Beyond Surface Characteristics: A New Health Text-Specific Readability Measurement

Hyeoneui Kim[1], Sergey Goryachev[1], Graciela Rosemblat[2], Allen Browne[2], Alla Keselman[2], Qing Zeng-Treitler[1]
[1]Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School, Boston, MA [2]National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

Accurate readability assessment of health related materials is a critical first step in producing easily understandable consumer health information resources and personal health records. Existing general readability formulas may not always be appropriate for the medical/consumer health domain. We developed a new health-specific readability pilot measure, based on the differences in semantic and syntactic features as well as text unit length. The tool was tested with 4 types of materials: consumer health texts, electronic health records, health news articles, and scientific biomedical journals. The results were compared with those produced by three commonly used general readability formulas. While the general formulas underestimated the difficulty of health records by placing them at the same grade levels as consumer health texts, our method rated health records as the most difficult type of documents. Our ratings, however, were highly correlated with general formulas ratings of consumer health, news, and journal articles (r=0.81~0.85, p<.0001).

Keywords: readability, consumer health informatics

Introduction

Readability of a text refers to the ease with which it can be read, and is usually expressed as a grade level.[1] Many government agencies have issued readability guidelines for their documents, in an attempt to ensure understanding by the general public. As consumers play an increasingly active role in managing their own health, availability of easily understandable health information resources becomes critical. Furthermore, presenting medical contents in a more understandable consumer-friendly manner has been recognized as one of the key requirements for successful implementation of personal health records (PHR).[2]

In order to produce more readable health materials, we must first be able to assess their readability level accurately. However, existing general readability measurements may not accurately evaluate the readability of medical texts,[3, 4] and the readability level of medical records in particular is usually underestimated.[4, 5] The goal of this study was to develop a health-domain specific approach to readability measurement. The accuracy of the approach was compared against existing readability formulas on a variety of materials, including Electronic Health Records (EHRs).

Background

Widely used existing readability formulas are designed on the basis of text unit length. The score is represented as a grade level, interpreted as a number of years of education needed to understand a given text. For example, the Flesch-Kincaid Grade Level (FKGL) formula converts Flesch Reading Ease scores into grade levels for ease of interpretation (Figure 1).[6]

$$0.39\left(\frac{total\_words}{total\_sentences}\right)+11.8\left(\frac{total\_syllables}{total\_words}\right)-15.59$$

Figure 1. The FKGL formula

The Simple Measure of Gobbledygook (SMOG) formula, frequently used to measure the readability of health information, is based on the number of sentences and polysyllabic words, or words that contain more than 3 syllables (Figure 2).[7]

$$1.0430\times\sqrt{\frac{polysyllable\_words}{sentences}\times 30}+3.1291$$

Figure 2. The SMOG formula

Gunning-Fog index (GFI) uses sentence length and the percentage of polysyllabic words (Figure 3).[8]

$$0.4\times\left(\left(\frac{words}{sentence}\right)+100\left(\frac{polysyllable\_words}{words}\right)\right)$$

Figure 3. The GFI formula

Several studies have demonstrated that the general use of readability measurements may not be optimal for health related materials. Gemoets et al. pointed out that the existing measures do not reflect completely the readability level of health texts.[9] Rosemblat et al. identified the "ability to

communicate the main point" and familiarity with vocabularies as additional factors that need to be considered in measuring health text readability.[3] Ownby pointed out that in addition to vocabulary complexity, sentence complexity and use of passive voice are the important determinant of text readability.[10] Zeng et al. also showed that EHRs, consumer health materials, and scientific journal articles display many syntactic and semantic aspects that are not taken into account by existing readability measurements.[4]

Materials and Methods

Materials. To create a new readability measurement, we first collected a sample of easy and difficult health-related text.

The easy sample consisted of 200 self-labeled easy-to-read health materials from various web information resources including MedlinePlus®[1] and the Food and Drug Administration consumer information pages[2]. They covered various topics on disease, wellness, and health policy. The average total number of characters per document was 4,737 (sd = 3,086) and the average FKGL was 8 (sd = 1.3).

The difficult sample consisted of 200 scientific biomedical journal and medical textbook articles. Topics included various diseases, wellness, biochemistry, and policy issues. The average total number of words per document was 15,027 (sd = 13,874) and the average FKGL was 16.3 (sd = 2.7).

Text features. Three types of text features were used in our readability measure.

1. As text length features, average numbers of: words per sentence, characters per word, and sentences per paragraph were used.

2. Syntactic features included parts of speech (POS), extracted using a natural language processing tool HITEx.[11] The POS categories were noun, verb, pronoun, proper noun, particle, article, determiner, symbol, punctuation, possessive, preposition, adverb, and adjective. For each POS category, the average number of the category type (e.g., noun, verb, adverb, etc) per sentence was calculated.

3. Semantic features included average term and concept familiarity scores provided by the Open Access and Collaborative (OAC) consumer health vocabulary (CHV)[3]. Three types of scores were employed by this study: context-based term score, frequency-based term score, and concept-based concept score. The two term scores reflect the string-level difficulty for consumers and the concept score reflects the concept-level difficulty for consumers. The scores measure the likelihood that a health term or concept will be understood by lay people, and have been validated with actual consumers.[12, 13]

Distance score. Our distance score readability measure was calculated based on how the text features – described in the previous section – of a test document differ from those of the easy sample.

The mean and standard deviation of each feature was pre-calculated for the easy and difficult samples, respectively. The mean feature values were also calculated for each test document.

On each feature, the distance between the test document and the easy sample was measured by the difference between their means. The features were measured with different scales, for example average numbers of POSs per sentence ranged between 0-10, while familiarity scores range between 0-1. Therefore standard deviations of the easy sample were used to normalize the distances. Since text features within syntactic, semantic, and text unit length categories are highly correlated, the weighted average of the distances was calculated for each category. We assigned weights based on the differences between the easy and difficult sample – a feature was given more weight if easy and difficult samples differed more on the feature. Three categorical distances were calculated first as below, where i is a text feature category, and j is an individual feature in a text feature category i.

$$D_i = \sum \left(\frac{\left| \overline{X}_{ij}^{test} - \overline{X}_{ij}^{easy} \right|}{STD_{ij}^{easy}} \times W_{ij} \right) \times \frac{1}{\sum W_{ij}}$$

Here, weight (W) is defined as below.

$$W_{ij} = \frac{\left| \overline{X}_{ij}^{difficult} - \overline{X}_{ij}^{easy} \right|}{STD_{ij}^{easy}}$$

The final distance score was then calculated as the sum of the three categorical distances. We did not assign different weights to the categories, because an easy document should be easy in all aspects.

Evaluation. The distance score method was tested on the total of 40 articles on the topic of Chronic Obstructive Pulmonary Disease (COPD) (to control for any potential topic-related bias):

[1] See http://medlineplus.gov/
[2] See http://www.fda.gov/opacom/morecons.html
[3] See http://www.consumerhealthvocab.org

- 10 consumer health text materials were collected from 4 reputable sources: Brigham and Women's Hospital[4], WebMD[5], Mayo Clinic[6], and American Association for Respiratory Care[7].
- 10 health related news articles were collected from Reuters and BBC news.
- 10 discharge summary reports came from the Brigham and Women's Hospital's electronic health record.
- 10 full-length text scientific journal articles came from medical journals such as New England Journal of Medicine and the Journal of American Medical Association.

As the EHRs and the journals are written for healthcare professionals, we expected them to be more "difficult" (requiring more years of education for understanding) than the consumer health texts and the news articles specifically written for lay audiences. Our perception of the texts' difficulty upon reviewing them was consistent with this assumption. Based on our assumptions, we expected that an accurate readability measure should rank EHR reports and journal papers as more difficult to read than consumer health and news articles. Examples of the four types of text are provided in Table 1.

Consumer Heath Text
"It is important to follow your doctor's instructions carefully, so that your lungs receive the right amount of medicine."
"If you have COPD, you might be more likely to get colds and flu. Because your heart can be strained, it will get bigger."
EHR
"Did well with med adjustment, no further episodes hypotension, card enz neg. Will d/c pt on home regimen minus the norvasc."
"Presented to urgent care 3 weeks PTA c/o similar symptoms and was given a 14 day course of levaquin and prednisone taper."
News Article
"Most people with a smoker's cough do not realize it could be a symptom of a fatal lung disease, according to a survey."
"People with severe chronic obstructive pulmonary disease (COPD), usually known as emphysema, are helped considerably when they take a combination of two inhaled drugs -- the long-acting beta-2 agonist salmeterol, which relaxes airways, and the inhaled corticosteroid fluticasone to fight inflammation, according to a new study."
Journal Article
"Variables found to be associated with death in the univariate analysis (P<.10) were entered into a stepwise logistic regression model to estimate adjusted odds ratios (ORs) and 95% CIs."
"The primary outcome measure was the rate of decline in the FEV1 after the administration of a bronchodilator, an indicator of the progression of COPD."

Table 1. Examples of the four types of text

We also measured the readability of these documents with the existing general purpose formulas (i.e., FKGL, SMOG, GFI) using open-source web based tools.

Results

We observed overlaps in the ranges of the readability scores between the text types. Median scores produced by each readability measurement are presented in Table 2, along with the minimum and maximum scores in parenthesis.

	CHT	EHR	NEWS	JNL
Distance Score	2.11 (1.46, 3.74)	6.43 (3.13, 7.80)	3.81 (1.58, 5.67)	5.01 (3.43, 7.38)
FKGL	11 (8, 13)	9.5 (6.7,11)	14 (12, 16)	15.5 (12, 17)
GFI	18.5 (15, 21)	18 (15, 19)	21.5 (19, 24)	24 (19, 26)
SMOG	10.41 (7.81, 12.53)	10.68 (9.24, 11.41)	12.87 (10.46, 15.65)	16.61 (12.72 , 18.8)

Table 2. The median, minimum, and maximum readability scores.

In general, all four measurements were consistent in the rating of the readability level of the consumer health materials, the news articles, and the scientific journals, by ranking the first as the easiest among the three, the scientific journals as the most difficult, and the news articles as somewhere in between (Figures 1-4). Significant discrepancies, however, were noted between the readability of the EHRs as measured by our distance score and the already existing tools. While FKGL, GFI, and SMOG ranked EHRs as easy and comparable in difficulty to consumer health texts, our distance score ranked them as the most difficult of the four text types.

In the measuring of EHR readability, none of the existing measures was significantly correlated with

[4] See http://www.brighamandwomens.org
[5] See http://www.webmd.com
[6] See http://www.mayoclinic.com
[7] See http://www.yourlunghealth.org/

our distance scores. However, they showed high correlations with the distance scores in measuring the readability of the consumer health texts, the news articles, and the journal articles (r = 0.81-0.85, p<.0001).

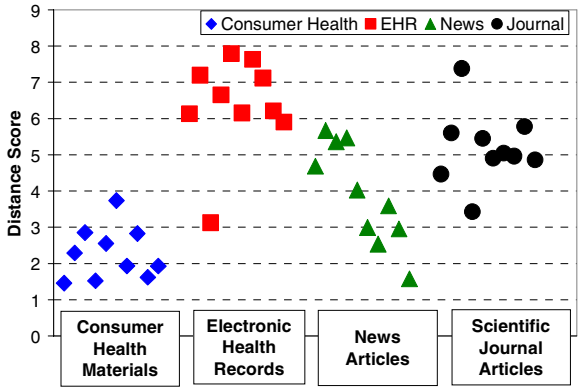

Figure 1. Readability measured by distance scores

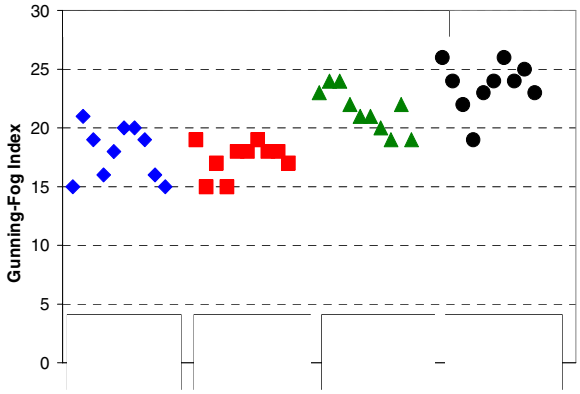

Figure 2. Readability measured by GFI

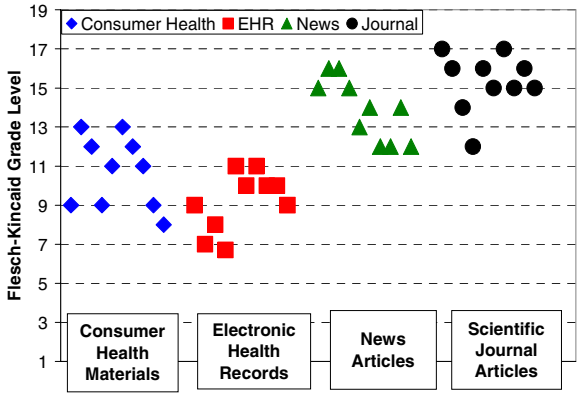

Figure 3. Readability measured by FKGL

Discussion

This study presents and evaluates a new health specific readability measure that can be used towards assessing and improving the readability of consumer health materials. Besides being health specific, our measurement differs from existing general purpose readability formulas in the inclusion of syntactic and semantic features, and in the calculation of distance from a sample of known easy texts.

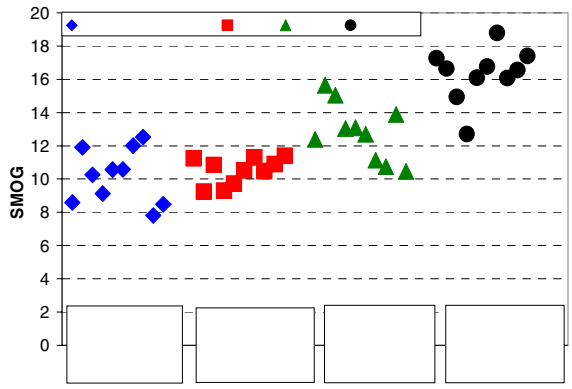

Figure 4. Readability measured by SMOG

The distance-based measure was evaluated against three formulas, commonly used in health communication and literacy research, FKGL, SMOG, and Gunning-Fog index. All algorithms were applied to 4 types of documents (consumer health materials, EHR, news, scientific journals). The main difference between our distance scores and the FKGL, SMOG, and Gunning-Fog index was found in EHR reports: EHR was ranked as the most difficult document type by our measure, but as the easiest (similar to consumer health materials) by the other formulas. Given that EHR reports are well recognized as difficult for consumers to understand, our measurement appears to be more accurate in the assessment of EHR readability.

The evaluation results also showed that our distance-based measurement is strongly correlated with FKGL, SMOG, and Gunning-Fog index, when EHRs were excluded. Since these existing formulas have been validated on general texts that share many characteristics with health-related narrative documents, the correlation provides some validation for our measurement. However, the text unit length based formulas failed to provide a remotely accurate measure for the readability of EHRs. It is our belief that readability metrics need to assess text readability consistently regardless of the document type.

Overall, the study suggests that the distance-base readability measure of consumer health texts is a useful alternative to general text-unit-length formulas.

All measures used in our study produced large overlaps in scores between news articles and journal articles (with mean scores lower for news articles). While some news articles can be fairly difficult (Table1), we believe their readability scores should

be more consistently lower than those of the scientific journal articles. This observation points to the room for improvement in all existing measures.

Other approaches to developing a health-specific readability formula include naïve-Bayes classifier, developed by Leroy et al.[14] It uses a modified "bag-of-words" approach and categorizes a document into one of the easy, intermediate, difficult readability levels. A common challenge faced by Leroy's classifier and our measure is the need for objective reference models for validation. For instance, Leroy used patient blogs as easy and hospital-provided education materials as intermediate samples for training and testing. Our claim of the distance score's better estimation of EHR difficulty was also based on general belief and empirical observation.

This study has several limitations that we plan to address in future endeavors. The difficult sample we used in this study should include more types of documents such as clinical guidelines and lecture notes. Cohesion is an important text feature that we did not measure in this study and would like to incorporate in the future study. Although the evaluation on 4 document types provides some validation of our new readability measurement, it needs to be tested for correlation with actual comprehension by a diverse body of consumers. Finally, it is not clear how to interpret the distance score, especially in the context of individual health literacy level. Clearly, a higher distance score indicates a more difficult document. Which score is appropriate for a specific health literacy level, however, remains to be studied.

Conclusion

We have developed a new health-specific readability measurement. The measurement was tested on 4 types of health document and compared to 3 existing formulas. Results suggest that our new measurement may provide a more accurate assessment of the readability of medical records, and it is consistent with existing formulas on other document types.

Acknowledgments

This work is supported by the NIH grant R01 LM07222 and by the Intramural Research Program of the NIH, NLM/Lister Hill National Center for Biomedical Communications.

References

[1] Zakaluk BL, Samuels SJ. (Eds.). Readability: it's past, present, and future. Newark: International Reading Association. 1988.

[2] Tang PC, Ash JS, Bates DW, Overhage M, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc. 2006;13(2):121-28.

[3] Rosemblat G, Loga R, Tse T, Graham L. Text features and readability: expert evaluation of consumer health text. MEDNET 2006;p. In press.

[4] Zeng-Treitler Q, Kim H, Goryachev S, Keselman A, Slaughter L, Anott Smith C. Text characteristics of clinical reports and their implications for the readability of personal health records. Medinfo 2007; p. In press.

[5] Chapman WW, Aronsky D, Fiszman M, Peter HJ. Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department. Proc AMIA 2000;131-5.

[6] Readability formula: Flesch-Kincaid grade. http://csep.psyc.memphis.edu/cohmetrix/readabilityresearch.htm. Retrieved Feb 23 2007.

[7] McLaughlin GH. SMOG grading: a new readability formula. J Reading. 1969;12;639-46.

[8] The Fog index: a practical readability scale. http://www.as.wvu.edu/~tmiles/fog.html Retrieved Feb 23 2007.

[9] Gemoets D, Rosemblat G, Tse T, Loga R. Assessing readability of consumer health information: an exploratory study. Medinfo 2004;11(Pt 2);869-73.

[10] Ownby RL. Influence of vocabulary and sentence complexity and passive voice on the readability of consumer-oriented mental health information on the internet. AMIA 2005; p585-8.

[11] 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.

[12] 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;3745/2005:184-92.

[13] Keselman A, Tse T, Crowell J, Browne A, Ngo L, Zeng Q. Assessing consumer health vocabulary familiarity: an exploratory study. MEDNET 2006; p. In press.

[14] Leroy G, Rosemblat G, Browne A. A balanced approach to health information evaluation: a vocabulary-based naïve-Bayes classifier and readability formulas. Unpublished manuscript.