Measuring the readability of medical research journal abstracts

Samuel Severance
University of Colorado
School of Education
Institute of Cognitive Science
Boulder, CO USA
severans@colorado.edu

K. Bretonnel Cohen
University of Colorado
School of Medicine
Biomedical Text Mining Group
Aurora, CO USA
kevin.cohen@gmail.com

Abstract

This study examines whether the readability of medical research journal abstracts changed from 1960 to 2010. Abstracts from medical journals were downloaded from PubMed.org in ten-year batches (1960s, 1970s, etc.). Abstracts in each decade underwent processing via a custom Python script to determine their Coleman-Liau Index (CLI) readability score. Analysis using one-way ANOVA found statistically significant differences between the mean CLI readability scores of each decade ($F(4, 6689135) = 12936.91$, $p<0.0001$). Post-hoc analysis using Tukey's method also found all pairwise comparisons between decades' mean CLI readability scores to be statistically significant ($p<0.001$). Readability scores increased from decade to decade beginning with a mean CLI score of 16.0813 in the 1960s and ending with a mean CLI score of 16.8617 in the 2000s. These results indicate a 0.7804 grade level increase in the difficulty of reading medical research journal abstracts over time and raises questions about the accessibility of medical research for broader audiences.

1 Introduction

A persistent issue in academic research centers on whether the knowledge published by researchers reaches and is understood by those it could benefit. The medical field takes up this issue in its efforts to translate research into practice, or the idea of "translational research" (Woolf, 2008). Ideally, practitioners can access and thoroughly comprehend research to better ensure new treatments and knowledge reaches patients and that patient care revolves around evidence-based practices (Pravikoff, Tanner, & Pierce, 2005; Woolf, 2008). Beyond seeking to leverage new research among medical practitioners, translational research also focuses on supporting patients in becoming more active and involved in their healthcare (Woolf, 2008). With the advent of the information age, patients and patients' family members have substantial opportunities to research their own medical conditions and their treatment options. Navigating and understanding medical research requires that it proves accessible in terms of its readability.

This study is a diachronic analysis of the readability of medical research. Specifically, this study seeks to answer whether the readability of medical research journal abstracts has changed from the 1960s to the 2000s. Results from this study may have implications for how researchers could communicate their findings to patients and how to address discrepancies between the reading level of medical journals and lay audiences' reading abilities.

2 Relevant Literature

2.1 Readability of health materials in relation to patients

Research on the readability of health materials in relation to patients has a strong presence in the literature. Health literacy researchers have found

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 127–133,
Beijing, China, July 30, 2015.

that the vast majority of textual information patients typically encounter—from informed consents to patient education materials—surpass the reading ability of patients (Rudd, Moeykens, & Colton, 1999). Such discrepancies may have profound negative influences on patient health outcomes (Paasche-Orlow & Wolf, 2010). Indeed, Baker et al. (1998) found an independent association between low health literacy and increased hospital admission rates where patients with low literacy became hospitalized twice as often as more literate patients. Additionally, patients with high functional health literacy become more involved in their care, including exploring options beyond those presented by a doctor, whereas patients with low functional health literacy tend to limit decisions regarding their care to only those presented to them by doctors (Smith et al. 2009). With implications for personal and community health, a study by Navarra et al. (2014) found that HIV-infected youth with below-grade-level reading skills did not completely adhere to their antiretroviral therapy.

Despite growing evidence of the role of health literacy in patient outcomes, the readability of medical information for patients has not improved over time, even for items intended for patients. The lack of readability of informed consents, in particular, has garnered attention in the literature (Mead & Howser, 1992; Rudd et al., 1999). An examination of the readability of informed consents from 1975 to 1982 at the Veterans Administration Medical Center found them to have a college reading level and that their reading difficulty may have actually increased over the time period examined (Baker & Taub, 1983). Fifteen years later, a study of surgical consents from across the US also showed similarly difficult reading levels with a given consent requiring an average reading level of 12.6 (Hopper et al., 1998). Beyond informed consents, other materials directly aimed at laymen also show readability issues. In an analysis of emergency first-aid instructions, Temnikova (2012a, 2012b) found ten separate categories of readability/complexity problems. Alamoudi and Hong (2015) found the readability of websites related to microtia and aural atresia lacking in terms of facilitating comprehension.

2.2 Identifying and addressing readability issues

A significant body of work focuses on addressing readability issues in health contexts. It makes the significance of the corpus-based study reported here clear: it shows that we can address readability problems, but first *we must know what the readability issues are*.

Elhadad (2006), for instance, shows which terms in a medical journal article a lay reader would likely not understood and presents an application that finds these terms and mines an appropriate definition from the Web. Achieving usable results with a small corpus, Elhadad and Sutaria (2007) presented a parallel-corpus-driven method for finding technical/lay equivalents of medical terms using measures of association. Leroy et al. (2010) pointed out that perceived and the actual difficulty of text influenced the willingness and ability to learn from health information. The researchers manipulated characteristics of health texts and measured perceived and actual difficulty, and found they could improve the perceived difficulty of text. Their technique also uncovered some problems with standard readability formulas. Using lexical and grammatical analysis of a medical corpus to develop a new metric to estimate text difficulty called "term familiarity," Leroy et al. (2012) performed an experiment where individuals showed slightly improved understanding for simplified documents. An evaluation of a writing assistance tool that assists with automated simplification related to term familiarity found that simplified text had strong beneficial effects on both perceived and actual difficulty, with better understanding and more learning after reading simplified text than after reading un-simplified text (Leroy, Kauchak & Mouradi, 2013). In another study, Leroy et al. (2013) examined the effects of lexical simplification and coherence enhancement on readability and showed that they interact in complex ways with both perceived and actual difficulty. Investigating linguistic features, specifically discourse features that correlate with the readability of texts for adults with intellectual disabilities, Fung et al. (2009) presented a tool for rating the readability of texts for these readers. Huenefaurth et al. (2009) compared different methods for evaluating text readability software for adults with intellectual disabilities, finding that multiple-choice questions with illustrations proved more useful than yes/no questions or Likert scales for evaluating simplification programs.

2.3 Work presented in context of relevant literature

Specific research utilizing a diachronic, corpus-based approach to examining the readability of medical journals did not turn up in a review of the literature. However, previous studies taking a diachronic approach to the readability of corpus data do have precedence. Indeed, the inspiration for this study comes from work by Štajner (2011). Štajner performed a diachronic analysis of the Brown "family" of corpora to examine changes in the readability of the English language over time. Similar to this study, Štajner utilized the Coleman-Liau Index as a measure of the readability of the Brown "family" of corpora.

3 Methodology

This study occurred in three main phases in order to answer the research question: How has the readability of medical journal abstracts changed between the 1960s and 2000s?

3.1 Obtaining a medical research corpus

The first phase of this study involved compiling a machine-readable corpus of medical research journal abstracts. PubMed.org contains a large volume of medical research journal abstracts and these provided the basis of a corpus. Abstracts were downloaded in groups by decade (see Table 1).

Table 1. Number of abstracts by decade.

Decade range	Number of abstracts
1960-1969	5324
1970-1979	313053
1980-1989	1049637
1990-1999	2017482
2000-2009	3327954

This study focused solely on journal abstracts dealing with research on human subjects with the assumption that a human-centered research corpus has more meaningful parallels to the potential interests of most patients.

3.2 Measure the readability of abstracts

The Coleman-Liau Index measure of readability (Coleman & Liau, 1975) formula is as follows:

$$CLI = 5.89\frac{c}{w} - 29.5\frac{s}{w} - 15.8$$

In this formula, c is equal to the total number of characters in a given text, w is equal to the total number of words in a given text, and s is equal to the total number of sentences in a given text. The CLI outcome measure is given as a grade-level readability score. For example, a grade of 10.5 would correspond to a text at a reading level of halfway through 10th grade.

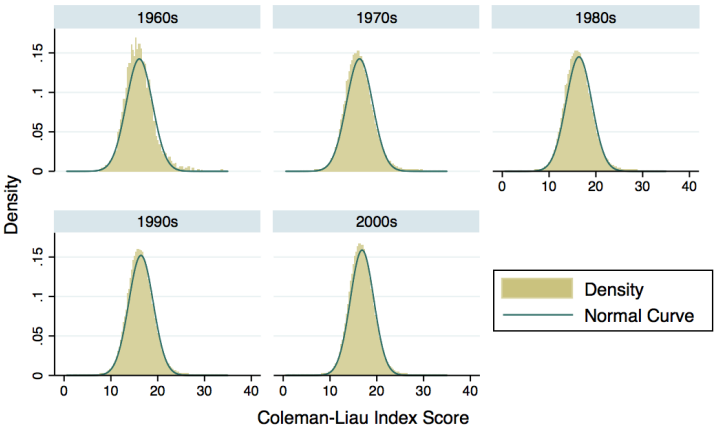

Figure 1: Distribution of CLI Scores by decade.

3.3 Statistical analyses

The next phase of the study involved creating a database and running analyses to determine the mean CLI scores for abstracts in each decade and whether the differences between these mean scores were statistically significant. A statistically significant difference in the mean CLI scores for each decade would indicate changes in the readability of medical journal abstracts over time.

In order to avoid type 1 errors, the analysis did not engage in a series of t-tests to compare the mean CLI scores for each decade. Rather a one-way ANOVA was deemed more appropriate after checking that the data met certain assumptions. Specifically, ANOVA requires that the data have an approximately normal distribution. Evidence for normality includes histograms of each decade's CLI scores with each distribution closely following a normal curve (see Figure 1 above). A Shapiro-Wilk test for normality could not be done because it has an upper limit of 2,000 to 5000 observations (Razali & Wah, 2011), and the data sets in this paper surpass that (see Table 1 above). However,

examination of the quantile-quantile plots (Figure 2 below) is consistent with the data being approximately normally distributed in each decade with only a small fraction of overall observations displaying deviations in the tails of some plots.

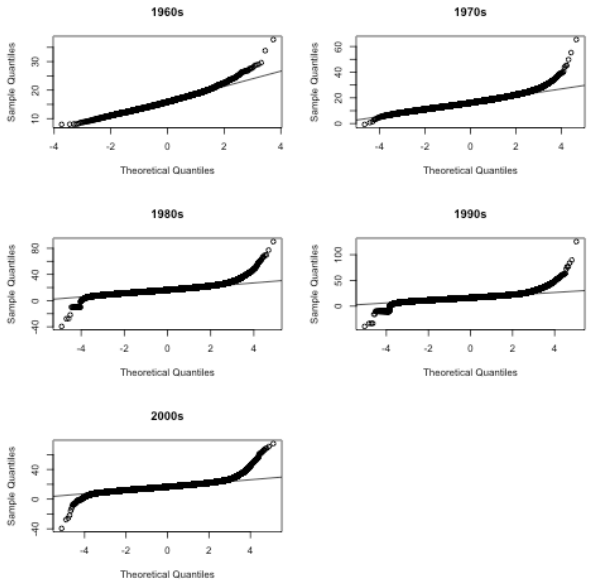

Figure 2: Quantile-quantile plots, by decade.

ANOVA also requires very similar variances for each group. Levene's test for homogeneity of variance, run on subsets created through random sampling of each decade, gave statistically significant values, which means the null hypothesis that the variances were the same could not be rejected. Another assumption for running ANOVA is that the data are independent. Although it is possible that some research articles may have been republished or had text cited in different decades, such instances likely were rare and not significant given the size of the corpus.

With the above assumptions addressed, the one-way ANOVA was carried out. Examination of the output indicated that a statistically significant difference did exist between the mean CLI scores for each decade. A one-way ANOVA, however, is an omnibus test and does not indicate between which groups the statistically significant difference exists, just that a statistically significant difference exists somewhere in the data. To determine between which decades there exists a statistically significant difference in mean CLI scores, a post hoc analysis using Tukey's method was carried out.

4 Results

The mean CLI scores for each decade were calculated (see Table 2.) The 1960s had a mean CLI score of 16.0813 with a 95% confidence interval (CI) of 16.00567 to 16.1569. The 1970s had a mean CLI score of 16.3123 with a 95% CI of 16.3024 to 16.32212. The 1980s had a mean CLI score of 16.3867 with a 95% CI of 16.38137 to 16.39194. The 1990s had a mean CLI score of 16.4302 with a 95% CI of 16.42657 to 16.43385. The 2000s had a mean CLI score of 16.8617 with a 95% CI of 16.85901 to 16.86446. Note that none of the 95% CIs overlap between decades.

Table 2. Mean CLI scores by decade.

Decade	Mean CLI Score	Number of Abstracts
1960s	16.0813	5324
1970s	16.3123	313053
1980s	16.3867	1049637
1990s	16.4302	2017482
2000s	16.8617	3327954

A one-way ANOVA indicated a statistically significant difference between the mean CLI scores for each decade ($F(4, 6689135) = 12936.91$, $p<0.0001$; see table 3). In determining which pairs of mean CLI Scores for each decade had a statistically significant difference, a pairwise comparison of means post hoc analysis using Tukey's method indicated that all possible combinations of CLI Scores for each decade had statistically significant differences ($p<0.001$).

5 Analysis

Having confirmed the statistical significance of the differences between all pairings of the mean CLI scores for each decade, we can consider the mean CLI scores for each decade statistically distinct

Table 3. One-way ANOVA results comparing mean CLI scores by decade.

Source	SS	df	MS	F-statistic	p-value
Between groups	353331.527	4	88332.8818	12936.91	<0.0001
Within groups	45673229.4	6689135	6.82797244		
Total	46026560.9	6689139	6.88079003		

from one another. Given this, we can make higher-level observations based on what patterns the individual means reveal as part of a group. More importantly, we can make assertions that allow us to answer our research question: How has the readability of medical journal abstracts changed between the 1960s and 2000s?

According to the results of this study, the average difficulty in readability of medical research journal abstracts increased over time. Specifically, readability scores increased from decade to decade beginning with a mean CLI score of 16.0813 in the 1960s and ending with a mean CLI score of 16.8617 in the 2000s. The mean CLI score, therefore, increased 0.7804 grade level units within the timespan examined. We should also note the high mean CLI scores for each decade. All scores fell within the level of readability expected for a grade level of 16 or a senior in college.

6 Future work

The work reported here discusses only one readability metric. Fleshing out the data with additional readability metrics would prove useful. Experimental assessment of comprehension by lay readers would be a useful addition to the metrics; for example, by asking them to read abstracts and answer questions. Specific subdomains of the biomedical literature may have their own readability issues, such as formulae and gene names, and identifying these might have implications for approaches to addressing specific readability issues.

7 Conclusion

This study sought to determine whether the readability of medical research journal abstracts changed between the 1960s and 2000s. The results here indicate an increase in difficulty of 0.7804 grade levels during this time period. Medical journal abstracts, we can conclude, have become more and more difficult to read.

For patients attempting to learn more about medical conditions or their treatment options through the reading primary literature, this task has become more difficult to achieve. Importantly, however, the high overall mean CLI scores for each decade indicate that this task likely has always proven difficult for patients. Medical journal abstracts have had readability scores equivalent to grade levels of 16 since the 1960s, well above the average American who reads between a 7th and 8th grade level (NCES, 2003) and certainly above the 9th-grade level considered "difficult" (USDHHS, 2000). This consistent difficulty mirrors other research showing a lack of progress in the readability of medical-related text (Rudd et al., 1999).

From this study's results and the US Department of Health and Human Services recommendations for the reading levels of medical information text, the readability gap between published medical research and the average American patient's reading ability appears equal to 7 grade levels. Bridging this chasm in accessibility will likely require interventions for both the researcher and patient. Shoring up the "health literacy" of Americans would involve a concerted effort to increase the average reading ability of patients. Purposefully addressing health literacy in K-12 education settings and Adult Basic Education settings may prove beneficial (Nielsen-Bohlman et al., 2004; Rudd et al., 1999). Such efforts, however, will likely not bridge the 7 grade level gap entirely. Instead, the medical research community should consider taking steps—for example, developing reading guides or parallel publications aimed at lay readers—to increase the readability of their research given patients' information needs and to support patient self-advocacy.

Despite a desire by patients to access and comprehend research that would increase their involvement in their own care, members of the medical research and publishing community continue to place a premium on complex writing skills putting such research out of the reach of most patients. Lakoff (1992) makes a strong case for academics in general being rewarded for difficult writing, and perhaps even being published for incomprehensible writing. With typical reading levels of almost 17, most scientific writing is now beyond the reading level of not only the average patient but also most health professionals who typically have a bachelor's degree equivalent to a grade level of 16.

Acknowledgments

We thank the participants in LING 5200, Computational Corpus Linguistics, in Fall 2014 for their input into this project. Noemie Elhadad and Gondy Leroy provided helpful comments on a late draft of the work.

References

Alamoudi, U., and Hong, P. (2015) Readability and quality assessment of websites related to microtia and aural atresia. *International Journal of Pediatric Otorhinolaryngology* 79(2):151-156.

Baker, M. T., & Taub, H. A. (1983). Readability of informed consent forms for research in a Veterans Administration medical center. *Jama*, *250*(19), 2646-2648.

Baker, D. W., Parker, R. M., Williams, M. V., & Clark, W. S. (1998). Health literacy and the risk of hospital admission. *Journal of general internal medicine*,*13*(12), 791-798.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283.

Elhadad, N. (2006) Comprehending technical texts: predicting and defining unfamiliar terms. *American Medical Informatics Association Symposium Proceedings,* pp. 239-243.

Elhadad, N., and Sutaria, K. Mining a lexicon of technical terms and lay equivalents. *BioNLP 2007,* pp. 49-56.

Feng, L., Elhadad, N., and Huenefauth, M. (2009) Cognitively motivated features for readability assessment. *EACL 2009,* pp. 229-237.

Hartley, J. (2004). Current findings from research on structured abstracts.*Journal of the Medical Library Association*, *92*(3), 368.

Hopper, K. D., TenHave, T. R., Tully, D. A., & Hall, T. E. (1998). The readability of currently used surgical/procedure consent forms in the United States.*Surgery*, *123*(5), 496-503.

Huenefauth, M., Feng, L, and Elhadad, N. (2009) Comparing evaluation techniques for text readability software for adults with intellectual disabilities. *ACM SIGACCESS conference on computers and accessibility,* pp. 3-10.

Joint Commission. (2007) What did the doctor say? Improving health literacy to protect patient safety. *Health Care at the Crossroads* series. http://www.jointcommission.org/nr/rdonlyres/d5248b2e-e7e6-4121-887499c7b4888301/0/improving_health_literacy.pdf

Lakoff, R.T. (1992) *Talking power: the politics of language.* Basic Books.

Leroy, G., Endicott, J. E., Kauchak, D., Mouradi, O., & Just, M. (2013). User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of medical Internet research*, *15*(7).

Leroy, G., Endicott, J. E., Mouradi, O., Kauchak, D., & Just, M. L. (2012). Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In *AMIA Annual Symposium Proceedings* (Vol. 2012, p. 522). American Medical Informatics Association.

Leroy, G., Helmreich, S., & Cowie, J. R. (2010). The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, *79*(6), 438-449.

Leroy, G., Kauchak, D., & Mouradi, O. (2013). A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International journal of medical informatics*, *82*(8), 717-730.

Meade, C. D., & Howser, D. M. (1991, December). Consent forms: how to determine and improve their readability. In *Oncology nursing forum* (Vol. 19, No. 10, pp. 1523-1528).

Navarra, A.M., Neu, N., Toussi, S., Nelson, J., & Larson, E.L. (2014) Health literacy and adherence to antiretroviral therapy among HIV-infected youth. *J. Assoc. Nurses AIDS Care.* 25(3):203-213.

National Center for Education Statistics (2003). National Assessment of Adult Literacy (NAAL). http://nces.ed.gov/naal.

Pravikoff, D. S., Tanner, A. B., & Pierce, S. T. (2005). Readiness of US nurses for evidence-based practice: many don't understand or value research and have had little or no training to help them find evidence on which to base their practice. *AJN The American Journal of Nursing*, *105*(9), 40-51.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Roter, D. L., Rudd, R. E., Keogh, J., & Robinson, B. (1986). Worker produced health education material for the construction trades. *International Quarterly of Community Health Education*, *7*(2), 109-121.

Rudd, R. E., Moeykens, B. A., & Colton, T. C. (1999). Health and literacy: a review of medical and public health literature. *Office of Educational Research and Improvement*.

Smith, S. K., Dixon, A., Trevena, L., Nutbeam, D., & McCaffery, K. J. (2009). Exploring patient involvement in healthcare decision making across different education and functional health literacy groups. *Social science & medicine*,*69*(12), 1805-1812.

Štajner, S. (2011, September). Towards a Better Exploitation of the Brown 'Family' Corpora in Diachronic Studies of British and American English Language Varieties. In *RANLP Student Research Workshop* (pp. 17-24).

Temnikova, I. (2012a) Improving emergency instructions. *Communicator* (pp. 48-53).

Temnikova, I. (2012b) *Text complexity and text simplification in the crisis management domain.* University of Wolverhampton doctoral thesis.

United States Department of Health and Human Services. (2000) Saying it clearly. http://www.talkingquality.gov/docs/section3/3_4.htm.

Weiss, B. D., Coyne, C., Michielutte, R., Davis, T. C., Meade, C. D., Doak, L. G., ... & Furnas, S. (1998). Communicating with patients who have limited literacy skills-Report of the National Work Group on Literacy and Health.*Journal of Family Practice*, *46*(2), 168-176.

Woolf, S. H. (2008). The meaning of translational research and why it matters. *JAMA*, *299*(2), 211-213.