HHS Public Access

Author manuscript

J Commun Healthc. Author manuscript; available in PMC 2021 July 23.

Published in final edited form as:

J Commun Healthc. 2020 ; 13(4): 1–13. doi:10.1080/17538068.2020.1822726.

Predicting the readability of physicians' secure messages to improve health communication using novel linguistic features: Findings from the ECLIPPSE study

Scott A. Crossley[a], **Renu Balyan**[b], **Jennifer Liu**[c], **Andrew J. Karter**[c], **Danielle McNamara**[b], **Dean Schillinger**[d]

[a]Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA, USA

[b]Department of Psychology, Arizona State University, Tempe, AZ, USA

[c]Kaiser Permanente Northern California, Oakland, CA, USA

[d]Division of General Internal Medicine and Health Communications Research Program, University of California San Francisco, San Francisco, CA, USA

Abstract

Background: Low literacy skills impact important aspects of communication, including health-related information exchanges. Unsuccessful communication on the part of physician or patient contributes to lower quality of care, is associated with poorer chronic disease control, jeopardizes patient safety and can lead to unfavorable healthcare utilization patterns. To date, very little research has focused on digital communication between physicians and patients, such as secure messages sent via electronic patient portals.

Method: The purpose of the current study is to develop an automated readability formula to better understand what elements of physicians' digital messages make them more or less difficult to understand. The formula is developed using advanced natural language processing (NLP) to predict human ratings of physician text difficulty.

Results: The results indicate that NLP indices that capture a diverse set of linguistic features predict the difficulty of physician messages better than classic readability tools such as Flesch

CONTACT Scott A. Crossley, scrossley@gsu.edu, Department of Applied Linguistics/ESL, Georgia State University, 25 Park Place, Suite 1500, Atlanta, GA 30303, USA.

Notes on contributor

Scott Crossley is a Professor of Applied Linguistics at Georgia State University. His primary research focus is on natural language processing and the application of computational tools and machine learning algorithms in language learning, writing, and text comprehensibility. His main interest area is the development and use of natural language processing tools in assessing writing quality and text difficulty.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Ethical approval statement

Written or oral informed consent was obtained from each patient and the study protocol conforms to the ethical guidelines of the "World Medical Association Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects" adopted by the 18th WMA General Assembly, Helsinki, Finland, June 1964 and amended by the 59th WMA General Assembly, Seoul, South Korea, October 2008, as reflected in a priori approval by the appropriate institutional review committee.

Kincaid Grade Level. Our results also provide information about the textual features that best explain text readability.

Conclusion: Implications for how the readability formula could provide feedback to physicians to improve digital health communication by promoting linguistic concordance between physician and patient are discussed.

Keywords

Health literacy; secure messaging; natural language processing; linguistics; diabetes; communication; machine learning; chronic care management; health care quality; electronic health records

Introduction

The purpose of the current study is develop an automated readability formula of physicians' secure messages (SMs) to patients based on advanced linguistic features to better understand what elements of physicians' written text make them more or less difficult to understand (i.e. a readability formula for physician messages). We focus specifically on digital communication (i.e. written electronic communications) between physicians and type 2 diabetes (DM2) patients in the form of SMs sent within online patient portals. We focus on DM2 patients because asynchronous SMs appear especially important for patients that require ongoing self-management on the part of the patient, as well as associated counseling, guidance and coordination of care on the part of the physician [1]. SM is frequently used in diabetes care to discuss such issues as emerging symptoms, medication concerns, test results, responses to treatment, patient education, appointment and referral requests and processes, and other administrative concerns [2].

To develop our linguistic formula to assess readability of physicians' writing, we use natural language processing (NLP) indices that capture the construct of text readability in terms of lexical sophistication, syntactic complexity, sentiment and cognition variables, and text cohesion. Our readability criterion is developed from expert ratings of text difficulty that categorize the SMs as being easier or more difficult to understand by struggling readers. We measure the performance of this formula by training it on a subset of SMs and extending the resulting model to a testing subset. Finally, we compare the performance of this new formula to a classic readability measure, the Flesch-Kincaid grade level formula (FKGL; [3]), as a reference.

Text readability

It is estimated that one third of American adults are struggling readers that read at basic or less than basic levels [4,5]. Low literacy skills impact important aspects of communication, including health-related information exchanges as found in written medical texts. Health literacy (HL) can include a number of demographic and individual difference factors including education level, culture, access to resources, socioeconomic status, and age, among others. HL also includes a patient's ability to obtain, process, comprehend, and communicate basic health information [6,7] and is highly correlated with literacy skills [8]. Limited HL is common in healthcare contexts and undermines healthcare communication

[9], especially among patients with chronic diseases such as DM2 [10]. Limited HL is a strong predictor of poorer overall health, reduced access to care, lower quality of care, worse disease control and complications, unfavorable healthcare utilization patterns, and higher mortality rates [11,12]. A growing body of evidence demonstrates that reducing the literacy demands of health communications directed to patients may reduce HL-related health disparities [1,13,14].

In terms of linguistic factors, medical texts that are linguistically simple enough to allow readers below the 6th grade level to process them are generally considered easier to read, while texts simple enough to allow readers at the 7th-9th grade levels to process them are considered of average difficulty. Texts at the 9th grade level and above are considered more difficult to read [15,16]. These guidelines have led agencies such as The National Institutes of Health and the American Medical Association to suggest that patient-oriented health texts be written at around the 6th to 8th grade levels [17,18]. However, with so many adults reading at lower levels, it is estimated that about 89 million people in the United States do not have strong enough HL skills to understand and process most health-related materials, even if written at these grade levels [19].

Medical texts that are composed above patients' reading skills can lead to negative health outcomes and health-related behaviors, as well as increased health care costs [16]. Providing patients with more comprehensible texts could lead to greater uptake and recall of medical information [20]. Several recommendations have been made to achieve this objective, and in particular, to improve the readability of patient-directed text materials [17].

Assessing text readability

In a meta-analysis of 155 previous studies, Wang et al [21] reported that six classic readability formulas based on linguistics features are commonly used in medical text analysis: Dale-Chall [22], Flesch Reading Ease (FRE; [23]), Flesch-Kincaid Grade Level (FKGL, [3]), Frequency of Gobbledygook [24], Simple Measure of Gobbledygook (SMOG, [25]), and Fry [26]. The most commonly used formula has been the FKGL, which was used in almost 60% of studies reviewed by Wang et al. [21]. However, research into the effectiveness of these classic readability formulas in medical domains [21,27,28] and text and discourse disciplines [29–34] have reported concerns about their performance and validity.

Assessments of health-related texts using classic readability formulas generally indicate the texts are written at levels too difficult for many adults to successfully comprehend. For instance, Walsh and Volsko [20] examined 100 consumer-health-information articles using three readability formulas and reported that most of the selected articles were written above the 7th-grade reading level. Other studies have found that a vast majority of diverse medical texts are too difficult to read for the intended audience including patient education brochures [35], health-related websites [18,36,37], consumer education materials, health newsletters, clinical trial records, clinical reports, journal articles, [38], patient education documents [39], letters sent from physicians to patients [40], certified web-based educational materials and resource clearinghouses for diabetes patients [41].

Notably, numerous problems in terms of prediction and reliability have been reported for classic readability formulas. As an example, Wang et al. [21] used six different readability formulas to assess depression-related health texts and found that each formula could vary by up to six reading grade level estimates for the same text. In another study, Wu et al. [27] used three readability formulas to compare the readability medical texts, finding that that readability formulas classified referral letters as the most difficult to read in comparison to the clinician notes (whereas the reverse should have been the case). Lastly, Zheng and Yu [28] examined generalist texts (i.e. Wikipedia articles) and specialist texts (i.e. electronic health records) using classic readability formulas and found, paradoxically, that electronic health records were estimated to be easier to read than the Wikipedia articles.

Finally, studies that have examined the effects of improving readability of health-related texts on patient comprehension using classic readability formulas have shown mixed results, with most having only minimal to no effects [2,42,43]. These results indicate that simply changing the length of sentences or introducing shorter words do not have strong effects on text comprehension.

Natural language processing and text readability

One theoretical explanation for the absence of stable effects in classic readability formulas is because the linguistic features of these formulas (e.g. word and sentence length) do not strongly map onto linguistic constructs predictive of reading comprehension [29–34]. These concerns have led many researchers to test different linguistic features and seek alternative readability formulas based on more advanced NLP features.

In the medical domain, researchers have developed models of readability using structural length measures similar to those found in classic readability formulas including word length, sentence length, characters per word, and sentences per paragraph [44–47]. However, these researchers have also supplemented these features with more advanced natural language processing (NLP) measures including lexical sophistication indices such as lexical diversity [44], word familiarity, and word frequency [46,47], semantic features such as average term and concept familiarity scores [45], syntactic features including epistemic modals, relative clauses, passive structures [44], and part of speech tags [45,46], and cohesion features such as word overlap ratios [46]

These newer NLP informed models of readability have shown better performance than classic readability formulas. For instance, Kim et al. [45] reported that their readability formula was more successful at categorizing electronic health records than classic readability formulas. Subsequent studies also showed that Kim et al's [45] formula better captured the readability of three medical documents types (referral letters, non-referral letters, and health articles) than classic readability formulas [27]. Additionally, the readability model developed by Zeng-Treitler et al. [46] was shown to be a better predictor than classic readability formulas for levels of readability (i.e. easy-to-read and hard-to-read medical texts), the difficulty of medical texts as rated by experts, and of medical documents types for which readability was calculated using cloze testing procedures. Similar results for more advanced NLP readability models have been reported outside of medical domains as well (e.g. [48–51])

Method

Tailoring written communication to patients is a difficult process, especially when little is known about what makes a medical text more difficult to comprehend linguistically, when existing readability formulas are problematic, and when physician training is absent. In addition, as medical communication moves online via electronic patient portals, emerging technologies may lead to a greater need to address digital literacy. In spite of this, little research has examined the readability of physicians' messages to patients, despite the fact that linguistic concordance in communication exchanges between physicians and patients is an important pathway to achieve shared meaning, especially for patients with limited health literacy (HL, [9,52]).

To make advances in establishing linguistic concordance, we must first understand what linguistic elements of physicians' secure messages (SM)s make them more or less difficult to understand (i.e. develop a readability formula specifically for physicians' message to patients). Thus, the purpose of the current study is to harness SMs derived from a patient portal, authored by a large sample of physicians working in a large health system to identify the linguistic features of physicians' SMs to patients that are associated with text complexity. To do so, we examine relations between linguistic features and a gold standard of expert ratings of physician text complexity.

Corpus

The SMs that comprise our corpus comes from the Diabetes Study of Northern California (DISTANCE) derived from electronic communication exchanges in Kaiser Permanente Northern California (KPNC), a nonprofit, fully integrated healthcare delivery system. DISTANCE has been described in detail elsewhere [53,54]. Briefly, our study sample is drawn from the DISTANCE [54], and includes data from a subset of 10,504 diabetes patients who had sent one or more English-language SMs to their primary care physicians between July 1, 2006 – Dec. 31 2015. The SMs sent by physicians to patients include messages about lab results, and patient and physician questions regarding scheduling requests, medication refills, and test results, all which match previous analyses of similar data sets [55].

Within DISTANCE, 1,136 primary care physicians sent SMs to patients. We aggregated those SMs by patient that included at least 150 words (the selected threshold for accurately analyzing the text complexity of the SMs, based on prior research; [56]) to create SM threads. Since our goal is to eventually examine concordance between physician and patient language, we used a limited random sampling technique for this study by only including SMs from physicians that had sent SMs to specific patients whose SMs were used in a previous analysis [54]. The previous analysis modeled patient HL and we wanted to collect matching physician complexity data to eventually model concordance. Our final sample comprised 724 unique message threads from 592 individual physicians sent to 486 unique patients. From this sub-sample, these physicians sent SMs to an average of 1.23 patients; 112 of these 592 physicians messaged at least two different patients.

Upon reading the de-identified SMs threads, it was apparent that some physicians employed automated text available in the electronic health record's portal to explain common symptoms, ailments and treatments. To estimate the frequency of automated text, two expert raters examined each message thread. The raters agreed that 220 of the SM threads contained some form of automated text embedded within the SM (involving ~30% of the SMs). We elected to retain SM threads that included automated text in the message threads for our subsequent linguistic analyses because (a) such text is representative of the language used by physicians when messaging patients and (b) it proved difficult to reliably exclude automated text segments using natural language processing (NLP) or other machine learning approaches.

To control for the fact that excess text length might influence the human ratings of readability, SM threads were randomly trimmed to contain approximately ~300 words. When trimming, no individual messages contained in the SM threads was truncated (i.e. we kept SMs intact). The average text length for the SM threads in the final corpus was 294.9 words (SD = 139.5). Among the SMs threads, 359 were composed by female and 365 by male physicians. The average age of the physicians was 53.8 years. The majority of SMs were written by Asian physicians (n = 335) followed by White (n = 291) and Black (n = 38) physicians. Twenty-two of these physicians identified themselves as Hispanic. Of the physicians, 136 reported the ability to write in a language other than English and 201 reported they spoke at least one additional language beyond English.

Expert ratings

The SM threads from the physicians to the patients were evaluated for understandability by two expert raters in terms of how difficult they were for 'struggling readers' to process and comprehend [6,7]. The evaluations thus focused on cognitive aspects of reading and did not focus on demographic or individual differences related to HL including education, culture, access to resources, socioeconomic status, or age. Each rater had an advanced degree, experience teaching literacy-based classes, expertise in medical discourse, and prior experience rating medical text samples. The raters were first provided with a definition of a struggling reader as:

> Individuals who struggle with reading expend considerable cognitive effort and attention in translating print into language, which expends cognitive resources that could better be utilized in constructing, interpreting, and evaluating meaning. They struggle with learning to map the writing system (i.e. the printed visual symbols individually and in combination) to the spoken form of the language (i.e. the phonetics, phonology, and morphology). Their reading behavior is characterized by slow, effortful processing of text. This can impact not only their recognition of individual words, but also ability to build meaning from sentences and paragraphs of text.

The raters were then instructed to use a scoring rubric for each SM thread based on a five-point Likert scale. The rubric asked raters to judge 'How easy would it be it for a struggling reader to understand this message?' The five point-Likert scale was:

1. Very Easy

2. Somewhat Easy
3. Neither easy nor difficult
4. Somewhat Difficult
5. Very Difficult

This is similar to the approach used by Kandula and Zeng-Treitler [38], who had expert raters judge the readability of health texts using a 1–7 scale (1 = can be understood by anyone with basic literacy; and 7 = can be understood only by someone with professional education in a health domain). The raters in our study were first trained on a subset of 30 physician SMs not included in the final corpus. After training on the 30 texts, the raters demonstrated an inter-rater reliability (Cohen's Kappa) of 0.61. Raters then independently assessed each physician SM thread. After independent ratings, any disagreements that reflected a deviation of >1 point on the Likert scale were adjudicated between the two raters. There were 34 such disagreements in the corpus (4.7% of SM threads). Inter-rater reliability after adjudication of all SM threads (Cohen's Kappa) was .57. Final scores for understandability of a message were based on an average score between the two raters. Distribution of the scores was normal.

Linguistic features

We next employed natural language processing (NLP) tools to calculate linguistic features for each SM thread to develop the Model of Text Readability in Physicians (MoTeR-P) using the expert ratings. The selected linguistic features overlapped with previously developed and validated text complexity measures associated with theories of reading and included text cohesion, lexical sophistication, sentiment analysis and syntactic complexity features. Lexical sophistication and semantic features were calculated using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; [57]) and the SEntiment ANalysis and Cognition Engine (SEANCE; [58]). Syntactic complexity was calculated using the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; [59]) and Coh-Metrix [60]. Text cohesion was calculated using the Tool for the Automatic Analysis of Cohesion (TAACO; [61]). Finally, we used Coh-Metrix to calculate traditional readability scores for each text (i.e. Flesch-Kincaid Grade Level and Flesch Reading Ease). A more detailed account of these tools and the linguistic features they report are provided in Dascalu [62], Crossley, Kyle, and McNamara [61], [58, 59], and Kyle and Crossley [57]. We provide a brief overview below.

SEANCE

SEANCE [58] is a sentiment analysis tool that relies on a number of pre-existing sentiment, social positioning, and cognition dictionaries. SEANCE can provide semantic information about the words in a text, which may help readers link common themes and ideas. SEANCE contains a number of pre-developed word vectors developed to measure sentiment, cognition, and social order. These vectors are taken from freely available source databases. For many of these vectors, SEANCE also provides a negation feature (i.e. a contextual valence shifter) that ignores positive terms that are negated (e.g. not happy). Some specific indices provided by SEANCE include component scores that measures the number of words

related to fear and disgust, the number of words related to objects and to the touching and moving objects, the number of words related to action, the number of words related to joy, and the number of words related to friends and family.

TAALES

TAALES [57] calculates over 200 indices related to the lexical sophistication of words in a text. Measure of lexical sophistication correlate with faster processing and more accurate word decoding. TAALES measures simple lexical information (i.e. number of word and n-gram types where n-grams refer to multi-word units), word and n-gram frequency (i.e. how many times a lexical item or combination of lexical items occur in a larger, reference corpus), word range (i.e. how many documents in a reference corpus that a lexical item appears), word properties (e.g. familiarity, concreteness, meaningfulness), and strength of association between words (e.g. are two words strongly associated like *ebb* and *flow*).

Frequency and range indices in TAALES are calculated from available corpora such as SUBTLEXus [63] and the Corpus of Contemporary American English (COCA; [64]). Beyond frequency, bigram (two-word phrases) and trigram (three-word phrases) indices also measure proportion scores (i.e. the proportion of common n-grams found in a reference corpus). Word property information indices come from the Medical Research Council (MRC) psycholinguistic database [65] or other freely available databases such as age of acquisition scores (i.e. at what age people think words are learned) reported by [66].

TAACO

TAACO [61] reports on over 150 indices related to text cohesion. Sensitivity to cohesion structures in a text can help readers process and understand paragraphs and larger discourse segments [43,67]. For many cohesion indices, TAACO integrates a part of speech tagger found in the Stanford Parser [68] and synonym sets reported by the WordNet lexical database [69]. TAACO provides lexical overlap at the sentence (local cohesion) and paragraph (global cohesion) level for function words (e.g. pronouns) content words, and arguments (i.e. similar words such as *familiar, family*, and *familiarity*) and part of speech tags (e.g. nouns and verbs) as well for synonym overlap.

TAASSC

TAASSC ([59, 70]) measures both coarse and fine-grained clausal and phrasal complexity, both of which are related to syntactic complexity. Greater syntactic complexity can make it more difficult to organize main ideas within a text and assign thematic roles, leading to difficulties in parsing sentences effectively [71,72]. At the more granular level, TAASSC reports on indices of clausal complexity, phrasal complexity, and verb argument constructions. These indices include the frequency of verb-argument constructions and average number of direct object dependencies (i.e. the number of dependents per direct object be they adjectives or prepositional phrases).

Coh-Metrix

Coh-Metrix calculates many syntactic features not available in TAASSC. For this study, we were interested in syntactic measures that calculate the mean number of words before the

main verb, the number of higher level constituents per phrases, and syntactic similarity between adjacent sentences (i.e. overlap in the structure and part of speech tags between sentences). Coh-Metrix also calculates a number of measure of lexical diversity (i.e. the variety of words produced) including D [73] and Maas [74]. Lastly, Coh-Metrix calculates Flesch Kincaid Grade Level [3]. The Flesch-Kincaid grade level formula is a recalculation of the Flesch Reading Ease formula [23] which calculated a text's readability based on sentence length and number of syllables per word. The formula is

$$\text{Flesch Kincaid Grade Level} = +(11.8 \times \text{number of syllables/number of words}) - 15.59$$

Statistical analysis

The SM threads were dichotomized as being of either 'low' or 'high' complexity. Secure message (SM) threads scored by expert raters from 1–3 were classified as easier to understand ($n = 471$), while SM threads scored from 4–5 were classified as more difficult to understand ($n = 253$). Prior to statistical analyses, the linguistic variables were pruned. First, we removed those variables that were non-normally distributed. Our thresholds for normal distribution were skewness and kurtosis values of ±2 ([75]). Second, we removed all variables that did not report at least a small effect size ($d = \sim .200$) with the dependent variable (i.e. the low/high classification). Third, variables were checked for multicollinearity (defined as $r > \pm.700$). If variables were multicollinear, the variable with the highest effect size was retained and the other variables were removed. After variable pruning, 85 variables remained.

These 85 variables were then used as variables in a linear discriminant analysis (LDA) to predict expert ratings of complexity and develop Model of Text Readability in Physicians (MoTeR-P). We tested the accuracy of the training model on an independent testing set that was not part of model building (a 66% set of SMs, $n = 478$). The model reported by this LDA was then used to predict group membership for the remaining 34% SMs (the test set, $n = 246$). Our data was not evenly balanced across easy and difficult categories because many texts were scored as 3 (neither easy nor difficult). To account for the imbalance in the data and its potential effects on the LDA algorithm, we set probability weights such that messages with an LDA probability weight of more than .55 were assigned to the easy category, instead of the default 0.50 probability. We also controlled for suppression effects in the algorithm by removing variables that showed flipped signs between their co-efficient scores and their t values. Lastly, to control for over-fitting, we set a threshold for a maximum number of predictors to inform the training model that corresponded to 1 predictor for every 20 items (i.e. a maximum of 24 predictors). We report results from extending the LDA to the testing set in terms of chi-square values, kappa scores, overall accuracy, specificity, and sensitivity. A separate LDA meant to provide a comparative analysis to predict expert ratings of text complexity was also conducted using Flesch Kincaid Grade Level.

Results

MoTeR-P

The final linear discriminant analysis (LDA) model for Model of Text Readability in Physicians (MoTeR-P) contained 24 features (see Table 1 for descriptive and statistical details for these features). The LDA model correctly classified 184 of the 246 messages in the test set, $X^2 = 50.977$, $p < .001$, for an accuracy of .749 (see Table 2 for the confusion matrix for the LDA model based on the Complexity Profile). The measure of agreement between the expert raters' secure message (SM) categorization type and that assigned by the model produced a Cohen's Kappa of 0.455, demonstrating moderate agreement. Sensitivity for the model was .674 while specificity was .788.

The coefficients for the linguistic features indicated that SMs judged as more difficult to understand contained more sophisticated lexical features including more difficult words (i.e. words that occurred in fewer texts, words with higher age of acquisition scores, less concrete words, and less familiar words) and displayed greater lexical diversity. Conversely, more difficult texts included more frequent tri-grams and tri-grams that occurred in a greater number of texts. Messages that were more difficult to understand also contained more frequent academic function words. Syntactically, SMs judged harder to understand had less syntactic overlap across sentences and had sentences that contained more dependencies. In terms of cohesion, more difficult texts had a greater number of function words (including pronouns) and verbs that overlapped across paragraphs. In terms of arguments (i.e. nouns), more difficult texts had less overlap at the sentence and paragraph level. In terms of sentiment, messages that were more difficult to understand contained more words related to fear and disgust, and fewer words related to joy and friends and family. The cognitive variables indicated that SMs that were more difficult to understand contained fewer words related to actions and objects (which correlates with word concreteness).

Comparative analyses

Significant differences were reported for Flesch-Kincaid Grade Level (FKGL) scores between easy to understand SMs ($M = 7.248$ grade level, $SD = 2.153$) and difficult to understand SMs ($M = 8.144$ grade level, $SD = 2.376$), $t = 4.722$, $p < .001$, $d = .388$. An LDA model based on the FKGL score correctly classified 160 of the 246 messages in the test set, $X^2 = 6.560$, $p = .010$, for an accuracy of .650 (see Table 3 for the confusion matrix for the Flesch Kincaid Grade Level model). The measure of agreement between the expert raters' message SM categorization type and that assigned by the Flesch Kincaid model produced a Cohen's Kappa of 0.154, demonstrating only weak agreement. Sensitivity for the model was .302 while specificity was .838.

Discussion

Limited health literacy (HL) is common in healthcare contexts and can undermine medical communication between physicians and patients. Importantly, unsuccessful communication on the part of physician or patient can have important effects on health outcomes [76–78]. One approach to this problem is to better tailor communication between physicians and their

patients with limited HL in the hopes of improving health outcomes and reducing health-related disparities [1,13,14]. As a first step, this study seeks to measure the readability of physicians' secure messages (SMs) sent to patients in an electronic health record in order to understand what linguistic aspects of the SMs make them easier or more difficult to understand by struggling readers.

The study provides further evidence that the difficulty of health-related written materials, in this case physicians' SMs, can be assessed using linguistic features beyond those captured by traditional readability formulas such as Flesch-Kincaid Grade Level (FKGL) and provides support that the ease of comprehending written text is impacted by a complex and wide array of linguistic features [27,44–46,48,49] that can be better captured by natural language processing (NLP) tools. Effectively predicting text difficulty can provide a foundation for understanding what linguistic elements of a text may make it more or less difficult to understand for struggling readers, like those patients with low HL. This foundation can inform our understanding of lower HL and could eventually be used by physicians, administrators, and healthcare systems to guide text production in a manner that improves health communications by meeting the needs of populations with a range of HL, using HL strategies in interpersonal communications to confirm understanding, providing easier access to health information/services, designing and distributing content that is easy to understand and act on, and preparing workforces to be health literate [79]. Determining if models like Model of Text Readability in Physicians (MoTeR-P) can be used to help compose more comprehensible SMs so as to better match patients' reading levels is an important next step. If greater linguistic concordance can be reached, future research can evaluate whether such concordance enhances patient portal communications, thereby promoting favorable patterns of adherence, patient safety and healthcare utilization [16,76,80–83].

Overall, this study found that physician SM difficulty was predicted by challenging words in the text. While FKGL also indexes word difficulty in its formula, it does so by relying solely on the average number of syllables in words, a proxy for word difficulty. In this study, we find that reading challenges in the SMs are predicted by a multitude of lexical indices including word frequency as well as age of acquisition, familiarity, concreteness, lexical diversity, and multi-word phrases (i.e. tri-grams). Similar findings have been reported in terms of lexical diversity [44] word familiarity, and multi-word phrase frequency [46] and word frequency [47]. As such, the use of more advanced NLP approaches not only affords more precise calculations of reading accuracy but also provides greater information about the lexical features in the text that lead to greater reading difficulty.

Our results also support the notion that physician SMs were harder to read if they contained more syntactic dependencies and less syntactic overlap across sentences. Neither of these features would be captured by the syntactic measures found in classic readability formulas, which focus solely on sentence length. Similar findings have been reported in previous studies of health text readability in that more readable texts contain fewer relative clauses (i.e. fewer embedded clauses) and fewer passive clauses [44].

A more principled problem with classic readability formulas is that they do not consider text cohesion features, which are important elements of text readability that logically connect sections of text [84–86]. At least one previous medical readability formula included a measure of cohesion, but this measure did not distinguish between texts categorized as easier or more difficult to read [46]. The findings from our study did find significant differences in cohesion features between easy and difficult texts. Specifically greater noun overlap increase text readability whereas function word and verb overlap decreased text readability. Intuitively this makes sense because the core meaning of many non-narrative texts lies more in the items discussed than the actions or the structural components surrounding those items. Thus, we would expect that greater overlap of nouns between sentences and paragraphs would increase cohesion whereas verb and function word overlap may not. It is also important to note that greater pronoun overlap was predictive of less readable text. Because pronouns have no explicit meaning and instead depend on readers' ability to determine the pronoun's previous referent (i.e. anaphoric resolution), it makes sense that struggling readers may find that greater overlap of pronouns between text segments makes the text less readable.

Lastly, classic readability formulas do not consider affective or cognitive variables. To our knowledge, no research has indicated that affective variables are important indicators of text readability. However, we included them because SMs between physicians and patients – where medical decisions related to personal health are concerned – may rely on positive affect on the part of the physician to reduce anxiety for the reader and, thus, indirectly make a text more readable for the patient. The readability formula developed in this study supports this assertion in that more readable texts contained fewer words related to fear and disgust and more words related to joy. In addition, since friends and family are an important component of any support network, especially medical support networks where friends and family members are more likely to be involved in decision making, the use of more words related to friends and family may help ground and internalize the message, making it more understandable. Additionally, mentions of friends and families may personalize messages, making them less objectionable and more readable. The cognitive variable related to object terms contained in the model overlap with our lexical variable of concreteness, providing additional support that more concrete/object-related words help produce a more readable text. The action word feature indicated that the production of more words related to actions (i.e. verbs) helps with readability. While difficult to interpret, at least one previous study [46] found that a greater number of verbs was associated with ease of text readability. It may be that more verbs help readers construct more accurate semantic relationships and/or greater temporal cohesion [87].

While we expected MoTeR-P to outperform FKGL, there are still interesting elements of the FKGL analysis to discuss. Specifically, unlike many other previous studies [5,18,36–39], our study found that the grade levels reported (7th grade or lower for the easy to understand texts and 8th grade or higher for the more difficult texts) were aligned with what is considered average text difficulty and related text difficulty thresholds [15,16]. This may be a result of the specific domain (email-type SMs) or may reflect active strategies on the part of physicians to attempt to make their SMs understandable for an average reader. Notably, sensitivity was low while specificity was high such that FKGL was more likely to classify

text, regardless of level, as easy. These results show that while the FKGL scores were in the average range for difficulty, the formula performed poorly at classifying texts as more or less difficult to read.

We recognize that achieving readability is only one step in a complex process of attempting to overcome a range of determinants of, and barriers to, optimal healthcare that tend to co-occur with limited HL. A recent conceptual model has illustrated the complex ways in which HL interacts with pathways connecting the social determinants of health to health disparities [88]. This model indicates that patient HL is situated not only as a mediator or effect modifier in a causal model, but also as an important factor in an ecological model that also includes characteristics and attributes of the health care system that define the contexts in which patients and physicians operate. We believe that the contribution of our current study is to provide health systems with an automated tool to begin to measure and systematically address one of the contextual factors that distinguish those healthcare organizations that are 'health literate' from those that are not: the ease with which their clinicians can be understood by patients who struggle with limited HL.

While we introduce and provide preliminary validation for a new readability formula for health-related text, we also recognize that text comprehension is greater than the sum of its linguistic parts. It is possible that we could increase the algorithm's prediction accuracy with the addition of specific linguistic factors that may be especially important to HL, including the difficulty of medical acronyms and medical terms [44,47] and the quality of both explanatory or elicitative content that may be contained in physician text [7]. Additionally, the linguistic features that we examined may be insufficient to address the needs of struggling readers in diverse clinical settings; linguistic indicators of empathy, interactivity and shared meaning, for example, could be tested in future research. Beyond linguistic features, it would also be important to incorporate other variables relevant to HL, including patients' background knowledge, age, overall reading ability and socio-cultural background [44]. Analyzing visual aspects of text, such as figures, charts, white space, and layout design [17], and measuring the content and function of the SMs (i.e. if the function of the SM was a scheduling request, a medication refill response, a report of test results, advice around self-management, or a response to a patient query) may influence the extent to which a secure message can be easily understood and acted on. While some of these non-linguistic features could prove feasible to collect and include in future readability formulas (e.g. demographic information, non-linguistic textual information), developing automated techniques to examine the content and function of SMs may be more challenging. However, incorporating such techniques has been the subject of recent research [89]. For instance, topic models could be developed using word embedding techniques such as latent Dirichlet allocation (LDA). It is possible that some topics would be more difficult for struggling readers than others. While assessing the function of a SM may prove more difficult, recent research in academic writing indicates success in teasing apart genre, discourse, and rhetorical moves that help indicate the function of writing [90,91]. Such approaches might be adapted to physicians' SMs to develop future iterations and further enrich the work we have presented here.

It should be noted that our reading criterion relied on expert ratings of what struggling readers would perceive as understandable. We are currently collaborating with diverse type 2 diabetes (DM2) patients recruited from a public hospital system – some of whom are struggling readers – to examine whether expert ratings of SM complexity are concordant with patient ratings of these same messages. Relatedly, while the use of patient portals is rapidly expanding among all racial and socioeconomic subgroups, there is some evidence that SM may be less accessible to lower literacy readers. As such, our category of 'easy to read' may not be generalizable to all struggling readers and may explain the relatively low inter-rater reliability reported for this study. Future work to replicate our model using SMs obtained from public-sector health systems that care for a disproportionately larger number of struggling readers may lead to modifications in linguistic indices included or adaptations to cut-points to better serve the needs of the lowest level readers. We also recognize that a model that generates binary classifications may not best represent all applications and, while MoTeR-P could be applied to health-related text beyond SMs, its applicability in such settings is unknown. Finally, the work to date is limited to primary care physicians and their SMs. Future work should explore MoTeR-P with other members of the healthcare team who engage in SMs, from medical assistants to nurses to sub-specialists. We would like to include background information of the healthcare team members as predictors of readability given the available data and the use of appropriate machine learning approaches.

Advances in computational linguistics will likely enable the development of even more comprehensive measures that build on our new readability measure to assess the likelihood that patients who are struggling readers understand physicians' communications. Future iterations might incorporate automated indicators of the content and function of emails, the quality of elicitative and explanatory content, the visual attributes of the communications, and the emotional tone employed. In parallel, the next steps in our research are first to apply MoTeR-P to the entire DISTANCE sample of physician SMs sent to patients in order to examine potential differences in SM readability based on physician characteristics, as well as differences both within and across physicians with respect to the extent to which they tailor SM complexity to meet the HL of individual patients. Second, because complex physician communication could lead to poorer patient adherence and worse health outcomes, we also intend to examine whether greater concordance between physician SM complexity as measured by MoTeR-P and patient HL readability formulas is predictive of a range of diabetes-related health outcomes. Third, we plan to explore whether physicians have characteristic 'signatures' with respect to achieving concordance across their patients so as to determine whether overall communication style is an important factor in predicting health outcomes above and beyond individual-level concordance. For example, it will be important to determine whether the patient outcomes of physicians whose communications reflect a 'universal precautions' pattern (wherein easily comprehensible language is employed regardless of a patient's HL, [92]) differ from those of physicians whose communications reflect a 'universal tailoring' pattern (wherein the complexity of a physician's language tends to match that of the patient's HL level). Fourth, depending on our findings, we plan to use MoTeR-P to inform physician training and feedback efforts as part of pre-graduate, post-graduate, and continuous medical education efforts. Fifth, we plan on exploring the degree to which the expert ratings of SMs align with DM2 patients' ratings of the same text, as a

means to further validate MoTeR-P. Finally, an additional application of MoTeR-P is to develop a readability tool that could be practically implemented within health systems. Such a tool would provide real-time feedback to physicians about the readability of their SMs, which could lead to greater SM readability and improve patient comprehension and related health outcomes. Such a system not only could provide summative feedback about the readability of a SM but also specific suggestions about how texts could be modified to make them more comprehensible. Early work at manually simplifying medical texts has demonstrated that linguistic modifications can make texts more readable [39], and other researchers have sought automatic methods to modify texts [46,93]. This work will provide a foundation for scaling readability formulas into a practical health system-wide application to both optimize care for patients of all HL levels, as well as help reduce HL-related health disparities.

The ECLIPPSE Project set out to harness secure messages sent by diabetes patients to their primary care physician(s) to develop literacy profiles that can identify patients with limited health literacy in an automated way that avoids time-consuming and potentially sensitive questioning of the patient. Given the time and personnel demands intrinsic to current health literacy instruments, measuring health literacy has historically been extremely challenging. An automated literacy profile could provide an efficient means to identify subpopulations of patients with limited health literacy. Identifying patients likely to have limited health literacy could prove useful for alerting clinicians about potential difficulties in comprehending written and/or verbal instructions. Additionally, patients identified as having limited health literacy could be supported better by receiving follow-up communications to ensure understanding of critical communications, e.g. medication instructions, and promote adherence and increased shared meaning [29]. As such, our research to develop automated methods for health literacy assessment represents a significant accomplishment with potentially broad clinical and population health benefits in the context of health services delivery.

Acknowledgments

Funding

This project was also funded by a grant by the National Institutes of Health, NIDDK Centers for Diabetes Translational Research (P30 DK092924), and by the NLM 5R01LM012355-04 (PI: Dean Schillinger, MD).

References

[1]. Schillinger D, Handley M, Wang F, Hammer H. Effects of self-management support on structure, process, and outcomes among vulnerable patients with diabetes: a three-arm practical clinical trial. Diabetes Care. 2009;32(4):559–566. [PubMed: 19131469]

[2]. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. Health Educ Behav. 2006;33(3):352–373. [PubMed: 16699125]

[3]. Kincaid JP, Fishburne RP, Rogers RL, Chisson BS. Derivation of new readability formulas (automated readability index, Fog count, and Flesch reading ease formula) for Navyenlisted personnel. Pensacola (FL): Navy Training Command Research Branch; 1975.

[4]. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med. 1996;1(1):7–28.

[5]. Weiss BD. Health literacy research: isnt' there something better we could be doing? Health Commun. 2015;30 (12):1173–1175. [PubMed: 26372029]

[6]. Grossman EG. Patient protection and affordable care act. Washington (D.C.): Department of Health & Human Services; 2010.

[7]. Schillinger D, McNamara D, Crossley S, Lyles C, Moffet HH, Sarkar U, et al. The next frontier in communication and the ECLIPPSE study: Bridging the linguistic divide in secure messaging. J Diabetes Res. 2017;2017: 1–9.

[8]. Nutbeam D. Defining and measuring health literacy: what can we learn from literacy studies? Int J Public Health. 2009;54(5):303–305. [PubMed: 19641847]

[9]. Schillinger D, Bindman A, Wang F, Stewart A, Piette J. Functional health literacy and the quality of physician–patient communication among diabetes patients. Patient Educ Couns. 2004;52(3):315–323. [PubMed: 14998602]

[10]. Schillinger D, Wang F, Rodriguez M, Bindman A, Machtinger EL. The importance of establishing regimen concordance in preventing medication errors in anticoagulant care. J Health Commun. 2006;11(6):555–567. [PubMed: 16950728]

[11]. Schillinger D, Chen AH. Literacy and language. J Gen Intern Med. 2004;19(3):288–290. [PubMed: 15009787]

[12]. Sudore RL, Yaffe K, Satterfield S, Harris TB, Mehta KM, Simonsick EM, et al. Limited literacy and mortality in the elderly: the health, aging, and body composition study. J Gen Intern Med. 2006;21(8):806–812. [PubMed: 16881938]

[13]. Baker DW, DeWalt DA, Schillinger D, Hawk V, Ruo B, Bibbins-Domingo K, et al. "Teach to goal": theory and design principles of an intervention to improve heart failure self-management skills of patients with low health literacy. J Health Commun. 2011;16 (sup3):73–88. [PubMed: 21951244]

[14]. DeWalt DA, Broucksou KA, Hawk V, Baker DW, Schillinger D, Ruo B, et al. Comparison of a one-time educational intervention to a teach-to-goal educational intervention for self-management of heart failure: design of a randomized controlled trial. BMC Health Serv Res. 2009;9(1):99. [PubMed: 19519904]

[15]. U.S. National Library of Medicine. How to write easy to read health materials. MedlinePlus; 2017. Available from: https://www.nlm.nih.gov/medlineplus/etr.html.

[16]. Weiss BD. Health literacy: A manual for clinicians. Chicago (IL): American Medical Association Foundation and American Medical Association; 2003; Available from: http://www.ihconline.org/toolkits/HealthLiteracy/AMAHealthLiteracyManualClinicians.pdf.

[17]. Badarudeen S, Sabharwal S. Assessing readability of patient education materials: current role in orthopaedics. Clin Orthop Relat Res. 2010;468(10):2572–2580. [PubMed: 20496023]

[18]. Kugar MA, Cohen AC, Wooden W, Tholpady SS, Chu MW. The readability of psychosocial wellness patient resources: improving surgical outcomes. J Surg Res. 2017;218:43–48. [PubMed: 28985876]

[19]. Institute of Medicine. Health literacy: A prescription to end confusion. Washington (D.C.): The National Academies Press; 2004.

[20]. Walsh TM, Volsko TA. Readability assessment of internet-based consumer health information. Respir Care. 2008;53(10):1310–1315. [PubMed: 18811992]

[21]. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Soc Admin Pharm. 2013;9 (5):503–516.

[22]. Chall JS, Dale E. Readability revisited: The new Dale-Chall readability formula. Cambridge (MA): Brookline Books; 1995.

[23]. Flesch R A new readability yardstick. J Appl Psychol. 1948;32:221–233. [PubMed: 18867058]

[24]. Gunning R The technique of clear writing. New York: McGraw-Hill; 1968.

[25]. McLaughlin GH. SMOG grading-A new readability formula. J Read. 1969;12(8):639–646.

[26]. Fry EB. A readability formula that saves time. J Read. 1968;11:513–516.

[27]. Wu DT, Hanauer DA, Mei Q, Clark PM, An LC, Lei J, et al. Applying multiple methods to assess the readability of a large corpus of medical documents. Stud Health Technol Inform. 2013;192:647–651. [PubMed: 23920636]

[28]. Zheng J, Yu H. Readability formulas and user perceptions of electronic health records difficulty: A corpus study. J Med Internet Res. 2017;19(3):e59. doi:10.2196/jmir.6962. [PubMed: 28254738]

[29]. Bruce B, Rubin A. Readability formulas: matching tool and task. In: Green GM, editor. Linguistic complexity and text comprehension: readability issues reconsidered. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988. p. 5–22.

[30]. Bruce B, Rubin A, Starr K. Why readability formulas fail. IEEE Trans ProfCommun. 1981;PC-24:50–52.

[31]. Davison A, Kantor RN. On the failure of readability formulas to define readable texts: A case study from adaptations. Read Res Q. 1982;17:187–209.

[32]. Graesser AC, McNamara DS, Kulikowich JM. Coh-Metrix: providing multilevel analyses of text characteristics. Edu Res. 2011;40(5):223–234.

[33]. Rubin A How useful are readability formulas? In Osborn J, Wilson PT, Anderson RC, editors. Reading education: Foundations for a literate America. Lexington, MA: Lexington Books; 1985. p. 61–77.

[34]. Smith F Understanding reading: A psycholinguistic analysis of reading and learning to read. New York, NY: Routledge; 2012.

[35]. Schumaier AP, Kakazu R, Minoughan CE, Grawe BM. Readability assessment of American Shoulder and Elbow Surgeons patient brochures with suggestions for improvement. JSES Open Access. 2018;2:150–154. [PubMed: 30675586]

[36]. Berland GK, Elliott MN, Morales LS, Algazy JI, Kravitz RL, Broder MS, et al. Health information on the Internet: Accessibility, quality, and readability in English and Spanish. JAMA. 2001;285(20):2612–2621. [PubMed: 11368735]

[37]. Boulos MNK. British internet-derived patient information on diabetes mellitus: is it readable? Diabetes Technol Ther. 2005;7(3):528–535. [PubMed: 15929685]

[38]. Kandula S, Zeng-Treitler Q. Creating a gold standard for the readability measurement of health texts. In AMIA Annual Symposium Proceedings, Vol. 2008, American Medical Informatics Association, p. 353; 2008.

[39]. Hill-Briggs F, Schumann KP, Dike 0.5-Step methodology for evaluation and adaptation of print patient health information to meet the< 5th grade readability criterion. Med Care. 2012;50(4):294. [PubMed: 22354210]

[40]. McAndie E, Gilchrist A, Ahamat B. Readability of clinical letters sent from a young people's department. Child Adolesc Ment Health. 2016;21(3):169–174. [PubMed: 32680354]

[41]. Kusec S, Brborovic O, Schillinger D. Diabetes websites accredited by the health On the Net foundation Code of Conduct: readable or not? Stud Health Technol Inform. 2003;95:655–660. [PubMed: 14664062]

[42]. Barton JL, Trupin L, Schillinger D, Evans-Young G, Imboden J, Montori VM, et al. Use of low-literacy decision aid to enhance knowledge and reduce decisional conflict among a diverse population of adults with rheumatoid arthritis: results of a pilot study. Arthritis Care Res (Hoboken). 2016;68(7):889–898. [PubMed: 26605752]

[43]. Meade CD, Byrd JC, Lee M. Improving patient comprehension of literature on smoking. Am J Public Health. 1989;79(10):1411–1412. [PubMed: 2782514]

[44]. Gemoets D, Rosemblat G, Tse T, Logan RA. Assessing readability of consumer health information: an exploratory study. In Medinfo, pp. 869–73; 2004.

[45]. Kim H, Goryachev S, Rosemblat G, Browne A, Keselman A, Zeng-Treitler Q. Beyond surface characteristics: a new health text-specific readability measurement. American Medical Informatics (AMIA) Annual Symposium. Washington, D.C, pp. 418–422; 2007.

[46]. Zeng-Treitler Q, Kandula S, Kim H, Hill B. A method to estimate readability of health content. Association for Computing Machinery; 2012.

[47]. Zheng J, Yu H. Assessing the readability of medical documents: A ranking approach. JMIR Med Inform. 2018;6(1):e17. [PubMed: 29572199]

[48]. De Clercq O, Hoste V, Desmet B, Van Oosten P, De Cock M, Macken L. Using the crowd for readability prediction. Nat Lang Eng. 2014;20(3):293–325.

[49]. Crossley SA, Greenfield J, McNamara DS. Assessing text readability using cognitively based indices. TESOL Quart. 2008;42(3):475–493.

[50]. Crossley SA, Skalicky S, Dascalu M, McNamara DS, Kyle K. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. Discourse Process. 2017;54 (5–6):340–359.

[51]. Pitler E, Nenkova A. (2008). Revisiting readability: A unified framework for predicting text quality. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 186–95.

[52]. Sarkar U, Schillinger D, Bibbins-Domingo K, Napoles A, Karliner L, Perez-Stable EJ. Patient-physicians' information exchange in outpatient cardiac care: time for a heart to heart? Patient Educ Couns. 2011;85 (2):173–179. [PubMed: 21035298]

[53]. Moffet HH, Adler N, Schillinger D, et al. Cohort Profile: The Diabetes Study of Northern California (DISTANCE)—objectives and design of a survey follow-up study of social health disparities in a managed care population. International journal of epidemiology. 2008;38(1):38–47. [PubMed: 18326513]

[54]. Ratanawongsa N, Karter AJ, Parker MM, et al. Communication and medication refill adherence: the Diabetes Study of Northern California. JAMA internal medicine. 2013;173(3):210–218. [PubMed: 23277199]

[55]. Yazdannik A, Yousefy A, Mohammadi S. Discourse analysis: A useful methodology for health-care system researches. J Educ Health Promot. 2017;6:111. [PubMed: 29296612]

[56]. Crossley SA. How many words needed? Using natural language processing tools in educational data mining. Proceedings of the 10th international conference on educational data mining (EDM). 2018: 630–633.

[57]. Kyle K, Crossley SA. Automatically assessing lexical sophistication: indices, tools, findings, and application. TESOL Quart. 2015;49(4):757–786.

[58]. Crossley SA, Kyle K, McNamara DS. Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social order analysis. Behav Res Methods. 2017;49(3):803–821. [PubMed: 27193159]

[59]. Kyle K Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication [Doctoral dissertation]. Georgia State University; 2016. Available from: http://scholarworks.gsu.edu/alesl_diss/35.

[60]. Graesser AC, McNamara DS, Louwerse MM, Cai Z. Coh-Metrix: analysis of text on cohesion and language. Behav Res Meth Instrum Comput. 2004;36:193–202.

[61]. Crossley SA, Kyle K, McNamara DS. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behav Res Methods. 2016;48(4):1227–1237. [PubMed: 26416138]

[62]. Dascalu M Analyzing discourse and text complexity for learning and collaborating (Studies in computational intelligence, Vol. 534). Berlin, Germany: Springer; 2014.

[63]. Brysbaert M, New B. Moving beyond Kucera 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. 2009;41:977–990. 10.3758/BRM.41.4.977 [PubMed: 19897807]

[64]. Davies M The 385+ million word corpus of Contemporary American English (1990–2008+): design, architecture, and linguistic insights. Int J Corpus Linguist. 2009;14:159–190.

[65]. Coltheart M The MRC psycholinguistic database. Q J Exp Psychol. 1981;33:497–505.

[66]. Kuperman V, Stadthagen-Gonzalez H, Brysbaert M. Age-of-acquisition ratings for 30,000 English words. Behav Res Methods. 2012;44:978–990. [PubMed: 22581493]

[67]. Gernsbacher MA. Language comprehension as structure building. Hillsdale (NJ): Erlbaum; 1990.

[68]. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55–60; 2014.

[69]. Miller GA. Wordnet: a lexical database for English. Commun ACM. 1995;38(11):39–41.

[70]. Kyle K, Crossley SA. Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. Modern Language Journal. 2018;102(2):333–349.

[71]. Graesser AC, Swamer SS, Baggett WB, Sell MA. New models of deep comprehension. In: Britton BK, Graesser AC, editor. Models of understanding text. Mahwah, NJ: Erlbaum; 1996. p. 1–32.

[72]. Mesmer HA, Cunningham JW, Elfrieda HH. Toward a theoretical model of text complexity for the early grades: learning from the past, anticipating the future. Read Res Q. 2012;47(3):235–258.

[73]. Malvern D, Richards B. Validation of a new measure of lexical diversity. In: Beers M, Bogaerde B, Bol G, editors. From Sound to sentence: studies on first language acquisition. Groningen: University of Groningen, Centre for Language and Cognition; 2000. p. 81–96.

[74]. Maas HD. Zusammenhang zwischen Wortschatzumfang und Lange eines Textes. Zeitschnftfur Literaturwtssenschaft und Linguistik. 1972;8:73–79.

[75]. George D, Mallery M. SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update. 10 ed. Boston: Pearson; 2010.

[76]. Bailey SC, Brega AG, Crutchfield TM, Elasy T, Herr H, Kaphingst K, et al. Update on health literacy and diabetes. Diabetes Educ. 2014;40(5):581–604. [PubMed: 24947871]

[77]. Castro CM, Wilson C, Wang F, Schillinger D. Babel babble: physicians' use of unclarified medical jargon with patients. Am J Health Behav. 2007;31(1):85–95.

[78]. Fang MC, Panguluri P, Machtinger EL, Schillinger D. Language, literacy, and characterization of stroke among patients taking warfarin for stroke prevention: Implications for health communication. Patient Educ Couns. 2009;75(3):403–410. [PubMed: 19171448]

[79]. Brach C, Keller D, Hernandez LM, Baur C, Parker R, Dreyer B, et al. Ten attributes of health literate health care organizations. NAM Perspectives. Discussion Paper, National Academy of Medicine, Washington, DC; 2012. doi:10.31478/201206a.

[80]. Bauer AM, Schillinger D, Parker MM, Katon W, Adler N, Adams AS, et al. Health literacy and antidepressant medication adherence among adults with diabetes: the diabetes study of Northern California (DISTANCE). J Gen Intern Med. 2013;28(9):1181–1187. [PubMed: 23512335]

[81]. Sarkar U, Karter AJ, Liu JY, Adler NE, Nguyen R, Lopez A, et al. The literacy divide: health literacy and the use of an internet-based patient portal in an integrated health system—results from the diabetes study of Northern California (DISTANCE). J Health Commun. 2010;15 (S2):183–196. [PubMed: 20845203]

[82]. Sarkar U, Karter AJ, Liu JY, Moffet HH, Adler NE, Schillinger D. Hypoglycemia is more common among type 2 diabetes patients with limited health literacy: The diabetes study of Northern California (DISTANCE). J Gen Intern Med. 2010;25(9):962–968. [PubMed: 20480249]

[83]. Schillinger D, Grumbach K, Piette J, Wang F, Osmond D, Daher C, et al. Association of health literacy with diabetes outcomes. JAMA. 2002;288(4):475–482. [PubMed: 12132978]

[84]. Britton BK, Gulgoz S. Using Kintsch's computational model to improve instructional text: effects of repairing inference calls on recall and cognitive structures. J Educ Psychol. 1991;83:329–345.

[85]. Kintsch W The role of knowledge in discourse comprehension: A construction-integration model. Psychol Rev. 1988;95(2):163–182. [PubMed: 3375398]

[86]. McNamara DS, Kintsch E, Songer NB, Kintsch W. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cogn Instr. 1996;14(1):1–43.

[87]. Duran ND, McCarthy PM, Graesser AC, McNamara DS. Using temporal cohesion to predict temporal coherence in narrative and expository texts. Behav Res Methods. 2007;39(2):212–223. [PubMed: 17695347]

[88]. Schillinger D The Intersections between social determinants of health, health literacy, and health disparities. Stud Health Technol Inf. 2020;41(269):22–41. doi:10.3233/SHTI200020.

[89]. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. Int J Med Inform. 2017;105:110–120. doi:10.1016/j.ijmedinf.2017.06.004. [PubMed: 28750904]

[90]. Cotos E, Huffman S, Link S. Understanding graduate writers' interaction with and impact of the research writing Tutor during Revision. J Writ Res. 2020;12(1):187–232.

[91]. Knight S, Shibani A, Abel S, Gibson A, Ryan P, Sutton N, et al. Acawriter: A learning analytics tool for formative feedback on academic writing. J Writ Res. 2020;12(1):141–186.

[92]. Mabachi NM, Cifuentes M, Barnard J, Brega AG, Albright K, Weiss BD, et al. Demonstration of the health literacy universal precautions Toolkit: Lessons for quality Improvement. J Ambulatory Care Manage. 2016;39(3):199–208. doi:10.1097/JAC.0000000000000102. [PubMed: 27232681]

[93]. Proulx J, Kandula S, Hill B, Zeng-Treitler Q. Creating consumer friendly health content: implementing and testing a readability diagnosis and enhancement tool. In 2013 46th Hawaii International Conference on System Sciences, IEEE, pp. 2445–53; 2013.

Table 1.

Descriptive statistics for variables included in LDA model.

Feature	*t*	*p*	*d*	Mean (SD): Difficult messages	Mean (SD): Easy messages	LDA co-efficient
Word range score, CW, SUBTLEXus	−7.897	< .001	−0.758	3.196 (0.172)	3.303 (0.123)	3.682
Word age of acquisition scores, AW, Kuperman	7.074	< .001	0.679	5.496 (0.39)	5.284 (0.26)	−0.291
Lexical diversity (D)	5.555	< .001	0.533	89.827 (23.659)	77.259 (23.543)	−0.002
FW overlap, paragraphs	4.854	< .001	0.466	8.225 (4.016)	6.379 (3.936)	−0.066
FW frequency, COCA academic	4.164	< .001	0.399	16831.399 (3494.824)		
15496.32 (3257.931)	−0.001					
Lexical diversity (Maas)	−4.022	< .001	−0.386	0.021 (0.003)	0.023 (0.004)	17.125
Bigram association strength (delta p), COCA academic	3.988	< .001	0.383	0.045 (0.012)	0.041 (0.01)	−7.369
Terms related to action	−3.474	< .001	−0.333	0.609 (0.133)	0.655 (0.142)	1.073
Syntactic similarity score	−3.429	< .001	−0.329	0.104 (0.031)	0.116 (0.039)	1.861
Word familiarity, AW, MRC	−3.408	< .001	−0.327	593.239 (4.894)	594.622 (3.825)	0.001
Fear and disgust words	3.275	< .001	0.314	0.118 (0.067)	0.097 (0.07)	−1.606
Verb overlap (binary), paragraphs	3.225	< .001	0.309	0.764 (0.257)	0.674 (0.311)	−0.890
Average number of direct object dependencies	3.051	< .010	0.293	1.253 (0.368)	1.148 (0.355)	−0.043
Words related to friends and family	−3.037	< .010	−0.291	0.191 (0.081)	0.217 (0.095)	2.099
Argument overlap (binary), paragraphs	2.853	< .010	0.274	0.579 (0.197)	0.524 (0.203)	−0.150
Words related to joy	−2.849	< .010	−0.273	0.617 (0.501)	0.788 (0.68)	0.225
Argument overlap, sentences	−2.708	< .010	−0.26	0.219 (0.078)	0.241 (0.09)	0.989
Trigram range, COCA fiction	2.697	< .010	0.259	0.028 (0.008)	0.026 (0.007)	−13.702
Trigram frequency, COCA newspaper	2.606	< .010	0.25	0.429 (0.118)	0.401 (0.111)	−0.985
Pronoun overlap, paragraphs	2.479	< .050	0.238	1.354 (0.66)	1.199 (0.648)	−0.175
Construction frequency, SD, COCA fiction	2.242	< .050	0.215	650503.988 (135651.74)	620837.079 (139117.53)	−0.001
Word concreteness, AW, MRC	−2.183	< .050	−0.209	2.642 (0.136)	2.669 (0.126)	0.910
Incidence of words related to objects	−2.086	< .050	−0.2	0.134 (0.053)	0.145 (0.061)	2.552
Argument overlap, paragraphs	−1.941	< .050	−0.186	0.383 (0.159)	0.413 (0.164)	1.177

AW = All words; CW = content words; FW = function words; SD = standard deviation; M = mean.

Table 2.

Confusion matrix for LDA results (test set).

	Difficult	Easy
Difficult	58	28
Easy	34	126

Table 3.

Confusion matrix for Flesch-Kincaid Grade Level results (test set).

	Difficult	Easy
Difficult	26	60
Easy	26	134