Applying Multiple Methods to Assess the Readability of a Large Corpus of Medical Documents

Danny T.Y. Wu[a], David A. Hanauer[b], Qiaozhu Mei[a], Patricia M. Clark[c], Lawrence C. An[c], Jianbo Lei[d], Joshua Proulx[e], Qing Zeng-Treitler[e], Kai Zheng[a, f]

[a] *School of Information, University of Michigan, Ann Arbor, MI, USA*
[b] *Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA*
[c] *Center for Health Communications Research, University of Michigan, Ann Arbor, MI, USA*
[d] *Center for Medical Informatics, Peking University, Beijing, China*
[e] *Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA*
[f] *Department of Health Management and Policy, University of Michigan, Ann Arbor, MI, USA*

Abstract

Medical documents provided to patients at the end of an episode of care, such as discharge summaries and referral letters, serve as an important vehicle to convey critical information to patients and families. Increasingly, healthcare institutions are also experimenting with granting patients direct electronic access to other types of clinical narratives that are not typically shared unless explicitly requested, such as progress notes. While these efforts have great potential to improve information transparency, their value can be severely diminished if patients are unable to read and thus unable to properly interpret the medical documents shared to them. In this study, we approached the problem by contrasting the 'readability' of two types of medical documents: referral letters vs. other genres of narrative clinician notes not explicitly intended for direct viewing by patients. To establish a baseline for comparison, we also computed readability scores of MedlinePlus articles—exemplars of fine patient education materials carefully crafted for lay audiences. We quantified document readability using four different measures. Differences in the results obtained through these measures are also discussed.

Keywords: Readability; Referral letters; Electronic health records; Natural language processing; MedlinePlus

Introduction

Encouraging active patient participation in their care processes holds great promise to improve patients' "understanding of their health, foster productive communication, stimulate shared decision making, and ultimately lead to better outcomes" [1]. In addition to oral communications, handout materials provided to patients and families at the end of each hospitalized stay or outpatient visit, such as discharge summaries and referral letters, increasingly serve as a critical vehicle for provider–patient communication because of the limited amount of 'face time' patients have to interact with their healthcare providers.

There has also been a growing interest among provider institutions in adopting patient-facing technologies such as patient portals to bring patients even closer to their care processes and their healthcare teams [2]. The OpenNotes Project supported by the Robert Wood Johnson Foundation in the U.S.,[1] for example, advocates for granting patients direct access through the internet to more types of clinical documents including assessments and plans, progress notes, and radiology reports [1, 3]. Such documents are not typically shared to patients unless explicitly requested. The basic tenets are that keeping patients informed not only about the ultimate medical judgments and treatment decisions, but also the processes that lead toward the judgments and decisions, will help them better understand their care as well as engender new opportunities for improvements such as better patient compliance, less misunderstandings, omissions, and errors, as well as revolutionary types of patient–provider relationships [1].

However, simply making the documents available to patients is not adequate to achieve the anticipated goals of improving information transparency. Obviously, shared medical documents will not be very useful if patients, particularly underserved populations with low literacy levels, are unable to read them and thus unable to properly interpret and put them into meaningful use. This problem will not only diminish the value of investments in opening up access to medical documents but may also result in unintended consequences such as confusion and unnecessary patient anxiety due to misinterpretation [1]. It is therefore important to understand what might constitute the obstacles to a patient's ability to comprehend medical text in order to inform strategies to address them. As an initial step, in this study, we focus on assessing 'readability' as a surrogate measure for 'comprehensibility,' which is in turn a surrogate measure of 'usefulness,' of medical documents that may be potentially shared to patients [4].

Assessing and subsequently improving the 'readability' of assorted types of reading materials has been an enduring interest among writers, publishers, educators, and researchers [e.g., 5, 6, 7, 8]. Numerous readability formulas were developed to assess the ease with which text can be read based on syntactic matrices such as average sentence length, average grade level of words, and average word length in syllables [e.g., 9, 10, 11]. In this study, we are not only interested in contrasting the readability of medical documents that are readily accessible to patients such as referral letters vs. those that are not, but also in learning how those general-purpose readability measures may perform on medical text.

Methods

Empirical Dataset

The corpus of medical documents analyzed in this study was retrieved from the electronic health records system used at the

[1] http://myopennotes.org/.

University of Michigan Health System (UMHS), a 925-bed quaternary academic medical center connected with 120 outpatient clinics and approximately 40 health centers. Each year, UMHS provides care in over 44,000 inpatient admissions and 1.8 million ambulatory visits.

The corpus is comprised of over 2 million free-text clinical documents belonging to decedent hematology/oncology patients seen at UMHS in the past three years. The use of the corpus in this study was reviewed and determined to be exempt as nonhuman subject research by our institutional review board. All possible genres of clinician notes are present including admission notes, progress notes, radiology reports, discharge summaries, and referral letters. Because UMHS does not currently endorse an 'Open Notes' policy, i.e., these narrative documents were not proactively shared to patients, most of them were meant to remain as 'internal' communications among medical professionals.

The only exceptions are referral letters, which are usually prepared by specialists describing their findings and recommendations for care to referring physicians—generalists in most cases. Because most referral letters are provided to (or copied to) patients who generally have less medical knowledge, it is reasonable to assume while these letters are being composed, there is anticipation that they may be read and could be understood outside of a healthcare encounter. We therefore hypothesized that the referral letters would have relatively better readability scores compared to other types of narrative clinician notes not explicitly composed for the purpose of direct viewing by patients.

A total of 76,012 referral letters and 2,118,463 other types of clinician notes are available in the corpus. We randomly selected 50,000 documents from each set to include in the analyses of this study. They are labeled as *RL* (Referral Letters) and *N-RL* (Non-Referral Letters), respectively, in the remaining parts of this paper.

Figure 1 – A Screenshot of MedlinePlus "Health Topics"

In addition, we analyzed "Health Topics" articles downloaded from MedlinePlus (a sample screenshot is shown in Figure 1) in order to establish a baseline for readability comparison. MedlinePlus is a consumer-oriented website operated by the U.S. the National Library of Medicine which is aimed at providing high-quality information to patients and families about diseases, conditions, and wellness issues.[2] Because all "Health Topics" articles are carefully crafted patient education materials specifically composed for lay audiences, we hypothesized that they would be rated with the best readability scores. As of June 30, 2012, when we prepared data for analysis, a total of 926 "Health Topics" were available at MedlinePlus. All of them were analyzed in this study, labeled as *MP* hereafter. Note that only the main body of the "Health Topics" articles was used (the highlighted portion in Figure 1). Other ancillary text, such as disclaimers, navigational aids, and links to additional readings and external websites, were removed prior to analysis. Non-English articles were not included.

Surface Metrics

We used several surface metrics to characterize the text features of the three types of documents of interest: average document length (average number of words per document), average sentence length, vocabulary size (number of distinct words across all documents), and vocabulary coverage. To compute the last surface measure, vocabulary coverage, we used the following two dictionaries:

- A medical dictionary based on all concepts and concept names, as well as tokenized distinct words appearing in the concept descriptions, from the 2010AB release of the Unified Medical Language System (UMLS)® Metathesaurus, which contains more than 150 controlled medical vocabularies including ICD, SNOMED CT®, LOINC, and MeSH;[3]
- An English dictionary combining multiple open-source dictionaries of the English language as well as basic concepts and terminologies commonly used in medicine (e.g., GNU Aspell [12] and OpenMedSpel [13]).

These two dictionaries, referred to as UMLS and the Basic Medical English Dictionary (BMED) in this paper, were initially developed in a prior study we conducted to examine the text features of clinician notes that were voice-dictated vs. typed through computer keyboards [14]. We believe these two dictionaries collectively should provide a reasonably comprehensive coverage of English words and known medical concepts and terminologies.

Readability Measures

We applied four different measures to assess the readability of the three types of documents of interest. Three of these measures have been widely used and empirically validated on non-medical content. They include:

- *Flesch-Kincaid Grade Level* (FKGL) which computes readability score based on a combination of word–sentence and syllable–sentence proportions [9];
- *Simple Measure of Gobbledygook* (SMOG) based on counting the frequency of polysyllables appearing in a piece of text [10];
- *Gunning-Fog Index* (GFI) based on sentence length and the percentage of complex words (i.e., words with three or more syllables) [11].

The formulas of FKGL, SMOG, and GFI are provided in Equations 1, 2, and 3, respectively. The outputs of these measures are all expressed as the numbers of "years of school

[2] http://nlm.nih.gov/medlineplus/.
[3] http://nlm.nih.gov/research/umls/.

education" that it requires in order to proficiently understand a piece of writing under evaluation.

$$0.39\times\left(\frac{\text{words}}{\text{sentences}}\right)+11.8\times\left(\frac{\text{syllables}}{\text{words}}\right)-15.59 \quad (1)$$

$$1.0430\times\sqrt{\frac{\text{polysyllables}}{\text{sentences}}\times 30}+3.1291 \quad (2)$$

$$0.4\times\left(\frac{\text{words}}{\text{sentences}}+100\times\left(\frac{\text{complex words}}{\text{words}}\right)\right) \quad (3)$$

In addition to these general-purpose measures, we also adopted a pilot readability-scoring algorithm proposed by Kim et al. that was specifically developed to evaluate medical documents produced in healthcare settings [15]. The algorithm, expressed in Equations 4 and 5, generates a relative measure by contrasting the text features of a target document to those of a reference corpus deemed 'easy to read' and a reference corpus deemed 'difficult to read.' The text features include length of text (average numbers of words per sentence, characters per word, and sentences per paragraph), syntactic features (parts of speech), and semantic features (empirically validated familiarity scores reflecting the difficulty level of a health term or concept to a lay audience) [11]. To optimize the algorithm's performance on medical content, the weight of a given text feature will be amplified if it significantly differs between the 'easy to read' and the 'difficult to read' reference copra (Equation 5).

$$D_i=\sum\left(\frac{\left|\bar{X}_{ij}^{test}-\bar{X}_{ij}^{easy}\right|}{STD_{ij}^{easy}}\times W_{ij}\right)\times\frac{1}{\sum W_{ij}} \quad (4)$$

$$W_{ij}=\frac{\left|\bar{X}_{ij}^{difficult}-\bar{X}_{ij}^{easy}\right|}{STD_{ij}^{easy}} \quad (5)$$

The output of the algorithm is a readability score in the range of -1 to 1; documents rated with higher scores are deemed easier to read. In the same study that proposed the algorithm, Kim et al. performed a preliminary validity evaluation using 10 discharge summary reports. The results showed that the new algorithm produces more reasonable scores when applied to clinical content compared to other general-purpose measures.

In this study, we extended Kim et al.'s algorithm validation by applying it to a large corpus of medical documents. In addition, we employed more sophisticated surface measures such as vocabulary coverage based on comprehensive medical and English dictionaries.

Analysis

Text appearing as bulleted lists was first converted into separate sentences before the analyses were performed. Stop words were kept intact in the documents as they may contribute to readability. To test if significant differences might exist in readability assessments across the three types of documents, we used the one-way analysis of variance (ANOVA) with *Bonferroni* correction [16]. We also plotted the distributions of the readability scores to examine the variance among the results generated by different measures. All statistical analyses were performed using Stata 11 (StataCorp LP, College Station, Texas, USA).

Results

The surface metrics are shown in Table 1. On average, referral letters are longest in terms of document length followed by non-referral clinician notes and then by MedlinePlus "Health Topics" articles. Among the three types of documents studied, clinician notes other than referral letters tend to have longer sentences. They also have a larger vocabulary size compared to that of the referral letters, which is in turn 27 times larger than that of the MedlinePlus articles.

As for vocabulary coverage, only 2.7% of the words used in the MedlinePlus articles cannot be found in the English and Basic Medical English Dictionary, whereas about three quarters of the words appearing in the medical documents, either in referral letters or other non-referral letter types, are not covered by the dictionary. Further, more than 66% of the words that healthcare professionals used in composing medical documents, referral letters included, are not found in UMLS and BMED combined. This finding suggests that these documents may contain a large number of nonstandard words, terminologies, or alternative forms of spelling such as nonstandard abbreviations and acronyms, which is consistently with what was found in an earlier study we conducted using the same dataset [14]. This fact will likely affect the readability of the documents by lay audiences, and possibly by healthcare professionals as well. Lastly, as anticipated, referral letters have higher vocabulary coverage compared to non-referral letter types of narrative clinician notes, suggesting that referral letters might be relatively easier to read.

Table 1 – Surface Metrics

Surface Metric	RL	N-RL	MP
Average document length	623.6	495.5	124.6
Average sentence length	10.9	13.7	12.2
Vocabulary size	184,448	205,283	6,772
Vocabulary coverd by UMLS	24.2%	22.7%	67.0%
Vocabulary coverd by BMED	21.8%	19.1%	97.3%
Vocabulary coverd by UMLS and BMED combined	33.7%	30.7%	99.5%

RL = Referral Letters; N-RL = Non-Referral Letters; MP = MedlinePlus Articles.

Table 2 presents the mean and standard deviation of the readability scores generated by the four different readability measures. To put it into perspective, we also selected a few examples from each of the document types (Table 3) to illustrate what might constitute documents 'easier to read' vs. 'harder to read' based on the algorithm by Kim et al.

Not surprisingly, the MedlinePlus articles were consistently rated as easiest to read. All three general-purpose readability measures, however, deemed the referral letters as the most difficult document type, whereas the algorithm proposed by Kim et al. found non-referral letter types of clinician notes tend to be harder to read. The ANOVA test shows that the differences in the mean readability scores across the three document types, *RL*, *N-RL*, and *MP*, respectively, are all statistically significant regardless of the readability measure used. Similar results were found by performing pairwise tests across the document groups.

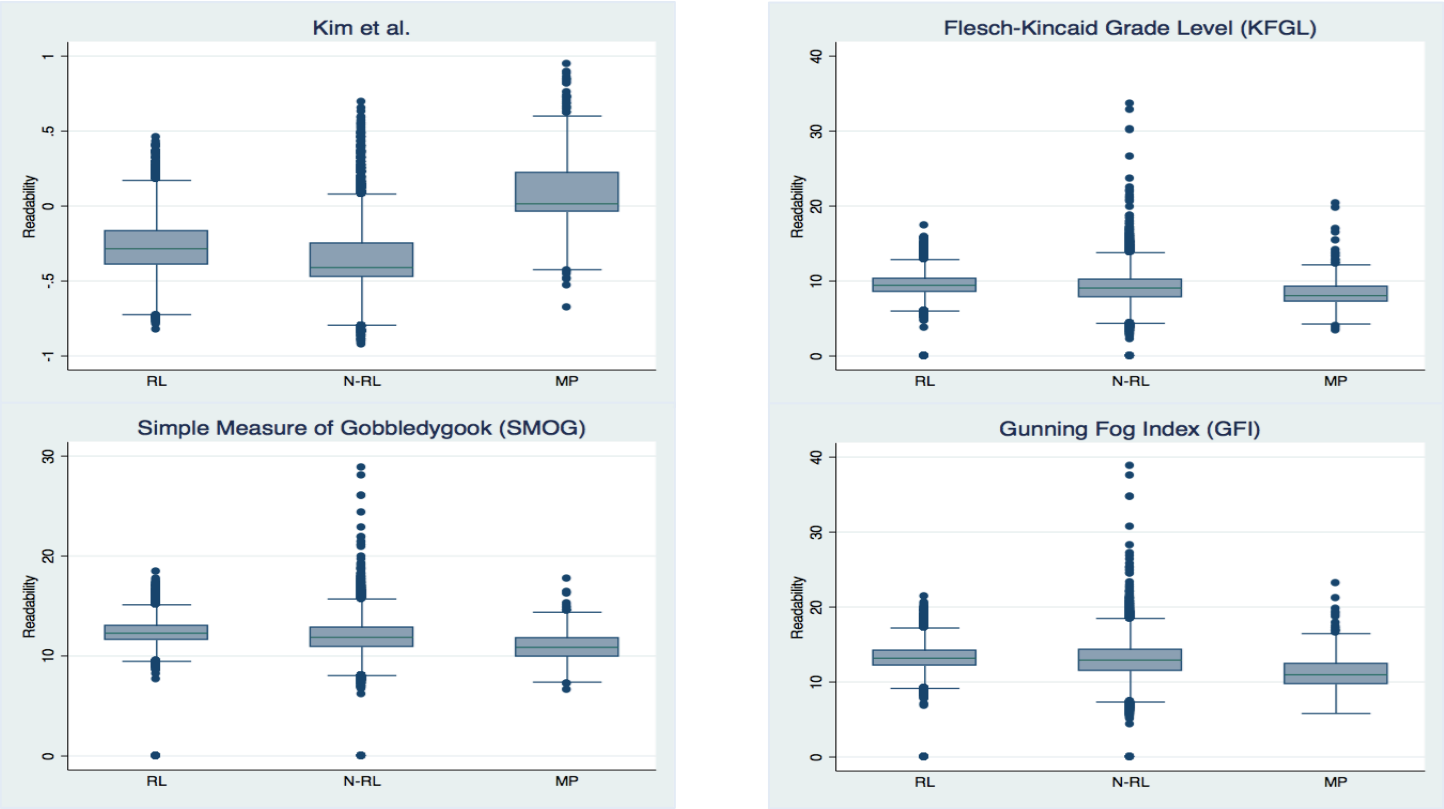

Figure 2 – Variance in Readability Scores

Table 2 – Readability Scores

Readability Measures	RL	N-RL	MP
Kim et al.	-0.27±0.15	-0.36±0.18👎	0.09±0.23👍
FKGL	9.44±1.30👎	9.09±1.89	8.29±1.77👍
SMOG	12.30±1.09👎	11.89±1.52	10.92±1.46👍
GFI	13.18±1.52👎	12.88±2.22	11.16±2.19👍

👍 Easiest to read; 👎 Most difficult to read.

Table 3 – Sample Documents

Type	Readability	Example
RL	Easier	"As you recall, he is a eighty-six-year-old gentleman with a history of a significant cataract in his right eye who presented for re-evaluation of his cataract."
RL	Harder	"She is seen today in followup for history of coagulase-negative staphylococcal line infection in the setting of the translumbar catheters for dialysis and TPN."
N-RL	Easier	"Will continue to follow and assess identified deficits and goals."
N-RL	Harder	"CHF with ischemic cardiac myopathy and ejection and an ejection fraction of 35%. PVOD with bilateral carotid stenosis."
MP	Easier	"The pattern of how you walk is called your gait. A variety of problems can cause an abnormal gait and lead to problems with walking."
MP	Harder	"You should be prepared to commit to three months of daily encouragement. Successful trips to the potty should be rewarded. Missteps shouldn't get as much attention."

Figure 2 illustrates the distributions of the readability scores obtained by using the four different readability measures. The non-referral letter types of clinician notes consistently exhibit the highest level of variance, which is intuitive as they contain multiple subtypes of clinical documents. Compared to the three general-purpose measures, the readability scores generated by the algorithm proposed by Kim et al. are more spread out, suggesting this new algorithm might have a better discriminative power in determining the readability of a medical document.

Discussion

Increasingly, healthcare institutions are exploring the feasibility of inviting patients to review and potentially co-construct clinical notes with their providers through patient-facing technologies such as patient portals. This concept has great potential to improve information transparency, foster new types of patient–provider relationships, and consequently lead to better patient satisfaction and health outcomes. Such anticipated benefits will not be attained, however, if patients are unable to read and interpret clinician notes and thus unable to use them effectively in patient–provider communication and informed decision-making.

Assessing the readability of medical documents is therefore a critical first step to assure that they are readable and are thus comprehensible when shared to patients and families. The results of this study show that the readability of the medical documents produced in healthcare environments has a great margin for improvements. For example, 66.3% of the words appearing in referral letters, a document type with a clear intention to be used in facilitating provider–provider as well as provider–patient communication, could not be found in UMLS and BMED combined. Such frequent use of nonstandard vocabularies can make the documents much more difficult to read, and may consequently result in misunderstandings and confusions when the documents are shared to patients and families, or when read by referring physicians.

Our results also show that traditional general-purpose readability measures based on simple metrics, such as sentence length and frequency of polysyllables, may not be able to produce meaningful scores to gauge how readable medical text is to lay audiences. For example, the three commonly used general-purpose measures all estimated that on average only 1 to 2 additional years of school education is needed to allow a person proficient in reading MedlinePlus articles to be able to proficiently read clinician notes. This is unlikely true given that nearly 80% of words appearing in the medical documents we examined are not even covered by common English and medical English dictionaries.

Interestingly, all general-purpose readability measures deemed referral letters more difficult to read than other types of clinician notes. This is anti-intuitive as the referral letters were composed specifically for the purposes of communicating information to other people (while non-referral letter types of clinician notes might not). Further, the surface metrics also suggest that non-referral letter types of medical documents tend to have longer sentences and contain more words that are not covered by known dictionaries. These observations suggest other text features not captured by general-purpose readability measures might be more important in determining the readability of medical content. In contrast, the algorithm proposed by Kim et al. seems to make more reasonable estimates about the readability of the three document types. We are unable to draw a definitive conclusion on this, however, within the scope of this study due to the lack of human validation.

This study has several limitations. First, all medical documents analyzed were retrieved from a single institution and a single patient care service (hematology/oncology). The results therefore may not be generalizable to medical content produced at other institutions and/or in other care settings. Second, we only used computational measures to estimate readability, which limits our ability to interpret the results obtained. In future work, it will be extremely valuable to engage real patients to validate these computational results solely based on text features. Third, for simplicity, we studied all non-referral letter types of documents as one group, and therefore were unable to identify nuances across different subtypes of medical documents such as those composed for distinct documentation, communication, or regulatory purposes; those created in different medical specialties and patient care contexts; and those composed by different types of clinicians.

Conclusion

In this study, we applied four different measures, in addition to several surface metrics, to assess the readability of a large corpus of medical documents composed by healthcare professionals. We used MedlinePlus "Heath Topics" articles as a baseline, and contrasted the readability of two distinct types of medical documents: referral letters easily accessible to patients vs. other genres of clinician notes that are not typically shared to patients unless explicitly requested. The results show that the readability of both types of medical documents, non-referral letters types of clinician notes in particular, is low compared to the patient education materials available at MedlinePlus. This fact may very likely undermine the value of sharing these documents to patients. Developing strategies to improve the readability of medical documents before making them available to patients is therefore warranted.

Acknowledgement

This project was supported in part by Grant # UL1RR024986 received from the U.S. National Center for Advancing Translational Sciences (NCATS), a component of the U.S. National Institutes of Health (NIH) and NIH Roadmap for Medical Research.

References

[1] Delbanco T, Walker J, Darer JD, et al. Open Notes: Doctors and patients signing on. Ann Intern Med. 2010;153:121–5.

[2] Tand PC, Lanksy D. The missing link: Bridging the patient-provider health information gap. Health Affairs 2005;24(5):1920–5.

[3] Delbanco T, Walker J, Bell SK, et al. Inviting patients to read their doctors' notes: A quasi-experimental study and a look ahead. Ann Intern Med. 2012;157(7):461–70.

[4] Garner M, Ning Z, Francis J. A framework for the evaluation of patient information leaflets. Health Expect 2012;15(3):283–94.

[5] Rosemblat G, Logan R, Tse T, et al. Text features and readability: Expert evaluation of consumer health text. Mednet 2006.

[6] Zeng-Treitler Q, Kim H, Goryachev S, et al. Text characteristics of clinical reports and their implications on the readability of personal health records. Stud Health Technol Inform. 2007;129(Pt 2):1117–21.

[7] Terblanche M, Burgess L. Examining the readability of patient-informed consent forms. Open Access Journal of Clinical Trials 2010;2:157–62.

[8] Adnan M, Warren J, Orr M. Assessing text characteristics of electronic discharge summaries and their implications for patient readability. HIKM 2010;108:77–84.

[9] Kincaid JP, Fishburne RP, Rogers RL, et al. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis, 1975.

[10] McLaughlin GH. SMOG grading: A new readability formula. J Reading 1969;12:639–46.

[11] Gunning R. *The Technique of Clear Writing*. New York, NY: McGraw-Hill International Book Co., 1952.

[12] Atkinson K. GNU Aspell. http://aspell.net/; accessed March 10, 2011.

[13] e-MedTools. OpenMedSpel - Opensource Medical Spelling. http://www.e-medtools.com/openmedspel.html; accessed March 10, 2011.

[14] Zheng K, Mei Q, Yang L, et al. Voice-dictated versus typed-in clinician notes: Linguistic properties and the potential implications on natural language processing. AMIA Annu Symp Proc. 2011;1630–8.

[15] Kim H, Goryachev S, Rosemblat G, et al. Beyond surface characteristics: A new health text-specific readability measurement. AMIA Annu Symp Proc. 2007;11:418–22.

[16] Dunn OJ. Multiple comparisons among means. J Am Statist Assoc.1961;56(293):52–64.

Address for correspondence

Kai Zheng, PhD
M3531 SPH II, 1415 Washington Heights
Ann Arbor, MI 48109
USA
kzheng@umich.edu