Text Characteristics of Clinical Reports and Their Implications on the Readability of Personal Health Records

Qing Zeng-Treitler[a], Hyeoneui Kim[a], Sergey Goryachev[a], Alla Keselman[b] Laura Slaughter[c] Catherine Smith[d]

[a]*Decision Systems Group ,Brigham and Women's Hospital, Harvard Medical School, U.S.A.*
[b]*Lister Hill National Center for Biomedical Communications, National Library of Medicine, U.S.A.*
[c]*Center for Shared Decision Making and Nursing Research, Rikshospitalet-Radiumhospitalet Medical Center, Norway.*
[d]*School of Library and Information Studies, University of Wisconsin-Madison, U.S.A.*

Abstract

Through personal health record applications (PHR), consumers are gaining access to their electronic health records (EHR). A new challenge is to make the content of these records comprehensible to consumers. To address this challenge, we analyzed the lexical, syntactic and semantic characteristics of three sets of health texts: clinical reports from EHR, known difficult materials and easy-to-read materials. Our findings suggest that EHR texts are more different from easy texts and more similar to difficult texts in terms of syntactic and semantic characteristics, and EHR texts more similar to easy texts and different from difficult texts in regard to lexical features. Since commonly used readability formulas focus more on lexical characteristics, this study points to the need to tackle syntactic and semantic issues in the effort to improve PHR readability.

Keywords:

Consumer Health, Readability, Personal Health Record, Consumer Health Vocabulary, Natural Language Processing

Introduction

Increasingly, consumers are taking an active role in their own health care by accessing and contributing to their personal health records (PHR). This role has become widely recognized by health care organizations and policy makers. One major source of PHR content is the institutional electronic health records (EHR), which are complex documents created by health care professionals for medical, legal, financial and administrative purposes. The organization, syntax, vocabulary and underlying conceptual knowledge employed by medical records are not easily comprehended by lay people. For consumers with an average level of health literacy, understanding HER content is challenging: Is "negative x-ray finding" good or bad? What does "FH of MI" mean? Which section(s) of the discharge summary describe my treatment plan?

For the PHR to fully realize its potential in helping consumers to manage complex health data and to facilitate informed decision making and self-care, its content should be easily accessible to consumers. A prominent panel of fellows of the American College of Medical Informatics recently published a white paper [1] stating "In order to be useful to the patient, the PHR must present data and accompanying tools in ways that enable the individual to understand and to act on the information contained in the record…..Both terminology and data presentation must be adapted to the individual using the PHR, so that they realize optimal benefits."

Considerable readability issues exist for today's PHRs which typically contain selected portions of an EHR and are aimed at a rather educated user group [2, 3]. The need to make the EHR information comprehensible will be more critical as an increasingly diverse patient population gains access to increasingly comprehensive records.

We have embarked on a project to translate EHR information into intelligible structure and plain language for PHR users. As one of the first steps, we collected and analyzed the lexical, syntactic and semantic characteristics of a set of clinical reports. To understand the implications these characteristics have on the content readability for consumers, same text analysis was also performed on two sample sets of health texts – a set of MEDLINE abstracts, which served as the examples of difficult texts and a set of easy-to-read educational materials and news articles, which served as the examples of easy texts.

The text characteristics of clinical reports were compared to those of the difficult and easy text samples. Through the comparison, we identified a number of syntactic and semantic features that potentially contribute to the low readability of EHR texts. We want to note that most of these identified factors are not accounted for by the popular readability measures such as Flesch-Kincaid and Fry [4], which focus on lexical features.

Background

Although more than a few health-specific literacy tests such as the Test of Functional Literacy in Adults (TOHFLA) have been developed [5], no health-specific readability measure is available. Recognizing the potential limitations of existing general-purpose readability measurements, we and other researchers began to examine various characteristics of health texts.

Our previous studies focused on the vocabulary aspect and resulted in the development of term and concept familiarity estimation methods [6]. In evaluation studies [6-8], our predicted term familiarity was shown to be well correlated with actual consumer comprehension and outperformed the word length and word list techniques employed by the general-purpose readability formulas.

In a 2006 report, Rosemblat and colleagues examined what text features health communication experts use to determine the readability of consumer-oriented health texts [9]. The two significant factors they identified were “vocabulary” (i.e. number of words that are likely to be familiar to readers) and “main point” (i.e. ability of readers to identify and understand the “take home” message). In this study, the presence and absence of these factors in the texts were established subjectively by the experts.

Also in 2006, Leroy and colleagues published a study that analyzed and compared the text characteristics of four types of documents: easy and difficult WebMD documents, patient blogs, and patient educational material, for surface and content-based metrics [10]. The easy and difficult WebMD documents were determined using the Flesch-Kincaid formula. They found a number of syntactic and semantic similarities and differences: for example, the easy WebMD pages are the most similar to patient blogs in terms of vocabulary difficulty, while the difficult WebMD pages and patient educational materials have similar counts for words per sentence.

No previous study has examined the readability of clinical reports in EHR systems, except one by Chapman et al which measured a set of dictated and transcribed x ray reports [11]. Although it was not the authors’ intention, we can observe from the results of Chapman’s study that the readability measure (Flesch-Kincaid) was greatly underestimating the difficult of these reports: the average grade level of these reports was reported to be 7.6. Based on our experience of natural language processing (NLP) of radiology reports, they are often difficult for lay researchers with graduate school education (equivalent to grade level 18 and above) to comprehend.

Materials and Methods

Materials

We collected three sets of health documents: EHR reports, difficult texts and easy texts.

The first set contains 40 EHR reports randomly selected from the clinical data repository of the Brigham and Women’s Hospital and Massachusetts General Hospital (Boston, MA, U.S.A.). We retrieved 10 outpatient clinic notes and 10 discharge summaries from each institution. The reports cover topics such as chief complaint, history of illness, laboratory finding, treatment, and discharge plan. The medical diagnoses appeared in the reports include common disease such as asthma, diabetes mellitus, pneumonia, and osteoarthritis. The average length of the documents is 3374 characters.

The second set is 40 abstracts of scientific journal papers randomly retrieved from the MEDLINE (www.pubmed.org). MEDLINE papers are intended for reading by researchers and clinicians, and typically require significant background knowledge in specialty areas (e.g. molecular biology or nephrology) to understand. The abstracts, thus, are good examples of difficult to read content. This set of documents included various topics such as abdominal pain, asthma, hypertension, paranoid schizophrenia. The average length of the documents is 1801 characters – abstracts are short by nature.

The third set is a convenience sample of 40 easy-to-read documents. We collected 27 (self-labeled) easy to read documents from multiple high-qualify consumer health web sites: 21 from the Food and Drug Administration (www.fda.gov), 4 from the National Institute of Mental Health (www.nimh.nih.gov), and 2 from the National Institute on Alcohol Abuse and Alcoholism (www.niaaa.nih.gov). We also selected thirteen records from the Reuter Health (http://www.reutershealth.com). The topics covered by these easy to read materials varied as well, including allergy, heart attack, breast feeding, alcoholism, depression. The average length of the documents is 4101 characters.

Methods

Each document was pre-processed by HITEx – a suite of open-source NLP tools that we have developed [12]. Each document was tokenized, split into sentences, and had part-of-speech (POS) tags assigned. Noun phrases were subsequently extracted and mapped to the Open-Access Collaborative (OAC) consumer health vocabulary[1].

For each parsed document, we first calculated the total number of characters, words, sentences and paragraphs. We considered a word to be any token that does not contain punctuation symbols. Paragraphs were defined depending on the document style: In the EHR reports we used, paragraphs are separated by a blank line; in the easy text sample, they are marked by line breaks. We then calculated the average word length (i.e. number of characters per word), average sentence length (i.e. number of words per sentence), and average paragraph length (i.e. number of sentences per paragraph).

[1] More detailed explanations and related publications of this vocabulary and the term/concept familiarity scores it provided can be found on the Consumer Health Vocabulary Initiative’s Web site (www.consumerhealthvocab.org).

Next, we calculated the frequency distribution of POS categories in each document. For the purpose of statistical analysis, some less frequent POS categories were merged (e.g. all punctuation categories were merged into one), reducing the total number of categories from 30 to 13.

Thirdly, we calculated the average term and concept familiarity scores for each document. These scores are obtained from the OAC consumer health vocabulary[1]. The OAC vocabulary provides three scores: a frequency-based term score, a context-based term score and a context-based concept score. The term scores reflect the string (surface)-level difficulty for consumers and the concept scores reflect the concept-level difficulty for consumers [8]. The scores have the range between 0 and 1, with 1 indicting perfect consumer familiarity (i.e. easiest) and 1 indicting complete consumer unfamiliarity (i.e. most difficult). We used these scores to gauge the semantic complexity of the contents.

Some terms did not mapped to OAC and not all OAC terms had the three scores assigned yet. When we calculated the weighted averages of the scores, these out-of-dictionary terms and terms with missing scores were excluded.

Finally, we calculated the Flesch-Kincaid Grade level for every document using a Microsoft Word[tm] built-in function.

In statistical analysis, mean and 25% and 75% quintiles of each text characteristic were first calculated. We then tested whether the EHR set shares the same text characteristics with either the difficult or the easy document set. The distributions of the characteristics were examined using the Shapiro-Wilkins W test (p <.001) to assess normality. When distributions are not normal, differences in distributions were tested using the Wilcox rank sum test. The text characteristics with normal distributions were tested for differences in means using the t-test.

Results

The text characteristics of the three sample sets (EHR, difficult, and easy text) were different in many aspects. The mean and 25%/75% quartile of the text characteristics are reported in Table 1-3.

Table 1. Means, Lower and Upper quintile, and Differences in Lexical Characteristics

	EHR	Easy	Difficult
Average Num of Characters per Word^	4.97 *(4.78, 5.23)*	4.71* *(4.40, 5.02)*	5.53* *(5.28, 5.83)*
Average Num of Words per Sentence	13.46 *(10.32, 16.35)*	16.93 *(12.38, 21.89)*	17.60** *(14.53, 20.41)*
Average Num of Sentences per Paragraph	3.57 *(1.96, 3.41)*	1.84 *(1.45, 2.0)*	15.50** *(13.0, 18.0)*

Lower and upper quintiles are in parentheses.
** compared to EHR, p <.05*
*** compared to EHR, p <.0001*
^ tested for means using t-test.

On the lexical level, the EHR sample falls between the easy and difficult texts in terms of word length (Figure 1); it has the shortest sentence length (Figure 2), which is not statistically different from that of the easy texts; it also has very few sentences per paragraph (Figure 3), which is not statistically different from that of the easy texts while being very different from that of the difficult texts.

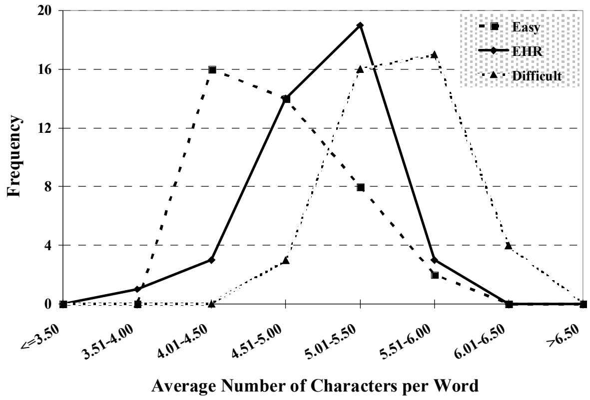

Figure 1. Frequency Distributions of the Number of Characteristics per Word

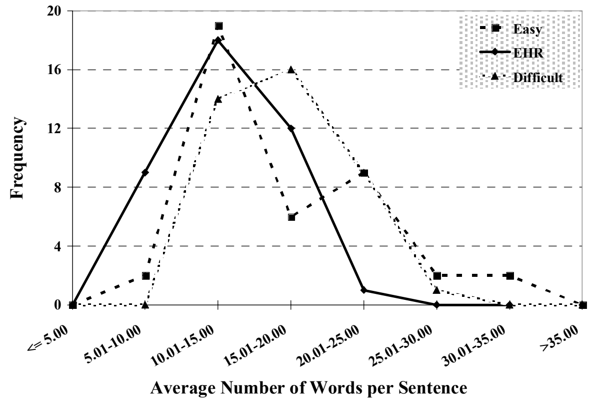

Figure 2. Frequency Distributions of the Number of Words per Sentence

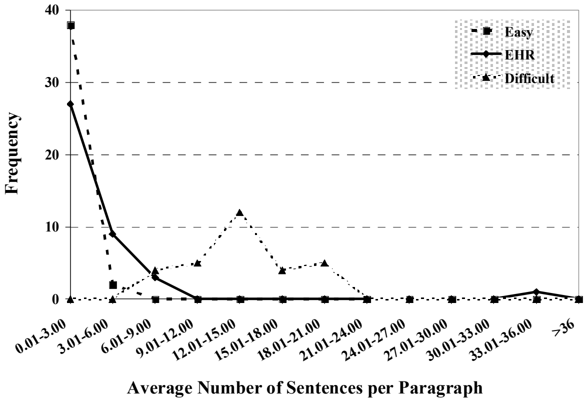

Figure 3. Frequency Distributions of the Number of Sentences per Paragraph.

On the syntactic level, the EHR sample differs from the easy texts statistically in most of the POS categories (Table 2). It does have some similarity with the difficult texts, for example, both have high noun and proper noun usage and less verb and pronoun usage.

Table 2. Means, Lower and Upper quintile, and Differences in Syntactic Characteristics (Parts of Speech Categories Per Sentence)

	HER	Easy	Difficult
Verb	1.62 *(1.31, 2.01)*	2.70** *(2.09, 3.21)*	1.93* *(1.41, 2.29)*
Noun^	3.30 *(2.27, 4.08)*	4.75** *(3.29, 6.08)*	5.48** *(4.47, 6.19)*
Proper Noun	3.32 *(1.62, 4.50)*	1.27** *(0.51, 1.71)*	3.01 *(2.0, 4.0)*
Punctuation	2.44 *(1.77, 2.70)*	2.37 *(1.69, 3.17)*	3.39** *(2.63, 4.0)*
Pronoun	0.36 *(0.15, 1.56)*	0.65** *(0.41, 0.87)*	0.09** *(0, 0.15)*
Adverb	0.33 *(0.16, 0.51)*	0.58** *(0.44, 0.67)*	0.40 *(0.24, 0.47)*
Adjective^	0.87 *(0.54, 1.16)*	1.41** *(0.95, 1.67)*	1.76** *(1.43, 2.02)*
Particle	2.02 *(1.33, 2.53)*	3.12 *(2.16, 3.94)*	3.09** *(2.48, 4.70)*
Determiner	0.03 *(0, 0.06)*	0.10** *(0.04, 0.12)*	0.05 *(0, 0.07)*
Proposition	0.02 *(0, 0.04)*	0.03 *(0, 0.05)*	0.02 *(0, 0.04)*
Symbol	0.09 *(0, 0.08)*	0.00** *(0, 0)*	0.06 *(0, 0.06)*
Modal	0.09 *(0.02, 0.10)*	0.32** *(0.25, 0.39)*	0.09 *(0, 0.13)*
Possessive	0.35 *(0.06, 0.53)*	0.47 *(0.27, 0.62)*	0.05** *(0, 0.10)*

Lower and upper quintiles are in parentheses.
** compared to EHR, p <.05*
*** compared to EHR, p <.0001*
^ tested for means using t-test.

On the semantic level, EHR's mean familiarity scores are the lowest in all three metrics (Table 3). Scores range from 0 to 1, terms/concepts with lower scores are considered less familiar to consumers and more difficult. Statistically significant differences were found between EHR and easy texts in every metric, while EHR and difficult texts only differ on the frequency-based term scores. The score distribution curves of the EHR and difficult texts overlap to a large extent (Figures 4-6).

Table 3. Means, Lower and Upper quintile, and Differences in Semantic Characteristics

	EHR	Easy	Difficult
Average Context-based Term Scores	0.60 *(0.58, 0.66)*	0.72** *(0.68, 0.80)*	0.62 *(0.53, 0.70)*
Average Frequency-based Term Scores^	0.63 *(0.59, 0.67)*	0.77** *(0.74, 0.80)*	0.67* *(0.63, 0.71)*
Average Context-based Concept Scores^	0.66 *(0.65, 0.68)*	0.73** *(0.71, 0.75)*	0.68 *(0.65, 0.71)*

Lower and upper quintiles are in parentheses.
** compared to EHR, p <.05*
*** compared to EHR, p <.0001*
^ tested for means using t-test.

It is a common belief that EHR is not easy for consumers to understand and there is ample empirical evidence supporting the view. The comparison of EHR text characteristics with the characteristics of the difficult and easy health text samples suggest that syntactic and semantics characteristics are key to explain EHR's low readability. On the other hand, the lexical features (word, sentence and paragraph lengths) failed to account for the difficult of EHR texts: EHR texts are similar to easy materials regarding these features.

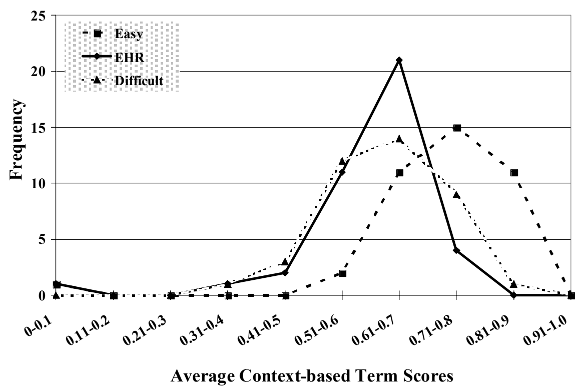

Figure 4. Distributions of the Average Context-based Term Scores

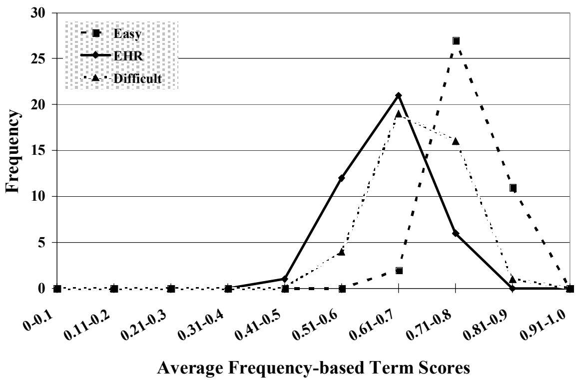

Figure 5. Distributions of the Average Frequency-based Term Scores

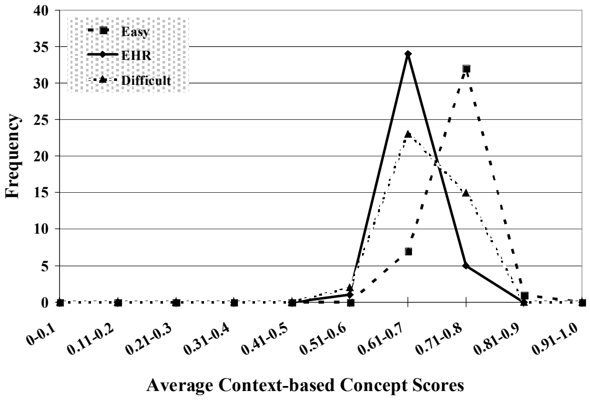

Figure 6. Distributions of the Average Context-based Concept Scores

The Flesch-Kincaid levels of the three samples also differ statistically (Table 4). The average grade of 8.23 is probably an accurate assessment of the easy materials. The 11.98 grade level assigned to the difficult materials is an underestimation, however, can be largely blamed on our use of the abstracts, which are short. The 9.68 grade level of EHR though, is clearly inaccurate.

Table 4. Means, Lower and Upper quintile, and Differences in Readability Scores

	EHR	Easy	Difficult
Flesch-Kincaid Grade Levels	9.68 *(8.55, 11.50)*	8.23* *(5.65, 12)*	11.98** *(12.0, 12.0)*

Lower and upper quintiles are in parentheses.
** compared to EHR, p <.05 ** compared to EHR, p <.0001*

Discussion

Although EHR reports primarily written and read by health professionals, consumers are gaining access to them through the proliferation of PHRs and the readability of EHR reports for consumers has been recognized as a problem. This paper presents an analysis of the lexical, syntactic, and semantic characteristic of EHR texts, and their implications on PHR readability.

Our statistical analysis suggest that EHR texts are more different from known easy texts and more similar to known difficult texts on the syntactic and semantic levels, while EHR texts are more similar to easy texts and different from difficult texts on the lexical level. Meanwhile, the commonly used readability focus on lexical features and not syntactic and semantic features, which we believe is the main cause of Flesch-Kincaid formula's failure to recognize the difficult of the EHR reports. Our findings points to the need to emphasize syntactic and semantic characteristics for the purpose of measuring and improving EHR readability.

Some of the limitations of this study are: Although the text samples we used are comparable to a couple of related studies described in the Background section in size, they are small. MEDLINE abstracts are good examples of difficult health texts; however, they are not a diverse representation of difficult health texts.

For future work, we plan to develop a new readability measure that is based on syntactic and semantic characteristics for PHR texts. We are also interested in exploring the role syntactic and semantic characteristics play in other health texts that consumers are exposed to. The ultimate goal of our research is to develop the means to improve the readability of health content for consumers.

Acknowledgments

This work is supported by the NIH grant R01 LM07222 and by the Intramural Research Program of the NIH, NLM.

References

[1] Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-6.

[2] Hassol A, Walker JM, Kidder D, Rokita K, Young D, Pierdon S, Deitz D, Kuck S, Ortiz E. Patient experiences and attitudes about access to a patient electronic health care record and linked web messaging. J Am Med Inform Assoc 2004;11(6):505-13.

[3] Winkelman WJ, Leonard KJ, Rossos PG. Patient-perceived usefulness of online electronic medical records: employing grounded theory in the development of information and communication technologies for use by patients living with chronic illness. J Am Med Inform Assoc 2005;12(3):306-14.

[4] Zakaluk BL, Samuels SJ. Readability: Its Past, Present, and Future: Intl Reading Assn; 1988.

[5] Baker D, Williams M, Parker R, Gazmarian J, Nurss J. Development of a Brief Test to Measure Functional Health Literacy. Patient Education and Counseling 1999;38:33-42.

[6] Zeng Q, Kim E, Crowell J, Tse T. A Text Corpora-Based Estimation of the Familiarity of Health Terminology. Lecture Notes in Computer Science 2005;3745 / 2005:184-192.

[7] Keselman A, Massengalea L, Ngo L, Browne A, Zeng Q. The Effect of User Factors on Consumer Familiarity with Health Terms: Using Gender as a Proxy for Background Knowledge about Gender-Specific Illnesses. In: International Symposium on Biological and Medical Data Analysis; 2006; 2006. p. In press.

[8] Keselman A, Tse T, Crowell J, Browne A, Ngo L, Zeng Q. Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study. In: MEDNET 2006; 2006; 2006.

[9] Rosemblat G, Logan R, Tse T, Graham L. How Do Text Features Affect Readability? Expert Evaluations on Consumer Health Web Site Text. In: MEDNET; 2006; Toronto, CA; 2006. p. In press.

[10]Leroy G, Eryilmaz E, Laroya BT. Health Information Text Characteristics. In: American Medical Informatics Association (AMIA) Annual Symposium; 2006; Washington DC; 2006. p. In press.

[11]Chapman WW, Aronsky D, Fiszman M, Haug PJ. Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department. Proc AMIA Symp 2000:131-5.

[12]Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30.

Address for correspondence

Qing Zeng-Treitler, Ph.D.

Decision Systems Group

Brigham and Women’s Hospital

Harvard Medical School

75 Francis Street

Boston, MA 02115

Email: qzeng@dsg.harvard.edu