification, including Naive Bayes [179], decision tree [68], and random forest [99], is often adopted. Confidence scores from the classifiers are ordered to arrive at a final rank of the candidate phrases. In the biomedical domain, [114] developed a system that assigns scores to noun phrases based on their degree of relevance to the main theme of the document using MeSH terms. [151] designed features that are specific to the medical domain. KEA++ [124] incorporated information from medical thesauri to extract candidate phrases and select key phrases. In contrast, we formulated important medical concept identification as a ranking problem, and does not require complex processing of the document to extract domain-specific features.

In information retrieval, deep learning models for ad hoc relevance ranking is related to this work. Deep Structured Semantic Model (DSSM) [74] was proposed

to learn vectors for the query and document. It then ranked the relevance of a document with a query by measuring cosine similarity of their vectors. Convolutional Deep Structured Semantic Model (C-DSSM) [152] replaced the feed forward network in DSSM with a convolutional network. [67] proposed Deep Relevance Matching Model (DRMM) to directly model the interaction between the terms in a query and document pair. The interactions were transformed into matching histograms and passed into a feed forward network to produce matching scores for the query terms. A term gating network aggregated the term matching scores to generate an overall matching score. [135] designed DeepRank, a network architecture that attempted to model the human judgment process. It first located query-centric contexts from the document, and built query context interaction matrices. A convolutional network or a 2-dimensional Gated Recurrent Unit was then used to learn representations for these local interactions. A recurrent network utilized the query terms' positions in the document to aggregate the local interaction representations. Finally a term gating network similar to DRMM aggregated the term level scores.

Our problem setting is different in that all of the candidate concepts that need to be ranked for importance come from the same document. Therefore, they all are relevant a priori. Modeling relevance by query document term interactions at the local level may be less effective. Moreover, since the internal document structure, interactions between concepts in particular, could provide information into their importance for patients, we leveraged both semantic and rich ontological relationships to model the document.

4.3.3 Methods

We designed our system to order medical concepts in an EHR note according to their importance to the patient. It first extracts candidate concepts using MetaMap, and obtains their corresponding CUIs and UMLS semantic types. These concepts

are then ranked by a deep neural network with context information from both the sentences they occurred in and the whole document.

The network consisted of three main components to model the candidate medical concept, a local sentential context, and a global document context. Figure 4.2 shows the architecture of our model. In this figure, it scores the concept "stage III kidney disease" in the excerpt from Figure 4.3.1, with sentence 4 as the local context, and all the concepts in the note as the global context.

A feed forward component modeled the medical concepts $\boldsymbol{c}$. It contained a series of dense layers followed by activation layers. Let $x_i^{(l)}$ denote the output of node i from layer l, $\boldsymbol{w}_i^{(l)}$ and $b_i^{(l)}$ the weights and biases at node i of layer l, f an activation function.

$$\begin{aligned} z_i^{(1)} &= \boldsymbol{w}_i^{(1)} c_i + b_i^{(1)} \\ z_i^{(l+1)} &= \boldsymbol{w}_i^{(l+1)} x_i^{(l)} + b_i^{(l+1)} \\ y_i^{(l+1)} &= f\left(z_i^{(l+1)}\right) \end{aligned}$$

The medical concept $\boldsymbol{c}$ to be ranked is fed into the first layer of this component. Rectified linear unit (ReLU) was used as the activation function in each layer:

$$f\left(z_i^{(l)}\right) = \max\left(0, z_i^{(l)}\right)$$

A convolutional component modeled the local sentential context that the medical concept in question occurred in. The sentences are represented by the sequence of medical concepts as extracted by MetaMap. Let $\boldsymbol{c}_{1,n}$ denote a sequence of n concatenated concepts:

$$\boldsymbol{c}_{1,n} = \boldsymbol{c}_1 \oplus \boldsymbol{c}_2 \oplus \ldots \oplus \boldsymbol{c}_n,$$

where $\oplus$ is the concatenation operator. Convolutional filters $\boldsymbol{w}$ are applied to a window of consecutive concepts to generate a feature. Let h_i be the feature that is generated from the concept sequence $\boldsymbol{c}_{i,j}$

$$h_i = f\left(\boldsymbol{w}\boldsymbol{c}_{i,j} + b\right),$$

where f and b are an activation function and a bias term, respectively. We used ReLU as the activation as in the feed forward layers. This operation is applied to all the consecutive concept sequences in a sentence to generate a feature map.

$$\boldsymbol{h} = [h_1, h_2, \ldots, h_s],$$

where $s = n - j + i$ for a convolutional filter of length $j - i$.

A max pooling layer is then applied on the feature map to obtain the maximum value of $\boldsymbol{h}$:

$$\hat{h} = \max(h_1, h_2, \ldots, h_s).$$

Multiple filters of different lengths are used in our model. The output from the max pooling layers of these filters are concatenated to form a representation of the local sentential context.

At the document level, we represented the global context $\boldsymbol{d}$ using a bag-of-concepts model. The concepts are aggregated based on their salience information $\boldsymbol{s}$.

$$\boldsymbol{d} = \boldsymbol{s}^{\intercal} \begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \\ \vdots \\ \boldsymbol{c}_n \end{bmatrix}$$

This document context vector is passed through a feed forward network similar to the medical concept component.

Finally, the three components are concatenated together with the salience information of the medical concept, before being passed to a fully connected dense layer. We captured a concept's salience using centrality measures from two graph representations of the document. The graphs were both constructed from the candidate concepts in an EHR note as the vertices. Two concepts were connected with an edge if there exists a medical relationship between them. These relationships were obtained from Systematized Nomenclature of Medicine Clinical Terminology (SNOMED CT), one of the source vocabularies in UMLS. For example, SNOMED CT asserts a "finding site" relationship between the concepts "heart failure" and "heart structure". From this graph, we derived a degree-based centrality measure for each concept. Another denser graph was built using the same vertices, with an adjacency matrix of concept similarities. An eigenvector-based centrality measure was computed from this graph. This metric was based on LexRank [53], which estimated similarity between two sentences from modified tf-idf vectors of the words. We adopted cosine similarity between two concepts using their embeddings. The two types of centrality measures captured the salience of the medical concept from two different perspectives: One from a expert curated ontology with rich structures, and the other from empirically derived embeddings that represented real world usage.

Input to the network included the embeddings of the medical concept and the embeddings of concepts both in the sentential context and the entire document. To prevent overfitting of the network, we employed dropout layers in each of the three components.

We optimized the neural network's parameters using a pairwise cross logistic loss. Let y denote the ground truth labels, $\hat{y}$ denote the scores computed by the network.

$$L(y, \hat{y}) = -\sum_{j=1}^{n}\sum_{k=1}^{n} \mathbb{1}_{[y_j > y_k]} \log(1 + \exp(\hat{y}_k - \hat{y}_j))$$

where $\mathbb{1}$ is an indicator function.

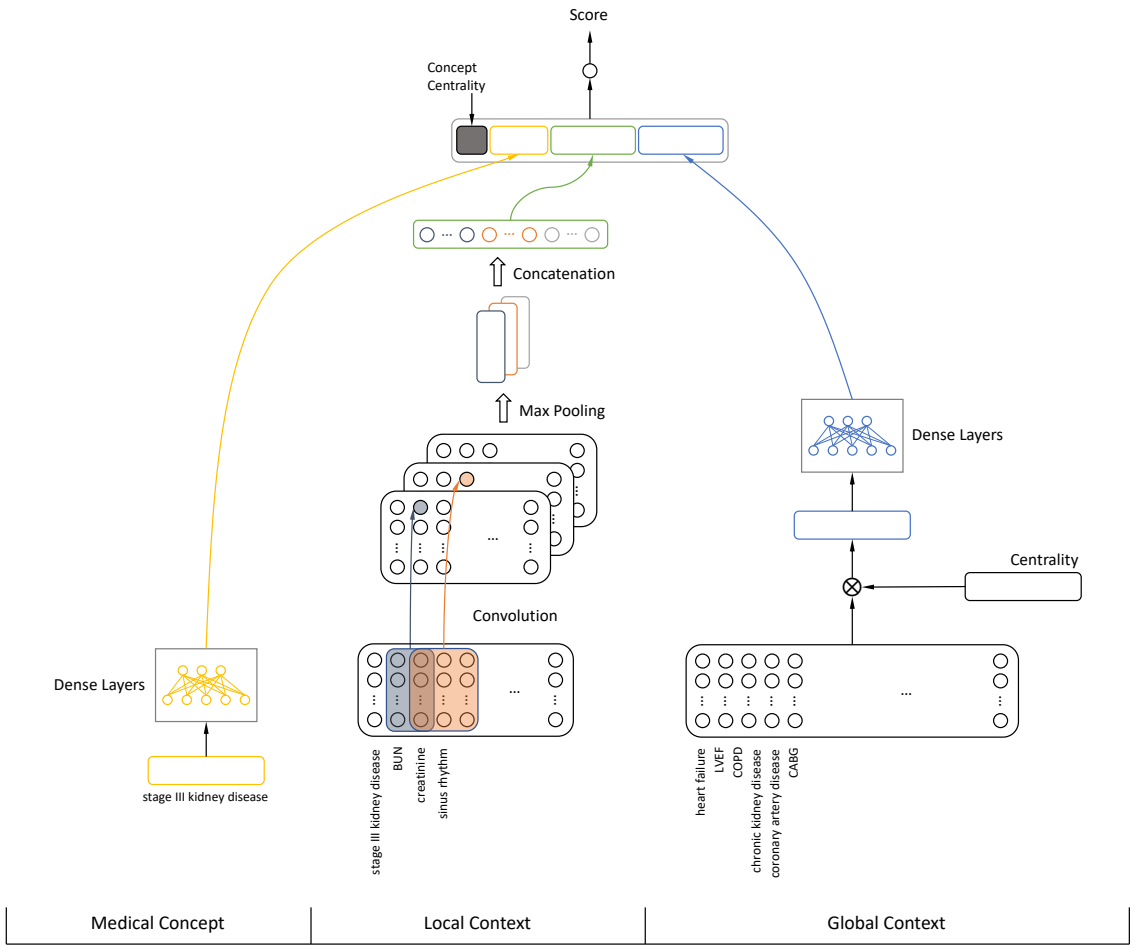

Figure 4.2. Architecture of the neural network model to rank important medical concepts. Three components model the candidate medical concept to rank ("stage III kidney disease" in this example), the local context that the candidate appears in ("stage III kidney disease", "BUN", "creatinine", "sinus rhythm", etc), and the global context of the EHR note ("heart failure", "LVEF", etc), respectively. The learned representations of these components are combined together with the candidate concept's centrality to produce a score.

We evaluate our system performance with Normalized Discounted Cumulative Gain (NDCG) [83]. NDCG can be calculated using Discounted Cumulative Gain (DCG) at position k (DCG_k), which rewards relevant documents more at the top of the retrieval list.

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(1+i)},$$

where rel_i is the graded relevance of the retrieval result at position i. NDCG at position k ($NDCG_k$) can then be computed by

$$NDCG_k = \frac{DCG_k}{IDCG_k},$$

where $IDCG_k$ is the ideal DCG, which is the score when a system correctly orders all the documents based on their relevance.

$$IDCG_k = \sum_{i=1}^{|REL_k|} \frac{rel_i}{\log_2(1+i)},$$

where REL_k is the ideal ordering of the relevant documents up to position k.

We compared our system's performance with several competitive learning to rank algorithms, including RankNet [23], ListNet [25], and LambdaMART [183]. To test the utility of these medical concepts, we applied our system to the education material retrieval task presented in the previous section.

4.3.4 Data

Finding impOrant medical Concepts most Useful to patientS (FOCUS) [29] is a collection of 90 EHR discharge summaries and progress notes from the University of Massachusetts Memorial Hospital outpatient clinics. Six different but common primary clinical diagnoses (cancer, chronic obstructive pulmonary disease, diabetes, heart failure, hypertension, and liver failure) were selected to create the corpus. After de-identification, every note was annotated by two physicians to identify terms

that are important to patients. In this work, an additional 51 notes were annotated following the same guidelines as the original FOCUS corpus. The notes were split into training (60%), development (20%), and test (20%) sets.

We adopted as concept representations a set of pre-trained 500-dimensional clinical concept embeddings (cui2vec) that was learned from a large collection of multimodal medical data [12]. The source data included an insurance claims database of 60 million members, a collection of 20 million clinical notes, and 1.7 million full text biomedical journal articles. This set consisted of embeddings for 108477 medical concepts, represented as UMLS CUIs. It achieved state-of-the-art performance on many concept relation identification benchmark tasks.

4.3.5 Results

4.3.5.1 Corpus

In each note in our corpus, the textual annotations from the physicians were converted to UMLS CUIs by MetaMap. MetaMap was also used to extract all other medical concepts in the original notes. Of the candidate concepts extracted by MetaMap, we only retained those with semantic types that occurred more than 10 times in the physicians' annotations. Since the important concepts were outnumbered by non-important concepts nearly 30 times, non-important concepts were subsampled to produce a balanced training data set.

We consolidated the two relevance rankings from different physicians by assigning to each annotated concept a weight that is equal to the reciprocal of the rank. The final graded relevance of each concept is ordered by the average of the weights from the two annotations. The inter-annotator agreement measured by Cohen's κ was 0.43, showing moderate agreement according to the interpretation scheme by [105].

Table 4.8. Statistics of the important medical concepts corpus.

	N or mean (SD)
Number of EHR notes	141
Average document length (tokens)	1023.9 (163.9)
Average number of unique medical concepts	272.0 (39.0)
Average number of ground truth important concepts	6.1 (4.4)

There are 39444 CUIs in the corpus, 1119 of which are important concepts. On average, each note contains 272 unique concepts, and 6.1 important ones. Table 4.3.5.1 summarizes the statistics of the corpus.

We analyzed the semantic types of the important terms annotated by the physicians. In total, they covered 88 of the 127 semantic types defined by UMLS. The most frequently annotated semantic types and a few examples are shown in Table 4.9.

4.3.5.2 System Performance

In the feed forward components, the number of the dense layers were selected between 3 and 6, while their dimensions were selected from 2^6, 2^7, 2^8, 2^9, and 2^{10}. The convolutional component employed filter widths of 2, 3, 4 with 100 feature maps each. A dropout rate of 0.5 was used in the dropout layer. Training was done using gradient descent with Adaptive Moment Estimation (Adam) [97].

Since only a small fraction of the candidate medical concepts are important, we subsampled the negative training examples during training to obtain a more balanced training dataset. After learning is completed, the full test dataset was used to evaluate our model.

Evaluation results on the test data set are shown in Table 4.10. To compare with the other methods, we used the pre-trained embeddings for the medical concepts and also included the embeddings for all the concepts in the context sentence. Our system outperformed RankNet, ListNet, and RankSVM on all of the metrics. Using paired t-test, improvement on NDCG@10 was statistically significant ($P < 0.05$). It also

Table 4.9. Semantic types of physician annotated important concepts.

Semantic type	Number	Examples
Disease or Syndrome	582	inferior mi, bacteriuria, nephritis, grave's disease, fibromyalgia
Pharmacologic Substance	536	Methotrexate, Insulin, Tylenol, Pantoprazole, PPSV23
Organic Chemical	420	Glipizide, Sucralfate, Atorvastatin, Harvoni, Nitroglycerine
Therapeutic or Preventive Procedure	187	Splenectomy, Bypass, Gastric, Cauterization, Immunizations
Finding	158	Diffusing capacity, Poor Oral Intake, Source, Severe, Low albumin
Diagnostic Procedure	133	Body plethysmography, Bone Marrow Biopsy, Screening Colonoscopy, uterine biopsy, Thoracoscopic lung biopsy
Qualitative Concept	107	Extremely, Poorly Differentiated, Stenosis, Ordered, Related
Amino Acid, Peptide, or Protein	107	BNP, alpha Foetoprotein, Insulin, Antibodies, Anticardiolipin, CT A
Neoplastic Process	103	HCC, Endometrial Adenocarcinoma, Sclerosis, Tuberous, Locally Advanced Cancer, Stage III Hodgkin Lymphoma
Sign or Symptom	101	Syncopes, paralysis left sided, Abdominal pain, epigastric, Leg pain, neoplasm pain

Table 4.10. System performance of identifying important medical concepts.

System	NDCG@1	NDCG@3	NDCG@5	NDCG@10
RankNet	0.09	0.11	0.11	0.15
ListNet	0.09	0.13	0.16	0.19
LambdaMART	0.30	0.30	0.30	0.32
RankSVM	0.13	0.13	0.14	0.14
Our system	0.29	0.27	0.29	0.35

Table 4.11. Performance of incorporating tailored important medical concepts as queries to retrieve educational materials.

System	P@10	MAP
Key concepts	22.5%	0.1424
Additional important concepts from		
RankNet	22.5%	0.1424
ListNet	23%	0.1425
LambdaMART	26.5%	0.1637
Rank SVM	22.5%	0.1438
Our system	26.5%	0.1642

achieved similar performance to LambdaMART, which is a competitive algorithm in learning to rank tasks. An advantage of our approach is that the representations can be reused in other tasks.

When the top 2 most important concepts as recognized by our system was added as additional query terms in the educational materials retrieval system to the Wikipedia-augmented method (described in Section 4.2.3), we see improvement in both the P@10 and MAP measures (Table 4.11).

4.3.6 Discussions

Medical concepts are abundant in EHR notes, which are shown to be a barrier to comprehension by patients. We developed a neural network based system to automatically identify concepts that are important for them to understand the notes.

Existing approaches usually target unfamiliar or difficult terms, whereas our work aims to reduce the medical concept overload problem. The unfamiliar terms, which are often approximated by low frequency words in a large corpus, are not necessarily critical for patients. Unfamiliar terms identified in this way are the same for all patients. However, in our dataset, 40% of the important medical concepts appeared in only one EHR note. Furthermore, although two documents with the same primary diagnoses share about 50 common medical concepts on average, there is less than one common concept between two notes. This suggests that the important terms are in general highly specific to the individual patients, with few shared ones.

Compared to prior work [29], which employed a multitude of sophisticated features, including term frequency, term structure, positional, lexical, part of speech, word embeddings, UMLS semantic type, Consumer Health Vocabulary based, and topical features, our system only required distributed representations of medical concepts and centrality features that can be derived from these representations, reducing the need for complex preprocessing.

The tailored concepts proved helpful when incorporated as query terms to retrieve patient educational materials. For example, in one note, the system identified hypertension as an important concept that was not in the original query when using the CRF based model. This concept is a comorbidity of the main problem experienced by the patient. Including it helped retrieve documents with topics on this condition, which were annotated as relevant for managing the main disease.

4.4 Summary

Patients reading their own medical notes in an EHR system frequently encounter difficult language. In this chapter, we studied approaches to retrieve EHR note-tailored online consumer-oriented health education materials. In our experiments, we have shown that using the full text of an EHR note is ineffective at retrieving relevant

education materials. Identifying key concepts of an EHR note as query terms result in significantly improved performance.

Medical concepts are often cited a major barrier to patient comprehension. We proposed a neural network based system to automatically identify concepts that are important for them to understand the notes. Unlike existing approaches that target unfamiliar or difficult terms, our work aims to reduce the medical concept overload problem by identifying important terms to patients, which are not necessarily rare words. Experiments show that this system outperformed three leading learning to rank algorithms.

CHAPTER 5

CONCLUSIONS

5.1 Overview

This chapter summarizes the dissertation, and is organized as follows. Section 5.2 describes the challenges of engaging patients in the current environment of wide availability of patient access to Electronic Health Records. Section 5.3 reiterates the main contributions of this dissertation. Finally, in Section 5.4 discusses the limitations of this work and future directions for improvements.

5.2 Summary

Patient-centered care has been established as a fundamental approach to improve the quality of health care in a seminal report by the Institute of Medicine published at the start of the century [132]. There is growing awareness that to achieve the best outcomes, patients and families must be more actively engaged in decisions about their healthcare and must have enhanced access to information and support [78]. In this work, we proposed innovative computational tools to facilitate patient engagement, an essential step towards realizing patient-centered care. Patient-centered care shifts the focus from the diagnosis to the patient, a shift that can result in significant improvements in clinical outcomes, patient satisfaction, and cost reduction.

A characteristic in the patient-physician interactions is the information asymmetry—a highly trained professional with domain expertise and a usually non-expert consumer. This characteristic presents significant challenges for patients to effectively communicate and engage with their health care providers and the health care system

Table 5.1. Validity measured by correlation with a reading comprehension test.

Instrument	Correlation
QuikLitE	0.52
SAHL-E	0.41
S-TOFHLA	0.40

Table 5.2. Document readability ranking performance on medical notes.

System	Concordance
FKGL	0.531
Our system	0.734

in general. We proposed methods to determine patient's health literacy level, measure readability of complex documents, identify important information for patients and provide educational materials specific to a patient.

In Chapter 2, we proposed a flexible and computationally inexpensive framework, QuikLitE, to create targeted health literacy instruments. Our assessments showed that the instruments instantiated using this method is both reliable (as demonstrated from high correlation between parallel instantiations in Section 2.4.2) and valid (as shown in Table 5.1). In addition, it does not suffer from the ceiling effect, where differences in literacy at higher levels cannot be distinguished.

Chapter 3 empirically demonstrated that readability formulas that are frequently used in health care research do not align well with lay readers' perceptions of text difficulty (Table 3.4). We therefore proposed a system to compare the readability of complex documents, such as EHR notes and educational materials. Experiments on different genres of medical texts all showed improvement over the traditional FKGL formula, statistically significant in many cases. Overall, the improvement over the baseline FKGL is more than 38% (Table 5.2).

In Chapter 4, we investigated retrieving educational materials that are specific to a patient's needs based on his or her EHR notes. A detailed comparison among

Table 5.3. Educational material retrieval performance for EHR notes.

Method	MAP
Full note	0.0091
Key concept	0.1424

several approaches revealed that identifying key concepts in a note is the most effective method to link educational documents (Table 5.3). Furthermore, we showed, using a physician annotated corpus, that the key medical concepts that are important for patients are not necessarily the unfamiliar or difficult ones, in contrast to many approaches that targeted the unfamiliar concepts. We designed a deep neural network model to rank medical concepts that are important for patients. It utilized embeddings of medical concepts induced from multiple sources and outperformed competitive baselines (Table 4.10). These important concepts also helped improve performance of retrieving educational materials from EHR notes.

5.3 Contributions

The major contributions of this work are as follows.

- A flexible framework that can dynamically generate health literacy instruments for a specific domain.

  We proposed QuikLitE, a framework for health literacy measurement instrument that can be tailored to individual patients. We showed that it is flexible, convenient, reliable, and valid.

- Empirical evidence that current readability measurement tools are inadequate at measuring users' perceived text difficulty.

  Our user studies using EHR notes and general medical text showed that grade levels predicted by the current tools widely used in the health care domain are inconsistent with the users' reported document difficulty.

- Method to measure complex document readability.

  We proposed a method to compare document readability, instead of classifying to pre-defined difficulty levels. Our experiments on various disease topics verified that this method is generalizable and effective at comparing the readability.

- Method to identify medical concepts that are important and tailored for patients.

  To help patients efficiently examine and review the vast amount of medical concepts in their own EHR notes, we proposed a neural network model to order medical concepts, without using sophisticated preprocessing of the text.

- Linking targeted educational materials for patients based on their medical records.

  We proposed methods to identify educational materials that can assist patients' comprehension of their medical records. Our method of generating queries from EHR notes retrieved significantly more relevant documents than what an inexperienced user may be able to using naive methods.

- Improving patient EHR comprehension by incorporating tailored medical concepts.

  The important medical concepts tailored to patients, when incorporated into queries generated from EHR notes, improve retrieval results over using textual queries.

5.4 Future Work

In this section, we summarize the limitations of our proposed methods and future directions for improvements.

5.4.1 Health Literacy Framework

Our QuikLitE framework requires a large corpus from the domain of interest to construct a specific health literacy instrument. Such corpora are generally readily available. However, scoring our test manually is challenging as it involves calculations of various weighted disagreements. This may limit its utility when a test is administered in a paper format for patients who are not comfortable with electronic devices. Future research could explore simpler scoring methods that are amenable to manual calculation, for example, discrete weights based on word frequencies. Future work could also explore other weighting schemes to better represent the individual needs of the user being tested, which could lead to opportunities to identify a test taker's knowledge gaps.

In our data set, the samples were biased toward educated white users in the general population. More tests may be needed to assess reliability and validity on underrepresented population and patient population of a particular health condition in future studies.

5.4.2 Document Readability Assessment

Unlike current methods that assign a label (for example, a grade level or a predefined set of easy, moderate, and difficult levels) as readability to a document, we adopted a ranking approach. This introduces a challenge for the users with a need to classify document difficulty levels. Future studies could investigate approaches that can integrate users' health literacy levels, such as measured using our proposed QuikLitE framework, with readability assessment to determine appropriateness of the documents.

Research in the general domain has explored features at many different levels, including surface, syntactic, semantic, and discourse levels [141, 139]. One of the goals of designing our method was to reduce dependency on such sophisticated processing

of the documents to extract features. Advances in neural network based methods provide a direction for future research to leverage their ability to learn representations directly from text.

5.4.3 Educational Materials Retrieval

Our proposed approaches generated one set of query terms for each document to retrieve educational materials that are relevant to an EHR note. However, as the medical records often contain a comprehensive history of the patient's health, many distinct aspects of the patient's conditions may be documented. Future research could investigate methods that can recognize these different aspects and generate multiple queries to improve the coverage of the retrieved results. In our dataset, there are also educational materials with topics that were not directly mentioned in the EHR notes. Medication is one example. The annotator included medications that the patient may need to take to manage the disease. Future work could investigate approaches that can incorporate external knowledge to address this issue.

In identifying medical concepts that are important to patients, our current model learned representations of concepts and documents separately and only leveraged their interactions as additional features. Future research could explore incorporating interactions between the candidate concept and other concepts in the note. Architecture designs that can influence learning through concept interactions may lead to better representations. Another direction for future work is to investigate approaches that can model the document structure to derive a better representation of the note, as opposed to the bag-of-concepts model we currently adopted.

BIBLIOGRAPHY

[1] Abrahamsson, Emil, Forni, Timothy, Skeppstedt, Maria, and Kvist, Maria. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)* (Gothenburg, Sweden, Apr. 2014), Association for Computational Linguistics, pp. 57–65.

[2] Agarwal, Nitin, Hansberry, David R, Sabourin, Victor, Tomei, Krystal L, and Prestigiacomo, Charles J. A comparative analysis of the quality of patient education materials from medical specialties. *JAMA Internal Medicine 173*, 13 (2013), 1257–1259.

[3] Altin, Sibel Vildan, Finke, Isabelle, Kautz-Freimuth, Sibylle, and Stock, Stephanie. The evolution of health literacy assessment tools: a systematic review. *BMC public health 14*, 1 (2014), 1207.

[4] Aronson, Alan R, and Lang, François-Michel. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc 17*, 3 (June 2010), 229–236.

[5] Azer, Samy A, AlSwaidan, Nourah M, Alshwairikh, Lama A, and AlShammari, Jumana M. Accuracy and readability of cardiovascular entries on wikipedia: are they reliable learning resources for medical students? *BMJ open 5*, 10 (2015), e008187.

[6] Badarudeen, Sameer, and Sabharwal, Sanjeev. Assessing readability of patient education materials: current role in orthopaedics. *Clinical Orthopaedics and Related Research® 468*, 10 (2010), 2572–2580.

[7] Baker, David W, Gazmararian, Julie A, Williams, Mark V, Scott, Tracy, Parker, Ruth M, Green, Diane, Ren, Junling, and Peel, Jennifer. Functional health literacy and the risk of hospital admission among medicare managed care enrollees. *American journal of public health 92*, 8 (2002), 1278–1283.

[8] Baker, David W, Williams, Mark V, Parker, Ruth M, Gazmararian, Julie A, and Nurss, Joanne. Development of a brief test to measure functional health literacy. *Patient education and counseling 38*, 1 (1999), 33–42.

[9] Balasubramanian, Niranjan, Allan, James, and Croft, W. Bruce. A comparison of sentence retrieval techniques. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2007), SIGIR '07, ACM, pp. 813–814.

[10] Barry, Michael J., and Edgman-Levitan, Susan. Shared decision making—the pinnacle of patient-centered care. *New England Journal of Medicine 366*, 9 (2012), 780–781.

[11] Bass III, Pat F, Wilson, John F, Griffith, Charles H, and Barnett, Don R. Residents' ability to identify patients with poor literacy skills. *Academic Medicine 77*, 10 (2002), 1039–1041.

[12] Beam, Andrew L, Kompa, Benjamin, Schmaltz, Allen, Fried, Inbar, Weber, Griffin, Palmer, Nathan, Shi, Xu, Cai, Tianxi, and Kohane, Isaac S. Clinical concept embeddings learned from massive sources of multimodal medical data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 25* (2020), 295–306.

[13] Begum, Nelufa, Donald, Maria, Ozolins, Ieva Z, and Dower, Jo. Hospital admissions, emergency department utilisation and patient activation for self-management among people with diabetes. *Diabetes research and clinical practice 93*, 2 (2011), 260–267.

[14] Bendersky, Michael, and Croft, W. Bruce. Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2008), SIGIR '08, ACM, pp. 491–498.

[15] Bendersky, Michael, Croft, W. Bruce, and Smith, David A. Two-stage query segmentation for information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2009), SIGIR '09, ACM, pp. 810–811.

[16] Biron, P., Metzger, M. H., Pezet, C., Sebban, C., Barthuet, E., and Durand, T. An Information Retrieval system for computerized patient records in the context of a daily hospital practice: the example of the léon bérard cancer center (France). *Appl Clin Inform 5*, 1 (2014), 191–205.

[17] Blei, David M., and Lafferty, John D. Visualizing topics with multi-word expressions. *arXiv:0907.1013 [stat]* (July 2009). arXiv: 0907.1013.

[18] Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent Dirichlet Allocation. *J. Mach. Learn. Res. 3* (Mar. 2003), 993–1022.

[19] Bodenheimer, Thomas, Lorig, Kate, Holman, Halsted, and Grumbach, Kevin. Patient self-management of chronic disease in primary care. *JAMA 288*, 19 (Nov. 2002), 2469–2475.

[20] Bodenreider, Olivier. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res. 32*, Database issue (Jan. 2004), D267–270.

[21] Boles, Catherine D, Liu, Ying, and November-Rider, Debra. Readability levels of dental patient education brochures. *American Dental Hygienists Association 90*, 1 (2016), 28–34.

[22] Brysbaert, Marc, and New, Boris. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods 41*, 4 (nov 2009), 977–990.

[23] Burges, Chris, Shaked, Tal, Renshaw, Erin, Lazier, Ari, Deeds, Matt, Hamilton, Nicole, and Hullender, Greg. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning - ICML 2005* (2005), ACM Press.

[24] Cao, YongGang, Liu, Feifan, Simpson, Pippa, Antieau, Lamont, Bennett, Andrew, Cimino, James J., Ely, John, and Yu, Hong. AskHERMES: An online question answering system for complex clinical questions. *J Biomed Inform 44*, 2 (Apr. 2011), 277–288.

[25] Cao, Zhe, Qin, Tao, Liu, Tie-Yan, Tsai, Ming-Feng, and Li, Hang. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning - ICML 2007* (2007), ACM Press.

[26] Cartright, Marc-Allen, Feild, Henry A., and Allan, James. Evidence finding using a collection of books. In *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing* (New York, NY, USA, 2011), BooksOnline '11, ACM, pp. 11–18.

[27] Cawsey, Alison J., Jones, Ray B., and Pearson, Janne. The evaluation of a personalised health information system for patients with cancer. *User Modeling and User-Adapted Interaction 10*, 1 (Feb. 2000), 47–72.

[28] Chapman, Kristina, Abraham, Charles, Jenkins, Valerie, and Fallowfield, Lesley. Lay understanding of terms used in cancer consultations. *Psycho-Oncology 12*, 6 (2003), 557–566.

[29] Chen, Jinying, Zheng, Jiaping, and Yu, Hong. Finding important terms for patients in their electronic health records: A learning-to-rank approach using expert annotations. *JMIR medical informatics 4* (Nov. 2016), e40.

[30] Cheng, Christina, and Dunn, Matthew. Health literacy and the internet: a study on the readability of australian online health information. *Australian and New Zealand journal of public health 39*, 4 (2015), 309–314.

[31] Chesser, Amy K, Keene Woods, Nikki, Wipperman, Jennifer, Wilson, Rachel, and Dong, Frank. Health literacy assessment of the stofhla: paper versus electronic administration continuation study. *Health Education & Behavior 41*, 1 (2014), 19–24.

[32] Chew, Lisa D, Bradley, Katharine A, and Boyko, Edward J. Brief questions to identify patients with inadequate health literacy. *Family Medicine 11* (2004), 12.

[33] Cimino, J. J., Elhanan, G., and Zeng, Q. Supporting Infobuttons with terminological knowledge. *Proc AMIA Annu Fall Symp* (1997), 528–532.

[34] Cimino, James J. Use, usability, usefulness, and impact of an Infobutton manager. *AMIA Annu Symp Proc* (2006), 151–155.

[35] Cimino, James J., Patel, Vimla L., and Kushniruk, Andre W. The patient clinical information system (PatCIS): technical solutions for and experience with giving patients access to their electronic medical records. *Int J Med Inform 68*, 1-3 (Dec. 2002), 113–127.

[36] Cliff, Barbara. The evolution of patient-centered care. *Journal of Healthcare Management 57*, 2 (2012), 86–88.

[37] Cline, Rebecca JW, and Haynes, Katie M. Consumer health information seeking on the internet: the state of the art. *Health education research 16*, 6 (2001), 671–692.

[38] Cohen, Jacob. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin 70*, 4 (1968), 213.

[39] Daumé III, Hal. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (Prague, Czech Republic, June 2007), Association for Computational Linguistics, pp. 256–263.

[40] Davis, Terry C, Crouch, MA, Long, Sandra W, Jackson, Robert H, Bates, Pat, George, Ronald B, and Bairnsfather, Lee E. Rapid assessment of literacy levels of adult primary care patients. *Family medicine 23*, 6 (1991), 433–435.

[41] Davis, Terry C, Long, Sandra W, Jackson, Robert H, Mayeaux, EJ, George, Ronald B, Murphy, Peggy W, and Crouch, Michael A. Rapid estimate of adult literacy in medicine: a shortened screening instrument. *Family medicine 25*, 6 (1993), 391–395.

[42] Davis, Terry C, Wolf, Michael S, Arnold, Connie L, Byrd, Robert S, Long, Sandra W, Springer, Thomas, Kennen, Estela, and Bocchini, Joseph A. Development and validation of the rapid estimate of adolescent literacy in medicine (realm-teen): a tool to screen adolescents for below-grade reading in health care settings. *Pediatrics 118*, 6 (2006), e1707–e1714.

[43] De Felippe, Nanci, and Kar, Farnaz. Readability of information related to the parenting of a child with a cleft. *Interact J Med Res 4*, 3 (Jul 2015), e14.

[44] Delbanco, Tom, Walker, Jan, Bell, Sigall K., Darer, Jonathan D., Elmore, Joann G., Farag, Nadine, Feldman, Henry J., Mejilla, Roanne, Ngo, Long, Ralston, James D., Ross, Stephen E., Trivedi, Neha, Vodicka, Elisabeth, and Leveille, Suzanne G. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. *Ann. Intern. Med. 157*, 7 (Oct. 2012), 461–470.

[45] Diamond, James J. Development of a reliable and construct valid measure of nutritional literacy in adults. *Nutrition Journal 6*, 1 (2007), 5.

[46] Diviani, Nicola, van den Putte, Bas, Giani, Stefano, and van Weert, Julia CM. Low health literacy and evaluation of online health information: a systematic review of the literature. *Journal of medical Internet research 17*, 5 (2015).

[47] Doupi, P., and van der Lei, J. Towards personalized internet health information: the STEPPS architecture. *Med Inform Internet Med 27*, 3 (Sept. 2002), 139–151.

[48] Duell, Paul, Wright, David, Renzaho, Andre MN, and Bhattacharya, Debi. Optimal health literacy measurement for the clinical setting: A systematic review. *Patient education and counseling 98*, 11 (2015), 1295–1307.

[49] Edlin, M. Consumer-directed health care. the goals: more choice, more control. *HealthPlan 43*, 2 (2002), 12.

[50] Elhadad, Noemie. Comprehending technical texts: predicting and defining unfamiliar terms. *AMIA Annu Symp Proc* (2006), 239–243.

[51] Eltorai, Adam EM, Ghanian, Soha, Adams Jr, Charles A, Born, Christopher T, and Daniels, Alan H. Readability of patient education materials on the american association for surgery of trauma website. *Archives of trauma research 3*, 2 (2014).

[52] Erby, Lori H, Roter, Debra, Larson, Susan, and Cho, Juhee. The rapid estimate of adult literacy in genetics (real-g): a means to assess literacy deficits in the context of genetics. *American journal of medical genetics Part A 146*, 2 (2008), 174–181.

[53] Erkan, Günes, and Radev, Dragomir R. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research 22* (2004), 457–479.

[54] Esch, Maria, Chen, Jinbo, Weise, Stephan, Hassani-Pak, Keywan, Scholz, Uwe, and Lange, Matthias. A query suggestion workflow for life science IR-systems. *J Integr Bioinform 11*, 2 (2014), 237.

[55] Flesch, Rudolph. A new readability yardstick. *Journal of applied psychology 32*, 3 (1948), 221.

[56] Fujii, Atsushi, Iwayama, Makoto, and Kando, Noriko. Introduction to the special issue on patent processing. *Information Processing & Management 43*, 5 (Sept. 2007), 1149–1153.

[57] Ganguly, Debasis, Leveling, Johannes, Magdy, Walid, and Jones, Gareth J.F. Patent query reduction using pseudo relevance feedback. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2011), CIKM '11, ACM, pp. 1953–1956.

[58] Gaudinat, Arnaud, Ruch, Patrick, Joubert, Michel, Uziel, Philippe, Strauss, Anne, Thonnet, Michèle, Baud, Robert, Spahni, Stéphane, Weber, Patrick, Bonal, Juan, Boyer, Celia, Fieschi, Marius, and Geissbuhler, Antoine. Health search engine with e-document analysis for reliable search results. *Int J Med Inform 75*, 1 (Jan. 2006), 73–85.

[59] Gazmararian, Julie A, Williams, Mark V, Peel, Jennifer, and Baker, David W. Health literacy and knowledge of chronic disease. *Patient education and counseling 51*, 3 (2003), 267–275.

[60] Goeuriot, Lorraine, Kelly, Liadh, Li, Wei, Palotti, Joao, Pecina, Pavel, Zuccon, Guido, Hanbury, Allan, Jones, Gareth, and Müller, Henning. ShARe/CLEF eHealth evaluation lab 2014, task 3: User-centred health Information Retrieval. In *CEUR Workshop Proceedings* (2014), vol. 1180, pp. 43–61.

[61] Gong, Debra A, Lee, Jessica Y, Rozier, R Gary, Pahel, Bhavna T, Richman, Julia A, and Vann Jr, William F. Development and testing of the test of functional health literacy in dentistry (tofhlid). *Journal of public health dentistry 67*, 2 (2007), 105–112.

[62] Grabar, N., and Hamon, T. Automatic extraction of layman names for technical medical terms. In *Healthcare Informatics (ICHI), 2014 IEEE International Conference on* (Sept 2014), pp. 310–319.

[63] Greene, Jessica, and Hibbard, Judith H. Why does patient activation matter? an examination of the relationships between patient activation and health-related outcomes. *Journal of general internal medicine 27*, 5 (2012), 520–526.

[64] Grossman, Stuart A, Piantadosi, Steven, and Covahey, Charles. Are informed consent forms that describe clinical oncology research protocols readable by most patients and their families? *Journal of Clinical Oncology 12*, 10 (1994), 2211–2215.

[65] Guariguata, L, Whiting, DR, Hambleton, I, Beagley, J, Linnenkamp, U, and Shaw, JE. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice 103*, 2 (2014), 137–149.

[66] Gunning, Robert. *The Technique of Clear Writing*. McGraw-Hill, New York, 1952.

[67] Guo, Jiafeng, Fan, Yixing, Ai, Qingyao, and Croft, W. Bruce. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16* (2016), ACM Press.

[68] HaCohen-Kerner, Yaakov, Gross, Zuriel, and Masa, Asaf. Automatic extraction and learning of keyphrases from scientific articles. In *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2005, pp. 657–669.

[69] Hansberry, David R, Agarwal, Nitin, and Baker, Stephen R. Health literacy and online educational resources: an opportunity to educate patients. *American Journal of Roentgenology 204*, 1 (2015), 111–116.

[70] Hansberry, DR, Agarwal, N, Gonzales, SF, and Baker, SR. Are we effectively informing patients? a quantitative analysis of on-line patient education resources from the american society of neuroradiology. *American Journal of Neuroradiology 35*, 7 (2014), 1270–1275.

[71] Haun, Jolie N, Valerio, Melissa A, McCormack, Lauren A, Sørensen, Kristine, and Paasche-Orlow, Michael K. Health literacy measurement: an inventory and descriptive summary of 51 instruments. *Journal of Health Communication 19*, sup2 (2014), 302–333.

[72] Henry, JaWanna, Pylypchuk, Yuriy, Searcy, T, and Patel, Vaishali. Electronic capabilities for patients among US non-federal acute care hospitals: 2012–2015. Tech. Rep. ONC Data Brief 38, Office of the National Coordinator for Health Information Technology, Washington, DC, 2016.

[73] Huang, Grace, Fang, Christina H, Agarwal, Nitin, Bhagat, Neelakshi, Eloy, Jean Anderson, and Langer, Paul D. Assessment of online patient education materials from major ophthalmologic associations. *JAMA ophthalmology 133*, 4 (2015), 449–454.

[74] Huang, Po-Sen, He, Xiaodong, Gao, Jianfeng, Deng, Li, Acero, Alex, and Heck, Larry. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13* (2013), ACM Press.

[75] Jiang, Jing, and Zhai, ChengXiang. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (Prague, Czech Republic, June 2007), Association for Computational Linguistics, pp. 264–271.

[76] Joachims, Thorsten. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, Cambridge, MA, 1999, ch. 11, pp. 169–184.

[77] Joachims, Thorsten. Training linear svms in linear time. In *Proc 12th ACM SIGKDD* (2006), ACM, pp. 217–226.

[78] Johnson, Bev, Abraham, Marie, Conway, Jim, Simmons, Laurel, Edgman-Levitan, Susan, Sodomka, Pat, Schlucter, Juliette, and Ford, Dan. Partnering with patients and families to design a patient- and family-centered health care system: Recommendations and promising practices, Apr. 2008.

[79] Jones, K, Parker, E, Mills, H, Brennan, D, and Jamieson, LM. Development and psychometric validation of a health literacy in dentistry scale (held). *Community Dent Health 31*, 1 (2014), 37–43.

[80] Jones, Kelly, Brennan, David, Parker, Eleanor, and Jamieson, Lisa. Development of a short-form health literacy dental scale (he ld-14). *Community dentistry and oral epidemiology 43*, 2 (2015), 143–151.

[81] Jones, R B, McGhee, S M, and McGhee, D. Patient on-line access to medical records in general practice. *Health bulletin 50* (Mar. 1992), 143–150.

[82] Jordan, Joanne E, Osborne, Richard H, and Buchbinder, Rachelle. Critical appraisal of health literacy indices revealed variable underlying constructs, narrow content and psychometric weaknesses. *Journal of clinical epidemiology 64*, 4 (2011), 366–379.

[83] Järvelin, Kalervo, and Kekäläinen, Jaana. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems 20*, 4 (oct 2002), 422–446.

[84] Kandula, Sasikiran, Curtis, Dorothy, Hill, Brent, and Zeng-Treitler, Qing. Use of topic modeling for recommending relevant education material to diabetic patients. *AMIA Annu Symp Proc 2011* (2011), 674–682.

[85] Kandula, Sasikiran, Curtis, Dorothy, and Zeng-Treitler, Qing. A semantic and syntactic text simplification tool for health content. *AMIA Annu Symp Proc 2010* (2010), 366–370.

[86] Kang, Tian, Elhadad, Noémie, and Weng, Chunhua. Initial readability assessment of clinical trial eligibility criteria. In *AMIA Annual Symposium Proceedings* (2015), vol. 2015, American Medical Informatics Association, pp. 687–696.

[87] Kelly, Liadh, Goeuriot, Lorraine, Suominen, Hanna, Schreck, Tobias, Leroy, Gondy, Mowery, DanielleL., Velupillai, Sumithra, Chapman, WendyW., Martinez, David, Zuccon, Guido, and Palotti, João. Overview of the ShARe/CLEF eHealth evaluation lab 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction* (2014), Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, Eds., vol. 8685 of *Lecture Notes in Computer Science*, Springer International Publishing, pp. 172–191.

[88] Kendall, Maurice G, and Smith, B Babington. The problem of m rankings. *The annals of mathematical statistics 10*, 3 (1939), 275–287.

[89] Keselman, Alla, Slaughter, Laura, Smith, Catherine Arnott, Kim, Hyeoneui, Divita, Guy, Browne, Allen, Tsai, Christopher, and Zeng-Treitler, Qing. Towards consumer-friendly PHRs: patients' experience with reviewing their health records. *AMIA Annu Symp Proc* (2007), 399–403.

[90] Keselman, Alla, and Smith, Catherine Arnott. A classification of errors in lay comprehension of medical documents. *Journal of biomedical informatics 45*, 6 (2012), 1151–1163.

[91] Keuleers, Emmanuel, and Brysbaert, Marc. Wuggy: A multilingual pseudoword generator. *Behavior research methods 42*, 3 (2010), 627–633.

[92] Khurana, Rahul N, Lee, Paul P, and Challa, Pratap. Readability of ocular medication inserts. *Journal of glaucoma 12*, 1 (2003), 50–53.

[93] Kim, Hyeoneui, Goryachev, Sergey, Rosemblat, Graciela, Browne, Allen C, Keselman, Alla, and Zeng-Treitler, Qing. Beyond surface characteristics: a new health text-specific readability measurement. In *AMIA* (2007).

[94] Kim, Youngho. *Searching based on query documents*. PhD thesis, University of Massachusetts, 2014.

[95] Kim, Youngho, Seo, Jangwon, and Croft, W Bruce. Automatic boolean query suggestion for professional search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (2011), ACM, pp. 825–834.

[96] Kim, Youngjun, Hurdle, John, and Meystre, Stéphane M. Using umls lexical resources to disambiguate abbreviations in clinical text. In *AMIA Annual Symposium Proceedings* (2011), vol. 2011, American Medical Informatics Association, p. 715.

[97] Kingma, Diederik P., and Ba, Jimmy. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

[98] Klare, George R. Assessing readability. *Reading research quarterly* (1974), 62–102.

[99] Krapivin, Mikalai, Autayeu, Aliaksandr, Marchese, Maurizio, Blanzieri, Enrico, and Segata, Nicola. Keyphrases extraction from scientific documents: Improving machine learning approaches with natural language processing. In *The Role of Digital Libraries in a Time of Global Change*. Springer Berlin Heidelberg, 2010, pp. 102–111.

[100] Kumaran, Giridhar, and Carvalho, Vitor R. Reducing long queries using query quality predictors. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2009), SIGIR '09, ACM, pp. 564–571.

[101] Kutner, Mark, Greenburg, Elizabeth, Jin, Ying, and Paulsen, Christine. The health literacy of america's adults: Results from the 2003 national assessment of adult literacy. nces 2006-483. *National Center for Education Statistics* (2006).

[102] Lafferty, John D., McCallum, Andrew, and Pereira, Fernando C. N. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning* (San Francisco, CA, USA, 2001), Morgan Kaufmann Publishers Inc., pp. 282–289.

[103] Laine, Christine, and Davidoff, Frank. Patient-centered medicine: a professional evolution. *Jama 275*, 2 (1996), 152–156.

[104] Lalor, John P, Wu, Hao, Chen, Li, Mazor, Kathleen M, and Yu, Hong. Comprehenotes, an instrument to assess patient reading comprehension of electronic health record notes: Development and validation. *Journal of medical Internet research 20*, 4 (2018).

[105] Landis, J R, and Koch, G G. The measurement of observer agreement for categorical data. *Biometrics 33* (Mar. 1977), 159–174.

[106] Lee, Chia-Jung, and Croft, W. Bruce. Generating queries from user-selected text. In *Proceedings of the 4th Information Interaction in Context Symposium* (New York, NY, USA, 2012), IIIX '12, ACM, pp. 100–109.

[107] Lee, Jessica Y, Rozier, R Gary, Lee, Shoou-Yih Daniel, Bender, Deborah, and Ruiz, Rafael E. Development of a word recognition instrument to test health literacy in dentistry: the reald-30–a brief communication. *Journal of public health dentistry 67*, 2 (2007), 94–98.

[108] Lee, Se-Jin, Lee, Wei-Na, Kim, Hyojin, and Stout, Patricia A. A comparison of objective characteristics and user perception of web sites. *Journal of interactive advertising 4*, 2 (2004), 61–75.

[109] Lee, Shoou-Yih Daniel, Stucky, Brian D, Lee, Jessica Y, Rozier, R Gary, and Bender, Deborah E. Short assessment of health literacy—spanish and english: a comparable test of health literacy for spanish and english speakers. *Health services research 45*, 4 (2010), 1105–1120.

[110] Lerner, E Brooke, Jehle, Dietrich VK, Janicke, David M, and Moscati, Ronald M. Medical communication: do our patients understand? *The American journal of emergency medicine 18*, 7 (2000), 764–766.

[111] Leroy, Gondy, Endicott, James E, Mouradi, Obay, Kauchak, David, and Just, Melissa L. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. *AMIA Annual Symposium proceedings. AMIA Symposium 2012* (2012), 522–531.

[112] Leroy, Gondy, and Kauchak, David. The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association 21*, e1 (2014), e169–e172.

[113] Leroy, Gondy, Miller, Trudi, Rosemblat, Graciela, and Browne, Allen. A balanced approach to health information evaluation: A vocabulary-based naïve bayes classifier and readability formulas. *Journal of the American Society for Information Science and Technology 59*, 9 (2008), 1409–1419.

[114] Li, Quanzhi, and Wu, Yi-Fang Brook. Identifying important concepts from medical documents. *Journal of biomedical informatics 39* (Dec. 2006), 668–679.

[115] Lin, Yuri, Michel, Jean-Baptiste, Aiden, Erez Lieberman, Orwant, Jon, Brockman, Will, and Petrov, Slav. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations* (2012), Association for Computational Linguistics, pp. 169–174.

[116] Lindau, Stacy T, Tomori, Cecilia, Lyons, Tom, Langseth, Lizbet, Bennett, Charles L, and Garcia, Patricia. The association of health literacy with cervical cancer prevention knowledge and health behaviors in a multiethnic cohort of women. *American journal of obstetrics and gynecology 186*, 5 (2002), 938–943.

[117] Mahdabi, Parvaz, Andersson, Linda, Keikha, Mostafa, and Crestani, Fabio. Automatic refinement of patent queries using concept importance predictors. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2012), SIGIR '12, ACM, pp. 505–514.

[118] Mák, Geneviève, Smith Fowler, Heather, Leaver, Chad, Hagens, Simon, and Zelmer, Jennifer. The effects of web-based patient access to laboratory results in british columbia: A patient survey on comprehension and anxiety. *J Med Internet Res 17*, 8 (Aug 2015), e191.

[119] Malatesha Joshi, R. Vocabulary: A critical component of comprehension. *Reading & Writing Quarterly 21*, 3 (2005), 209–219.

[120] Mc Laughlin, G Harry. Smog grading-a new readability formula. *Journal of reading 12*, 8 (1969), 639–646.

[121] McKeown, Kathleen R., Chang, Shih-fu, Cimino, James, Feiner, Steven K., Friedman, Carol, Gravano, Luis, Hatzivassiloglou, Vasileios, Johnson, Steven, Jordan, Desmond A., A, Desmond, Klavans, Judith L., Kushniruk, Andre, Patel, Vimla, and Teufel, Simone. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *In Proceedings of The First ACM+IEEE JCDL* (2001), pp. 331–340.

[122] McLean, Stuart, and Kramer, Brandon. The creation of a new vocabulary levels test. *Shiken 19*, 2 (2015), 1–11.

[123] Meara, Paul, and Buxton, Barbara. An alternative to multiple choice vocabulary tests. *Language testing 4*, 2 (1987), 142–154.

[124] Medelyan, Olena, and Witten, Ian H. Domain-independent automatic keyphrase indexing with small training sets. *Journal of the American Society for Information Science and Technology 59*, 7 (2008), 1026–1040.

[125] Metzler, Donald, and Croft, W. Bruce. A Markov Random Field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2005), SIGIR '05, ACM, pp. 472–479.

[126] Metzler, Donald, Strohman, Trevor, Zhou, Yun, and Croft, W. B. Indri at TREC 2005: Terabyte track, 2004.

[127] Mikolov, Tomas, Chen, Kai, Corrado, Greg S., and Dean, Jeffrey. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR, 2013* (2013).

[128] Miller, George A. Wordnet: a lexical database for english. *Communications of the ACM 38*, 11 (1995), 39–41.

[129] Morris, Nancy S, MacLean, Charles D, Chew, Lisa D, and Littenberg, Benjamin. The single item literacy screener: evaluation of a brief instrument to identify limited reading ability. *BMC family practice 7*, 1 (2006), 21.

[130] Mossanen, Matthew, True, Lawrence D., Wright, Jonathan L., Vakar-Lopez, Funda, Lavallee, Danielle, and Gore, John L. Surgical pathology and the patient: a systematic review evaluating the primary audience of pathology reports. *Hum. Pathol.* (July 2014).

[131] Nath, Charlotte Reese, Sylvester, Shirley Theriot, Yasek, Van, and Gunel, Erdogan. Development and validation of a literacy assessment tool for persons with diabetes. *The Diabetes Educator 27*, 6 (2001), 857–864.

[132] of Medicine (US). Committee on Quality of Health Care in America, Institute. *Crossing the quality chasm: a new health system for the 21st century*. National Academy Press, 2001.

[133] O'Neil, Braden, Gonçalves, Daniela, Ricci-Cabello, Ignacio, Ziebland, Sue, Valderas, Jose, et al. An overview of self-administered health literacy instruments. *PloS one 9*, 12 (2014), e109110.

[134] Paasche-Orlow, Michael K, Taylor, Holly A, and Brancati, Frederick L. Readability standards for informed-consent forms as compared with actual readability. *New England journal of medicine 348*, 8 (2003), 721–726.

[135] Pang, Liang, Lan, Yanyan, Guo, Jiafeng, Xu, Jun, Xu, Jingfang, and Cheng, Xueqi. DeepRank: A new deep architecture for relevance ranking ininformation retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17* (2017), ACM Press.

[136] Parker, Ruth M, Baker, David W, Williams, Mark V, and Nurss, Joanne R. The test of functional health literacy in adults. *Journal of general internal medicine 10*, 10 (1995), 537–541.

[137] Pendlimari, Rajesh, Holubar, Stefan D, Hassinger, James P, and Cima, Robert R. Assessment of colon cancer literacy in screening colonoscopy patients: a validation study. *Journal of surgical research 175*, 2 (2012), 221–226.

[138] Peters, Ellen, Meilleur, Louise, and Tompkins, Mary Kate. *Numeracy and the affordable care act: Opportunities and challenges*. Institute of Medicine of the National Academies, 2013.

[139] Petersen, Sarah E., and Ostendorf, Mari. A machine learning approach to reading level assessment. *Computer Speech & Language 23*, 1 (jan 2009), 89–106.

[140] Piñero-López, María Ángeles, Modamio, Pilar, Lastra, Cecilia F, and Mariño, Eduardo L. Readability analysis of the package leaflets for biological medicines available on the internet between 2007 and 2013: An analytical longitudinal study. *Journal of medical Internet research 18*, 5 (2016), e100.

[141] Pitler, Emily, and Nenkova, Ani. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Honolulu, Hawaii, Oct. 2008), Association for Computational Linguistics, pp. 186–195.

[142] Pluye, Pierre, Grad, Roland M., Dunikowski, Lynn G., and Stephenson, Randolph. Impact of clinical information-retrieval technology on physicians: a literature review of quantitative, qualitative and mixed methods studies. *Int J Med Inform 74*, 9 (Sept. 2005), 745–768.

[143] Polepalli Ramesh, Balaji, Houston, Thomas, Brandt, Cynthia, Fang, Hua, and Yu, Hong. Improving patients' electronic health record comprehension with NoteAid. *Stud Health Technol Inform 192* (2013), 714–718.

[144] Pyper, Cecilia, Amery, Justin, Watson, Marion, and Crook, Claire. Patients' experiences when accessing their on-line electronic patient records in primary care. *Br J Gen Pract 54*, 498 (2004), 38–43.

[145] Ramesh, Balaji Polepalli, Houston, Thomas K, Brandt, Cynthia, Fang, Hua, and Yu, Hong. Improving patients' electronic health record comprehension with noteaid. In *MedInfo* (2013), pp. 714–718.

[146] Rawson, Katherine A, Gunstad, John, Hughes, Joel, Spitznagel, Mary Beth, Potter, Vanessa, Waechter, Donna, and Rosneck, James. The meter: a brief, self-administered measure of health literacy. *Journal of general internal medicine 25*, 1 (2010), 67–71.

[147] Redish, Janice. Readability formulas have even more limitations than klare discusses. *ACM J. Comput. Doc. 24*, 3 (Aug. 2000), 132–137.

[148] Richman, Julia A, Lee, Jessica Y, Rozier, R Gary, Gong, Debra A, Pahel, Bhavna T, and Vann Jr, William F. Evaluation of a word recognition instrument to test health literacy in dentistry: the reald-99. *Journal of public health dentistry 67*, 2 (2007), 99–104.

[149] Rivas, A. R., Iglesias, E. L., and Borrajo, L. Study of query expansion techniques and their application in the biomedical Information Retrieval. *ScientificWorldJournal 2014* (2014), 132158.

[150] Sabbahi, Dania A, Lawrence, Herenia P, Limeback, Hardy, and Rootman, Irving. Development and evaluation of an oral health literacy instrument for adults. *Community dentistry and oral epidemiology 37*, 5 (2009), 451–462.

[151] Sarkar, Kamal. Automatic keyphrase extraction from medical documents. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009, pp. 273–278.

[152] Shen, Yelong, He, Xiaodong, Gao, Jianfeng, Deng, Li, and Mesnil, Grégoire. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion* (2014), ACM Press.

[153] Silver, N Clayton, and Dunlap, William P. Averaging correlation coefficients: should fisher's z transformation be used? *Journal of Applied Psychology 72*, 1 (1987), 146.

[154] Smith, Barry, and Fellbaum, Christiane. Medical WordNet: A new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th International Conference on Computational Linguistics* (Stroudsburg, PA, USA, 2004), COLING '04, Association for Computational Linguistics.

[155] Snyder, Thomas D, de Brey, Cristobal, and Dillow, Sally A. Digest of education statistics 2015. *National Center for Education Statistics* (2018).

[156] Sørensen, Kristine, Van den Broucke, Stephan, Fullam, James, Doyle, Gerardine, Pelikan, Jürgen, Slonska, Zofia, and Brand, Helmut. Health literacy and public health: A systematic review and integration of definitions and models. *BMC Public Health 12*, 1 (Jan 2012), 80.

[157] Stahl, Steven A, and Fairbanks, Marilyn M. The effects of vocabulary instruction: A model-based meta-analysis. *Review of educational research 56*, 1 (1986), 72–110.

[158] Štajner, Sanja, Evans, Richard, Orasan, Constantin, and Mitkov, Ruslan. What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility* (2012), Citeseer, pp. 14–22.

[159] Swearingen, Christopher J, McCollum, Lauren, Daltroy, Lawren H, Pincus, Theodore, DeWalt, Darren A, and Davis, Terry C. Screening for low literacy in a rheumatology setting: more than 10% of patients cannot read" cartilage,"" diagnosis,"" rheumatologist," or" symptom". *JCR: Journal of Clinical Rheumatology 16*, 8 (2010), 359–364.

[160] Taki, Sarah, Campbell, Karen J, Russell, Catherine G, Elliott, Rosalind, Laws, Rachel, and Denney-Wilson, Elizabeth. Infant feeding websites and apps: A systematic assessment of quality and content. *Interactive journal of medical research 4*, 3 (2014), e18–e18.

[161] Tang, Paul C, and Lansky, David. The missing link: bridging the patient–provider health information gap. *Health Affairs 24*, 5 (2005), 1290–1295.

[162] Tarnowski, Kenneth J, Allen, Denise M, Mayhall, Christine, and Kelly, Patricia A. Readability of pediatric biomedical research informed consent forms. *Pediatrics 85*, 1 (1990), 58–62.

[163] Terwee, Caroline B, Bot, Sandra DM, de Boer, Michael R, van der Windt, Daniëlle AWM, Knol, Dirk L, Dekker, Joost, Bouter, Lex M, and de Vet, Henrica CW. Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology 60*, 1 (2007), 34–42.

[164] Thomas, Garry R, Eng, Lawson, de Wolff, Jacob F, and Grover, Samir C. An evaluation of wikipedia as a resource for patient education in nephrology. In *Seminars in dialysis* (2013), vol. 26, Wiley Online Library, pp. 159–163.

[165] Turtle, Howard, and Croft, W. Bruce. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst. 9*, 3 (July 1991), 187–222.

[166] Van Oosten, Philip, Tanghe, Dries, and Hoste, Veronique. Towards an improved methodology for automated readability prediction. In *7th Conference on International Language Resources and Evaluation (LREC 2010)* (2010), European Language Resources Association (ELRA), pp. 775–782.

[167] Vargas, Christina R, Chuang, Danielle J, Ganor, Oren, and Lee, Bernard T. Readability of online patient resources for the operative treatment of breast cancer. *Surgery 156*, 2 (2014), 311–318.

[168] Vargas, Christina R., Koolen, Pieter G. L., Chuang, Danielle J., Ganor, Oren, and Lee, Bernard T. Online patient resources for breast reconstruction: an analysis of readability. *Plast. Reconstr. Surg. 134*, 3 (Sept. 2014), 406–413.

[169] Vogeli, Christine, Shields, Alexandra E, Lee, Todd A, Gibson, Teresa B, Marder, William D, Weiss, Kevin B, and Blumenthal, David. Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. *Journal of general internal medicine 22*, 3 (2007), 391–395.

[170] Wallace, Lorraine S, Ergen, William F, Cassada, David C, Freeman, Michael B, Grandas, Oscar H, Stevens, Scott L, and Goldman, Mitchell H. Development and validation of the rapid estimate of adult literacy in vascular surgery (real_vs). *Annals of vascular surgery 23*, 4 (2009), 446–452.

[171] Wallace, Lorraine S, Keenum, Amy J, Roskos, Steven E, Blake, Gregory H, Colwell, Strant T, and Weiss, Barry D. Suitability and readability of consumer medical information accompanying prescription medication samples. *Patient education and counseling 70*, 3 (2008), 420–425.

[172] Wang, Lih-Wern, Miller, Michael J, Schmitt, Michael R, and Wen, Frances K. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Research in Social and Administrative Pharmacy 9*, 5 (2013), 503–516.

[173] Ward, Brian W, and Schiller, Jeannine S. Prevalence of multiple chronic conditions among us adults: Estimates from the national health interview survey, 2010. *Preventing chronic disease 10* (2013).

[174] Wiljer, David, Bogomilsky, Sima, Catton, Pamela, Murray, Cindy, Stewart, Janice, and Minden, Mark. Getting results for hematology patients through access to the electronic health record. *Can Oncol Nurs J 16*, 3 (2006), 154–164.

[175] Williams, Mark V, Baker, David W, Honig, Eric G, Lee, Theodore M, and Nowlan, Adam. Inadequate literacy is a barrier to asthma knowledge and self-care. *Chest 114*, 4 (1998), 1008–1015.

[176] Williams, Mark V, Baker, David W, Parker, Ruth M, and Nurss, Joanne R. Relationship of functional health literacy to patients' knowledge of their chronic disease: a study of patients with hypertension and diabetes. *Archives of internal medicine 158*, 2 (1998), 166–172.

[177] Williamson, James Matthew Lloyd, and Martin, AG. Analysis of patient information leaflets provided by a district general hospital by the flesch and flesch–kincaid method. *International journal of clinical practice 64*, 13 (2010), 1824–1831.

[178] Wilson, Meg. Readability and patient education materials used for low-income populations. *Clinical Nurse Specialist 23*, 1 (2009), 33–40.

[179] Witten, Ian H., Paynter, Gordon W., Frank, Eibe, Gutwin, Carl, and Nevill-Manning, Craig G. KEA: Practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries - DL 1999* (1999), ACM Press.

[180] Woodmansey, Karl. Readability of educational materials for endodontic patients. *Journal of endodontics 36*, 10 (2010), 1703–1706.

[181] Wu, Amery D, Begoray, Deborah L, MacDonald, Marjorie, Wharf Higgins, Joan, Frankish, Jim, Kwan, Brenda, Fung, Winny, and Rootman, Irving. Developing and evaluating a relevant and feasible instrument for measuring health literacy of canadian high school students. *Health Promotion International 25*, 4 (2010), 444–452.

[182] Wu, Danny TY, Hanauer, David A, Mei, Qiaozhu, Clark, Patricia M, An, Lawrence C, Proulx, Joshua, Zeng, Qing T, Vydiswaran, VG Vinod, Collins-Thompson, Kevyn, and Zheng, Kai. Assessing the readability of clinicaltrials. gov. *Journal of the American Medical Informatics Association 23*, 2 (2016), 269–275.

[183] Wu, Qiang, Burges, Christopher J. C., Svore, Krysta M., and Gao, Jianfeng. Adapting boosting for information retrieval measures. *Information Retrieval 13*, 3 (sep 2009), 254–270.

[184] Xu, Hua, Stetson, Peter D, and Friedman, Carol. A study of abbreviations in clinical notes. In *AMIA annual symposium proceedings* (2007), vol. 2007, American Medical Informatics Association, p. 821.

[185] Xue, Xiaoibng, and Croft, W. Bruce. Transforming patents into prior-art queries. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2009), SIGIR '09, ACM, pp. 808–809.

[186] Zeng, Qing, Kim, Eunjung, Crowell, Jon, and Tse, Tony. A text corpora-based estimation of the familiarity of health terminology. In *Biological and Medical Data Analysis*. Springer Berlin Heidelberg, 2005, pp. 184–192.

[187] Zeng, Qing T., and Tse, Tony. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc 13*, 1 (Feb. 2006), 24–29.

[188] Zeng-Treitler, Qing, Goryachev, Sergey, Kim, Hyeoneui, Keselman, Alla, and Rosendale, Douglas. Making texts in electronic health records comprehensible to consumers: a prototype translator. *AMIA ... Annual Symposium proceedings. AMIA Symposium* (Oct. 2007), 846–850.

[189] Zheng, Jiaping, and Yu, Hong. Key concept identification for medical information retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (2015), pp. 579–584.

[190] Zheng, Jiaping, and Yu, Hong. Readability formulas and user perceptions of electronic health records difficulty: A corpus study. *J Med Internet Res 19*, 3 (Mar 2017), e59.

[191] Zhu, Dongqing, Wu, Stephen, Carterette, Ben, and Liu, Hongfang. Using large clinical corpora for query expansion in text-based cohort identification. *J Biomed Inform 49* (June 2014), 275–281.