Initial Readability Assessment of Clinical Trial Eligibility Criteria

Tian Kang, BS, Noémie Elhadad, PhD, Chunhua Weng, PhD
Department of Biomedical Informatics, Columbia University, New York, NY 10032

Abstract

Various search engines are available to clinical trial seekers. However, it remains unknown how comprehensible clinical trial eligibility criteria used for recruitment are to a lay audience. This study initially investigated this problem. Readability of eligibility criteria was assessed according to (i) shallow and lexical characteristics through the use of an established, generic readability metric; (ii) syntactic characteristics through natural language processing techniques; and (iii) health terminological characteristics through an automated comparison to technical and lay health texts. We further stratified clinical trials according to various study characteristics (e.g., source country or study type) to understand potential factors influencing readability. Mainly caused by frequent use of technical jargons, a college reading level was found to be necessary to understand eligibility criteria text, a level much higher than the average literacy level of the general American population. The use of technical jargons should be minimized to simplify eligibility criteria text.

Introduction

As the gold standard for generating the most rigorous medical evidence regarding the effectiveness and efficacy of new medical therapies, clinical trials are fundamental for advancing medical research and public health. However, recruitment has remained the biggest persistent obstacle[1,2]. With the pervasive Internet access, the rise of web-based patient-trial matching and patient-screening methods presents great opportunities to overcome recruitment barriers. There is an increasing need to engage the general public in participating in clinical trials through “active matching,” where patients or health consumers are empowered to review clinical trial eligibility criteria and select clinical trials to participate in. According to the search logs of popular online clinical trial search engines, in addition to researchers, patients and health consumers have been increasingly searching and browsing clinical trials.

The success of active patient-trial matching hinges on the readability of clinical trial summaries, particularly the eligibility criteria language that clinical trial seekers use for determining their match to a trial. Because of the different professional training levels between patients and the researchers who wrote the clinical trial summaries, we hypothesize that required reading level of the trial texts are higher than the health literacy of out-of-domain readers. In particular, eligibility criteria, which determine the potential match for a patient, contain complex, specific content primarily written for researchers. In this paper, we examine the readability of clinical trial eligibility criteria.

Text readability assessment is an established field of research. As defined by Dale and Chall et al.[3], readability is “the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at optimal speed, and find it interesting”. Most readability measures have been established as assessment of shallow features of text, such as the number of words per sentence and the number of syllables per word. The SMOG readability assessment formula, which measures the percentage of polysyllables in texts, is a popular method for semantic assessment[4]. These widely adopted assessments are easy to compute on any text but lack the ability to represent the complexity of a text accurately[5], as this requires the use of both semantic and syntactic features. Measures of readability, which do assess both semantic features and syntactic features, have also been adopted to achieve more accurate assessment results. The New Dale-CHALL formula uses a combination of average sentence length and the proportion of the words in the text deemed “difficult” based on a 3000-word list containing what have been identified as the most commonly used English words[6]. As an added benefit, Natural Language Processing (NLP) enables automated readability assessment. For instance, Yngve measures the depth of nodes in the parse tree using NLP methods as the indicator of structure complexity[7]. These well-accepted methods for readability assessment have been validated in general reading materials and have been in use for years. Feng et al. performed a readability measurement for people with intellectual disabilities (ID) using a linear regression model to fit a set of cognitively motivated features[8]. Methods like Support Vector Machine[9] were also popular in this kind of research. However, supervised learning approaches constantly suffer from one big limitation, the need for large-scale labeled corpora.

Compared to what is known about general English’s readability, our understanding of biomedical text’s readability, especially that of clinical trial summaries, remains limited. Many studies of medical texts focus on online health-related materials and evaluate if these texts’ readability meets the recommendations of NIH, AMA and USDHHS standards (less than 6 grade) using general readability formulas[10-15]. A review of all readability and comprehension instruments used for print and web-based cancer materials[16] concluded that although readability formulas are predictive measures of text comprehensibility, they have been criticized for "only relying on word and sentence factors and for ignoring possible effects for reader motivation, design, and graphics on readability and comprehension". Kim et al.[17] developed a new health text-specific readability measurement, which is based on the distance of text features from the health related material to the predefined "easy" and "difficult" documents. It takes both syntactic features and semantic features into consideration and generates a sum of those distances as the indicator of readability. This method has been applied to several studies since its inception.

Since clinical trials constitute a critical part of the biomedical domain, the readability of clinical trials has started to receive attention. For example, Wu et al. conducted a readability assessment for description text in clinical trials from ClinicalTrials.gov using both general purpose and medical specific readability assessment measures[18]. They concluded that the descriptions were the most difficult to read when compared to other corpora e.g. electronic health records, MedlinePlus Health Topics articles. Ross et al.[19] conducted the first formal analysis of computer-interpretations of eligibility criteria complexity in clinical trials in order to facilitate phenotype studies. They randomly selected 1000 clinical trials from ClinicalTrials.gov and manually analyzed their complexity and semantic patterns. The results concluded that 93% of these free-text criteria were mainly comprehensible for professionals and 85% had significant semantic complexity for computational representations. Still, little is known about the readability of eligibility criteria and its impact on clinical trial recruitment and evidence adoption. Aiming to fill this knowledge gap and guide future improvement of consumer-facing clinical trial search engines, this paper contributes the first comprehensive readability assessment of clinical trial eligibility criteria text. In this study, we measured the human reading level necessary to understand the eligibility criteria in ClinicalTrials.gov. We measured both general readability and health-domain specific readability of eligibility criteria and discussed the implications of our findings for clinical trial designers and clinical trial search engine developers.

Methods

Datasets

Three corpora were used for comparison in this project. We retrieved clinical trial summaries from the world's largest clinical trial registry, ClinicalTrial.gov,[20] as our target corpus. To generate a better understanding of the results, we also performed assessments on two health-related corpora, i.e., Reuters News and PubMed[21]. Reuters News is intended for lay people. PubMed is intended for technical readers. Samples taken from the three corpora are shown in **Table 1**. We select a set of disease topics to include disease-specific descriptions from the three corpora. We downloaded the flat file of the database from the ClinicalTrials.gov website using their API and extracted trial summary text for the same list of health topics in two health-related corpora (e.g., Alzheimer, Type 2 Diabetes Mellitus, and Cardiovascular Diseases). We matched these disease topics to the condition field of clinical trial summaries supplied by ClinicalTrials.gov. We extracted 120,977 clinical trials on the aforementioned health topics. Then we collected 3,144 Reuters stories along with their corresponding PubMed articles.

Table 1. Sample text on the topic of "lung cancer" from the three corpora

PubMed articles	*…Marijuana smoke contains many of the same constituents as tobacco smoke, but whether it has similar adverse effects on pulmonary function is unclear…The Coronary Artery Risk Development in Young Adults (CARDIA) study, a longitudinal study collecting repeated measurements of pulmonary function and smoking over 20 years…*
Reuter News	*... A few hits on the bong now and then don't seem to have any detrimental effects on lung health, suggests a new study. Researchers found that multiple measures of lung function actually improved slightly as young people reported using more marijuana …*
Eligibility Criteria	*Ages Eligible for Study: 18 Years and older* *Genders Eligible for Study: Both* *Inclusion Criteria:* • *Pathologically confirmed, by biopsy or cytology, non-small cell lung carcinoma diagnosed within 3 months prior to study enrollment.* • *T1, N0, M0 or T2, N0, M0.* • *…* *Exclusion Criteria:* • *Evidence of distant metastasis (M1) and/or nodal involvement (N1, N2, N3).*

Feature evaluation

Figure 1 shows our methodology framework. Trying to inspect the readability comprehensively, we selected several representative features for each aspect of the texts (i.e., text unit length, syntactic complexity and lexical complexity) for a thorough readability assessment.

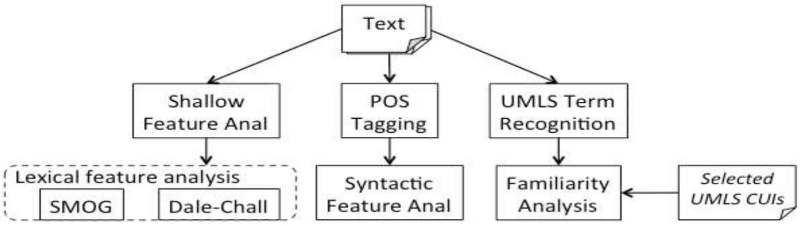


Figure 1. Overview of the readability assessment framework

(A) Shallow features:

1. aWPS: average number of words per sentence.
2. aSPW: average number of syllables per words.
3. %3+S: percentage of words in text with 3+ syllables ("polysyllables")

(B) Syntactic features:

1. aNP: average number of noun phrases per sentences.
2. aVP: average number of verb phrases per sentences.
3. aSBr: average number of "SBAR"s per sentences. ("SBAR" is defined as a "clause introduced by a subordinating conjunction." It is an indicator of sentence complexity[9].)

All of the syntactic features were generated by doing Part-Of-Speech using the Stanford parser[22]. The results for syntactic features provide us with a general view of the syntactic complexity of the documents.

(C) Lexical features:

Two separate readability scores were computed for the eligibility criteria of each clinical trial. The Simple Measure of Gobbledygook (SMOG) uses number of polysyllabic words (i.e., words containing 3 or more syllables) and number of sentences to calculate the reading level of the texts. It gives a general view of word complexity. The formula for calculating SMOG is:

$$grade = 1.0430\sqrt{number\ of\ polysyllables \times \frac{30}{number\ of\ sentences}} + 3.129$$

The new Dale-Chall formula was also used. This measurement is considered more accurate than the original formula as it is based on the use of "easy" and "difficult" words rather than syllables or simply letter counts. It labels words found on an established word list containing 3000 statistically common-used English words as "easy" words. If the word from the text is not included in the word list, then this word is labeled as "difficult" and a corresponding penalty is added to the final score. The final score can then be mapped to the grade level and the required grade reading level for readers can be determined[6]. The mapping from SMOG grade to U.S. standard educational level and the mapping from final score to education levels are shown **Table 2**.

The calculation method is as follows:

Raw score = 0.1579 × percentage of "difficult words" + average sentence length

If percentage of "difficult words" >5%

Adjusted score = Raw Score + 3.6365,

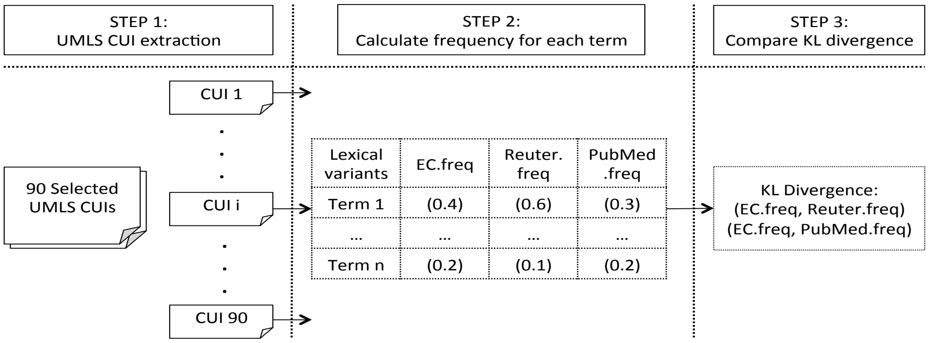
Otherwise Adjusted score = Raw score

SMOG GRADE	EDUCATIONAL LEVEL
6.0 and below	Low-literate
7.0 to 8.0	Junior High School
9.0 to 11.0	Some High School
12	High school graduate
13.0 to 15.0	Some College
16	University degree
17.0 to 18.0	Post-graduate studies
19.0 and above	Post-graduate degree
Dale-Chall GRADE	**EDUCATIONAL LEVEL**
4.9 and Below	Grade 4 and Below
5.0 to 5.9	Grades 5 - 6
6.0 to 6.9	Grades 7 - 8
7.0 to 7.9	Grades 9 - 10
8.0 to 8.9	Grades 11 - 12
9.0 to 9.9	Grades 13 - 15 (College)
10 and Above	Grades 16 and Above (College Graduate)

Table 2. Mapping of Readability Levels across Standards

2. Term familiarity prediction

We hypothesized that the technical terminology used in eligibility criteria might be one of the major obstacles preventing lay readers from fully understanding the content. Thus, a health-domain specific method published by Elhadad [21] was applied to evaluate the familiarity of words used in EC texts based on word frequency. The words with high frequencies in "easy" texts are usually found to elicit a higher recognition than words with lower frequencies. We know that the Reuter Health news stories are targeted at lay readers, thus when a word shows a high frequency (in its all morphological variants) in this corpus, we can define it as "familiar" ("easy") for lay readers. In contrast, the words with high frequencies in PubMed articles, but rarely seen in the news stories, are more likely to elicit a higher requirement for professional knowledge, then this word tends to be more "unfamiliar" ("difficult") for lay readers. Therefore, if the term usage in EC texts is more similar to Reuter News, then we could conclude that the terms used in EC texts are generally not difficult for lay readers; otherwise, the demands of the EC texts are more like technical papers. **Figure 2** illustrates the 3-step approach we applied to evaluate term familiarity.



Lexical variants	EC.freq	Reuter. freq	PubMed .freq
Term 1	(0.4)	(0.6)	(0.3)
...	...	...	...
Term n	(0.2)	(0.1)	(0.2)

Figure 2. Our analysis schema for familiarity evaluation of terms

In order to compare the clinical trial corpora to both the Reuter news and PubMed corpora, 188 frequently used (> 5 occurrence) UMLS concepts were extracted from the latter two corpora by an in-house entity-recognizer named HealthTermFinder. It was possible for each concept to contain several lexical variants mapping to the same CUI (e.g. CUI: C0027051, lexical variants: heart attack, myocardial infarctions). We generated a frequency distribution of all lexical variants for each CUI in each corpus and then applied Kullback–Leibler divergence to evaluate pairwise differences between Reuter news/ EC texts, and \PubMed articles/ EC texts. When the difference between eligibility criteria texts and news was smaller, we defined the word usage in EC texts as relatively "easy" for lay readers. For example, for CUI C0027051, we determined the frequency of use of "heart attack" and "myocardial infarctions." Commonly, the latter is a more technical term used in the health domain and so we expect a lower frequency in the Reuter news corpus, e.g. 0.01. Say we found that the frequency distribution of C0027051 in the news corpus was (0.99, 0.01); then we would calculate the frequency of the same CUI in the PubMed article. We expect that the frequency of "myocardial infarctions" would be much higher in technical papers than in the news corpus, e.g. (0.5, 0.5). When we apply KL divergence between the two distributions of the compared (news and article) and the test (clinical trials) corpus, those having similar distributions of word frequency can be detected. After manually filtering the concepts detected by HealthTermFinder that were not discriminative enough to be distinguished as technical or general text, such as C0014522 - "at any time," and also those not existing in our selected eligibility criteria texts, 90 UMLS concepts were finally selected. The KL divergence of Q from P is defined as:

$$D_{KL}(P||Q) = \sum P(i) ln \frac{P(i)}{Q(i)}$$

3. Stratified analysis using clinical trial metadata

Considering that context may help with comprehension, we also took the complete text of each clinical trial into consideration. Besides the eligibility criteria, ClinicalTrial.gov has also archived the background information, the purpose, and some other detailed metadata of each clinical trial. To examine clinical trial metadata associated with readability of eligibility criteria, we classified the selected trials by their recruitment conditions and study results and chose SMOG and Dale-Chall scores as the indicators of required reading level. Further, we investigated clinical trial metadata that might be associated with a variance in reading levels. We tried to classify the trials using some other metadata (e.g., country that submitted the trial, the year that the trial was started). We selected two lexical features as indicators for the stratified analysis in that, 1) they are easy to compute, and 2) though the general readability formulas suffer from a lack of ability to precisely measure biomedical text readability, they are still able to indicate the discrepancies in required readability among different texts and enable stratified analyses. The text processing was carried out in Perl (v5.18.2), and the statistical analysis was performed using R 3.1.1.

Results

1. Features Evaluation

The overall readability measurement results are shown in **Table 3**. The first two columns reflect the results for the two standard corpora used for comparison with the EC texts. Examining the shallow features, we can see that generally PubMed articles are more complex than their corresponding news stories, using longer sentences and more complex words. This is consistent with our intuition about scientific papers. A more in-depth look at the syntactic features reveals that the PubMed articles tend to use more nouns in each sentence while Reuter news stories tend to use more verbs and subordinate clauses. This result also is consistent with the characteristics of each type of literature. The news style, used for news reporting in mass media, tends to use subject-verb-object construction and vivid, active prose, to inform the public compared to the corresponding scientific papers, which follow the style of technical writing, having longer sentences including a series of nouns or noun phrases. Also, at the word level, as a rule, journalists try not to rely on jargon and will not use a long word when a short one will work the same, because their readers are the masses and so have completed various levels of education. Scientists, however, are more likely to use professional terminology. These contrasting features are reflected in the lexical feature results –both methods used graded PubMed articles much higher than the news. For Reuter news, according to Table 1 and 2, both SMOG and Dale-Chall grades indicate a reading level requirement of around grade 13-14, reflecting the consistency of the two kinds of measurement. However, discrepancy regarding the reading grade level exists between the measurement results (grade 13-14, college) and the claim made by Reuter news (grade 12, high school). This can be explained by the findings in a review of readability[16] -- SMOG usually results in a score one or two grade levels higher because it is based on stricter criteria for readability, like other similar assessment methods. Overall, PubMed articles tend to use longer and more unfamiliar terms than the news, requiring readers to have at least a college degree to fully comprehend them.

Table 3. Overall results of readability measurements

	Features	Reuter	PubMed	CT texts	EC texts	EC texts (95 CI)
Shallow Features	aWPS	21.24	23.30	16.34	8.86	(8.83, 8.90)
	aSPW	1.46	1.65	1.80	1.87	(1.87, 1.88)
	aPPS	3.05	5.28	3.98	2.46	(2.45, 2.46)
Syntactic Features	aNP	11.36	12.97	7.12	4.09	(4.07, 4.11)
	aVP	5.89	3.45	2.51	1.61	(1.60, 1.61)
	aSBr	1.10	0.41	0.29	0.15	(0.15, 0.15)
Lexical Features	SMOG	13.02	16.14	14.62	12.61	(11.88, 11.90)
	Dale-Chall	9.82	11.72	11.73	11.64	(12.08, 12.10)

**** The 95% Confidence Interval for averages of the scores of eligibility criteria from the entire database were calculated using 1000 bootstrap***

After ensuring that we had a full understanding of the results for the two standard corpora, we interpreted the corresponding results for the EC texts. Two kinds of EC texts were assessed here. As mentioned above, we were also concerned whether other parts of the trials documents might help readers to understand the EC of clinical trials besides the EC texts we were focusing on. Thus, the third column in **Table 3** represents the assessment results for the complete document text of the clinical trials (including purpose, detailed description, eligibility criteria, etc.), which we refer to as "CT texts" here. Finally, the fourth column is the assessment results for the eligibility criteria text ("EC text") alone. Compared to the Reuter and PubMed corpora, EC texts tend to use much shorter sentences (aWPS) and all 3 syntactic features indicate a particularly simple syntactic structure, especially in the EC text.

The reason for these results can be explained by the sample texts in **Table 1**. Eligibility criteria basically consist of short descriptions or the constraints for the characteristics of the target cohorts, and most of time they are not even a complete sentence, leading to particularly low results for aNP, aVP, and aSBr. In contrast, the syntax of the CT text is closer to that of the standard corpora and the assessment results with respect to lexical aspects (aSPW, aPPS, SMOG and Dale-Chall) further indicate that the CT texts are more similar to PubMed articles, indicating a college reading level requirement. However, discrepancy appeared between the results for the EC text in SMOG and those in Dale-Chall. The SMOG grade indicated that this corpus requires only a high school reading level, even lower than that of Reuter news, while the Dale-Chall grade indicated that it needs at a least college level reading skill to understand, meaning the EC text is as difficult as that in the PubMed corpus. The most likely explanation for this result is that while EC text uses many short terms, they are unfamiliar to the public, but this cannot be accounted for when using SMOG methods, because SMOG only takes the number of polysyllables into consideration. For example, some technical terms, e.g. "heparin", are much shorter than some familiar words, e.g. "characteristics", but are more difficult to understand for lay readers. Moreover, we also observed a large amount of professional abbreviations present in EC texts (e.g. "AD", "KPS"), which also have fewer syllables, but present more difficulties for readers with low health literacy. The Dale-Chall scoring system, in contrast to SMOG, takes familiarity of the words into major consideration, perhaps explaining why it determined EC texts to be as difficult to read as scientific papers. Thus, in conclusion, for EC texts, the syntactic structure may not cause problems for readers, but the large number of unfamiliar terms (e.g. technical jargon) used can be a major obstacle for lay readers, who lack a high level of health literacy. The other parts of the text in clinical trials literature were found to require an even higher reading level (see **Figure 3.** Each dot represents a clinical trial. Most of the dots are above the diagonals), consistent with the recent results for the readability assessment of the detailed descriptions in CT texts from clinicaltrials.gov by Wu et al.[18] Specifically, in their paper, they reported that detailed descriptions required an education level of grade 18 to understand. Therefore, it seems that other parts of the CT texts cannot offer much help to readers with respect to understanding eligibility criteria.

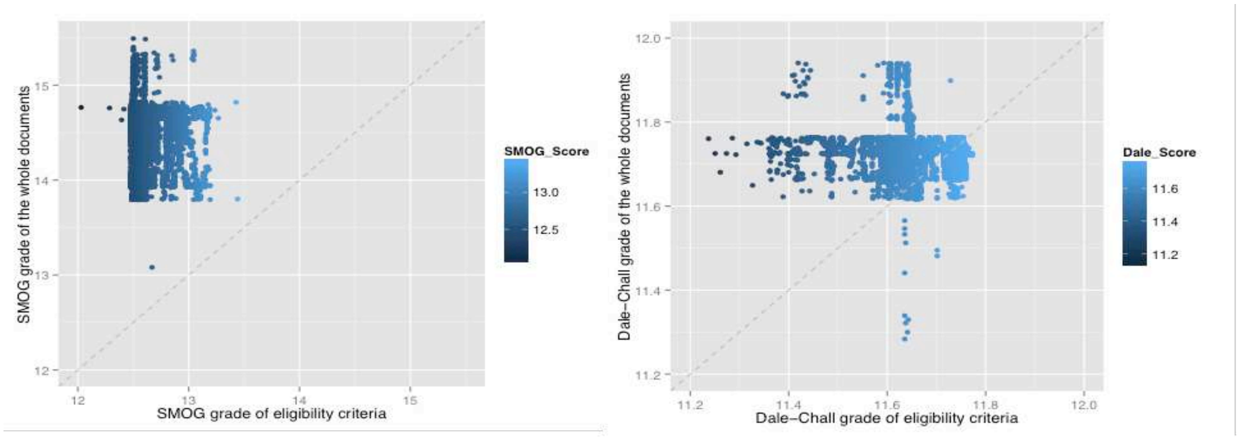


Figure 3. Lexical feature comparison between eligibility criteria and the according whole documents. Left: SMOG Grade; Right: Dale-Chall Grade. The x-axis is the SMOG grade of the whole clinical trial summary and the y-axis is the SMOT grade of the eligibility criteria text.

2. Term familiarity evaluation

The results for lexical features in Table 3 have led us to conclude that the terms used in EC texts are as unfamiliar for the public as those used in scientific papers, even though sometimes the words or terms are not that long and complex. Those findings agree with our hypothesis that technical jargon is one of the major obstacles to lay readers comprehending EC texts. To further confirm this hypothesis, we performed a term familiarity evaluation. As described in the methods section, we chose 90 UMLS concepts to evaluate for term familiarity by computing the KL divergence of the EC texts from both the PubMed and the Reuter corpora. If the divergence of term usage between EC texts and PubMed articles is smaller than that of the news articles, then the conclusion would be that EC are organized and composed with unfamiliar and more technical terms like professional papers. Otherwise, the reverse would be true. For instance, CUI C0027051 has the lexical variants "heart attack" and "myocardial infarction." In our eligibility criteria text, "heart attack" occurred 434 times while "myocardial infarction" occurred 5810 times, accounting for 7% and 93% of the text, respectively. In Reuters, this set of frequencies was 100% and 0%; while in the PubMed corpora, it was 50% and 50%. The results for KL divergence showed that for this set of lexical variants, the eligibility criteria text was closest to PubMed articles. The 90 comparison results after 180 times performing KL divergence are shown in **Figure 4**. Only eight sets of lexical variants from the same CUI could not determine which

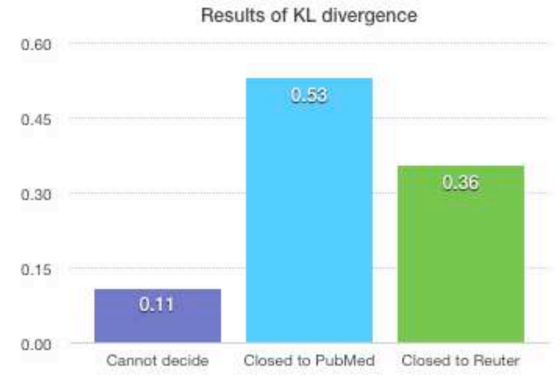


Figure 4 Results of KL divergence for 90 chosen UMLS CUIs to perform word familiarity prediction.

corpus our eligibility criteria was closer to. Among the rest of the valid values obtained, about 60 % of the concept usage results indicated a more technical lexical variant of the same concept was likely to be chosen to elaborate the EC. Again, this kind of term usage pattern is likely to increase the difficulty for out-of-domain readers to understand the criteria.

3. Clinical Trial Stratification analysis

The readability assessment of the EC texts from clinicaltrials.gov has come to a conclusion, but in order to make use of our results to improve the readability of clinical trials, we extended our research one step further and did a series of preliminary analyses to guide future work and in the hopes that we could find factors related to the changes in required reading level. We stratified the trials according to different features: source country, study type, recruitment conditions and whether they had final results. From all 120,977 files used for general analysis, we filtered out the data that had missing records for these features, which left 46,137 trial files to be examined in our stratified analysis.

Start year and study results.

Only 4,449 of the 46,137 files, less than 1/10 of all the clinical trials examined, included final study results. This is depicted in **Figure 5** – where one dot represents one documented trial. There are several possibilities that may lead to no result being recorded. Often it's either because the trial was not completed yet or the studies were terminated halfway. As the graphs also show, the red dots (trials with recorded results) are all located on the right side, indicating that the clinical trials with recorded results are more likely to be the trials started more recently. One way to explain this fact could be that the related knowledge and findings have grown tremendously in recent years, or the technology to document related records has advanced. For this reason, we were not surprised to observe a significant increase in numbers of records in the early 21st century, with the variance in required reading level of the eligibility criteria being enlarged since then as well.

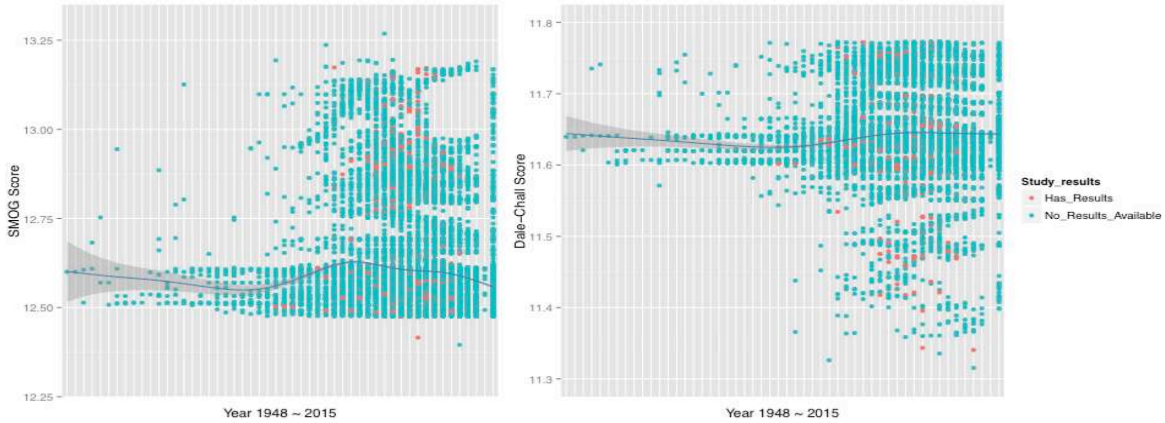


Figure 5. Distribution of lexical features among years according two readability measures. The red dots present the trials with final results, and the green points represent this without results records

Study type.

The two graphs in the upper row describe the readability of lexical features for eligibility criteria, and show no significant distinction among the different types of trials. However, when evaluating the whole documents (CT texts; two graphs in the lower row of **Figure 5**), the red and purple dots clearly split into separate groups, with the group of purple dots, representing interventional trials, clearly requiring above-average literacy to comprehend. This separation indicates that the discrepancy in required reading levels between the two major kinds of trials is more distinct than when considering EC texts alone. The results also show that, the observational studies were much easier to read than the interventional studies, which is consistent with the internal complexity of the interventional studies. However, as the number of submitted clinical trials began to increase rapidly, the readability of the two types converged to be more similar.

This very interesting phenomenon requires further study to determine what factors influenced this convergence. In comparing the results for EC texts alone (upper row in Fig 5) and the whole trial documents (lower row), it's obvious that the complexity of the interventional studies with respect to reading is not mainly caused by the eligibility criteria text but by other parts of the trial documents, e.g., detailed descriptions of the studies. **Figure 6**

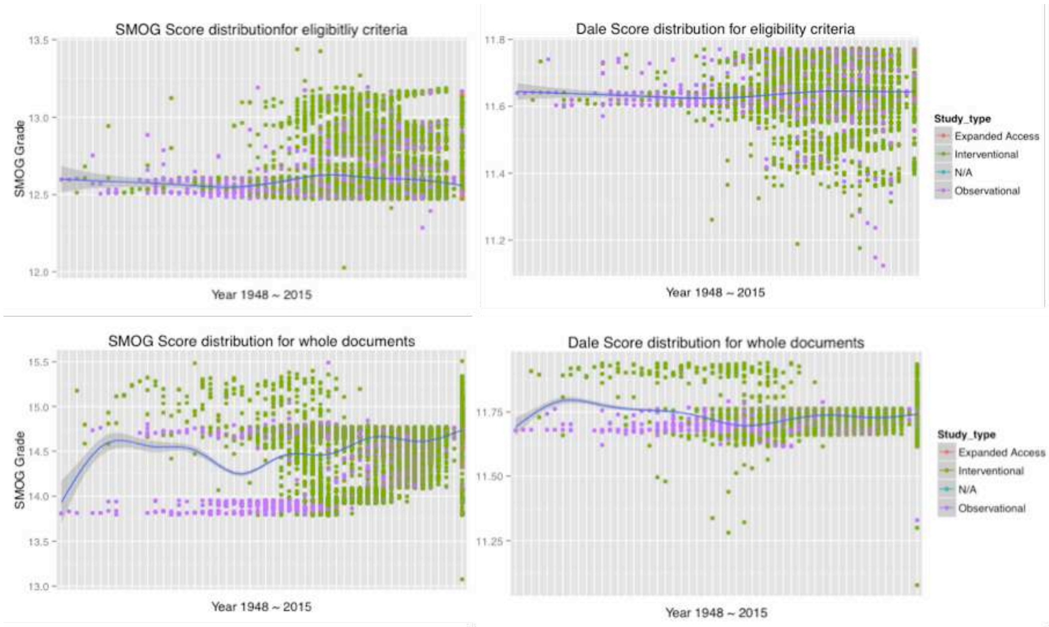


Figure 6. The lexical features classified by study type for eligibility criteria. Each row represents a kind of text, and each column represents a kind of lexical feature.

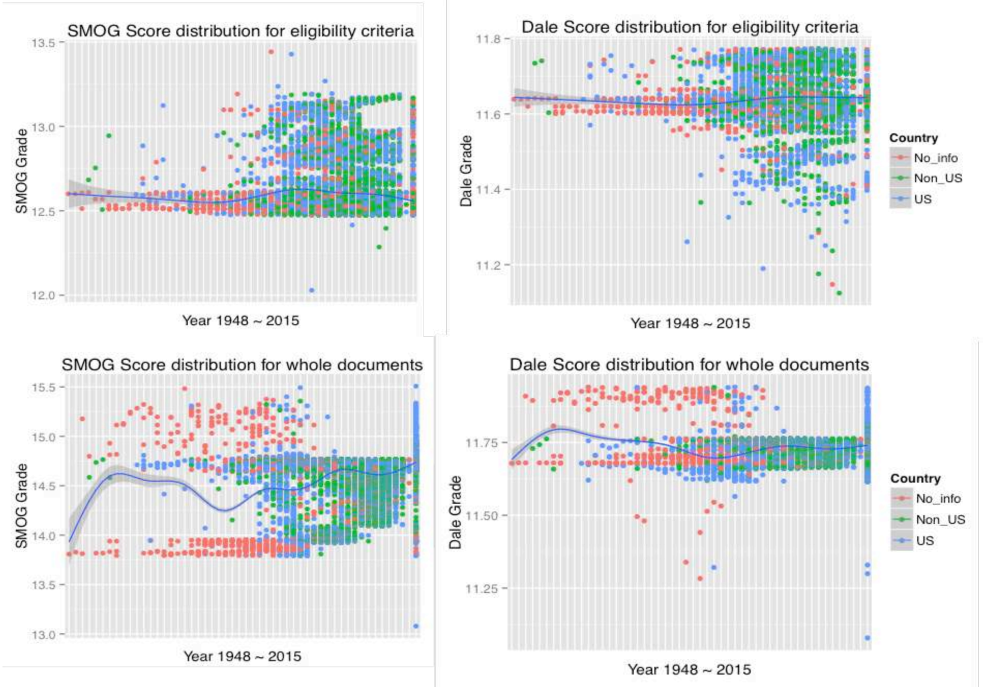


Figure 7. The lexical features classified by source countries. Each row represents a kind of text, and each column represents a kind of lexical feature.

reveals that observational (purple) and interventional (green) studies comprise the majority of the documents. It also shows that, as the 21st century approached, the number of interventional studies increased dramatically, a reasonable result given the advances made in medical science.

Source country.

Before the study, we hypothesized that whether English was the first language for the researchers who wrote those trial summaries might also influence the readability of the trial documentations. Since the number of U.S. and Non-U.S studies in the corpus were relatively equivalent (39% and 45%), we were able to classify trials for source country into one of only two groups: U.S. or non-U.S. (includes many non-English speaking countries, e.g. Germany, China, Japan). We then explored the differences in the readabilty of the trials according to country where the trial was being performed. It's interesting to notice that when stratified the whole documents (the two graphs in lower row of **Figure 7**), the trials with no location recorded showed either much higher or lower readability than the average level. We sorted all the trials by their startyears and noticed that most of the trials short of country information were started before 2000, and account for a large part of all documents before 1990s. We manually reviewed some trial documents with extrem readability scores and noticed that the low scores (easy to read) were always gained because the incompleteness of the trial documents, while the extremely high scores (difficult to read) were caused by multiple reasons, e.g. internal complexity of the study disease. One intersting thing we also found is that most of the low-readability trials without country information while started in recent year (after 2000) were those sponsored by large international pharmaceutical companies (which cause no country information recorded), and this kind of trials tend to be more difficult for readers compared to those conducted by federal-funded research institutes, universities or hospitals .

Discussion

Implications of the findings

According to our results, to gain a full comprehension of contents of EC, the readers have to be at least college education level, which is higher than the average American, who has achieved a high-school level. Therefore, in order to allow the many kinds of people who are looking for information in this online database, especially out-of-domain readers, to understand EC, there is a need for the website and researchers to be more careful in expressions and take those health consumers who are at a lower education level into consideration. Particularly, it was found that the main difficulty in understanding the contents is not a result of the syntactic features, but of the lexical ones, meaning that it is not a lot of long, complex sentences confusing people, but rather, the technical terms frequently used that increases the reading level required for target readers. Hence, the readability problem might be solved by using fewer technical jargons including clinical terminology and abbreviations, and instead considering lexical variants designed for patients, like consumer health vocabularies.

To realize this goal, we need to establish a comprehensive consumer health dictionary. However at present, no existing consumer health terminology can be called comprehensive or complete. It is a non-trivial effort to develop a comprehensive terminology that can keep up with the face pace of the constantly evolving consumer vocabulary. Most of the existing consumer vocabulary lists are based in a particular disease domain or user community. Given this situation, using existing consumer terminology to translate from clinical terminology is not realistic. Therefore, for further development, either existing "partial" consumer health vocabulary lists could be used to translate clinical trials in different fields, which lists could then be linked together, or an attempt could be made to find a method to establish a new comprehensive consumer health terminology, which would be a huge, long-term research effort.

Limitations

This study has several limitations. First, we only used computational measures to estimate readability. In the future, it would be valuable to engage enough real users, including both patients and clinicians, to validate our computational results and to obtain their feedback on their real-life experience. Second, the stratified analysis only serves as preliminary work for future guidance. The interpretation of the stratified analysis results here is just some reasonable guessing and is one of many possibilities without solid statistical tests to confirm. In the future, more relative study should be carried out based on refined stratification to find more interactive information.

Conclusions

In this study we presented an initial assessment of the readability of clinical trials eligibility criteria in ClinicalTrials.gov. We used both general popular readability formulas and health-domain specific methods, covering three aspects of readability: the unit length of a text, the syntactic complexity, and the lexical complexity. Moreover, we also conducted a comparison with two other health-related corpora to assess the term use tendency of the clinical trials. The overall results showed that eligibility criteria texts are beyond the comprehension capacity of the general American population, and the main reason for that is the frequent use of professional terminology and technical jargons. To guide future works on improvement, we also classified the trial texts by different factors and study the readability, trying to find some properties of the trials that might impact the readability. In the results, we found that in the beginning of the 21st century, the number of documented clinical trials increased greatly. Of all the properties we have studied, different study types of clinical trials might result in different levels of readability: generally, interventional studies require a higher education level to understand than observational studies. We also found that clinical trials sponsored by pharmaceutical companies tend to have higher reading level requirement compaired to federally-funded studies. Overall, our study presented a systematic evaluation and analysis of this online clinical trials database and contributed evidence for future improvement.

Acknowledgments

This study was sponsored by the U.S. National Library of Medicine grant **R01LM009886** (PI: Weng) and U.S. National Center for Advancing Translational Science grant **UL1 TR000040** (PI: Ginsberg).

References

1. Lovato LC, Hill K, Hertert S, Hunninghake DB, Probstfield JL. Recruitment for controlled clinical trials: literature summary and annotated bibliography. Controlled clinical trials 1997;18:328-52.
2. McDonald AM, Knight RC, Campbell MK, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials 2006;7:9.
3. Dale E, Chall JS. The concept of readability. Elementary English 1949;26:19-26.
4. McLaughlin GH. SMOG grading: A new readability formula. Journal of reading 1969;12:639-46.
5. Davison A, Kantor RN. On the failure of readability formulas to define readable texts: A case study from adaptations. Reading research quarterly 1982:187-209.
6. Chall JS. Readability revisited: The new Dale-Chall readability formula: Brookline Books Cambridge, MA; 1995.
7. Yngve VH. A model and an hypothesis for language structure. Proceedings of the American philosophical society 1960:444-66.
8. Feng L, Elhadad N, Huenerfauth M. Cognitively motivated features for readability assessment. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics; 2009: Association for Computational Linguistics. p. 229-37.
9. Schwarm SE, Ostendorf M. Reading level assessment using support vector machines and statistical language models. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; 2005: Association for Computational Linguistics. p. 523-30.
10. AlKhalili R, Shukla PA, Patel RH, Sanghvi S, Hubbi B. Readability Assessment of Internet-based Patient Education Materials Related to Mammography for Breast Cancer Screening. Academic radiology 2014.
11. Edmunds MR, Denniston AK, Boelaert K, Franklyn JA, Durrani OM. Patient Information in Graves' Disease and Thyroid-Associated Ophthalmopathy: Readability Assessment of Online Resources. Thyroid 2014;24:67-72.
12. Kong KA, Hu A. Readability Assessment of Online Tracheostomy Care Resources. Otolaryngology--Head and Neck Surgery 2014:0194599814560338.
13. Phillips NA, Vargas CR, Chuang DJ, Lee BT. Readability Assessment of Online Patient Abdominoplasty Resources. Aesthetic plastic surgery 2014:1-7.
14. Sharma N, Tridimas A, Fitzsimmons PR. A readability assessment of online stroke information. Journal of Stroke and Cerebrovascular Diseases 2014;23:1362-7.
15. Patel CR, Sanghvi S, Cherla DV, Baredes S, Eloy JA. Readability Assessment of Internet-Based Patient Education Materials Related to Parathyroid Surgery. Annals of Otology, Rhinology & Laryngology 2015:0003489414567938.
16. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. Health Education & Behavior 2006;33:352-73.
17. Kim H, Goryachev S, Rosemblat G, Browne A, Keselman A, Zeng-Treitler Q. Beyond surface characteristics: a new health text-specific readability measurement. AMIA Annual Symposium Proceedings; 2007: American Medical Informatics Association. p. 418.
18. Wu D, Hanauer D, Mei Q, et al. Assessing the readability of ClinicalTrials.gov. J Am Med Inform Assoc 2015.
19. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. AMIA Summits on Translational Science Proceedings 2010;2010:46.
20. Health UNIo. ClinicalTrials. gov. 2012.
21. Elhadad N. Comprehending technical texts: Predicting and defining unfamiliar terms. AMIA Annual Symposium proceedings; 2006: American Medical Informatics Association. p. 239.
22. Klein D, Manning CD. Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1; 2003: Association for Computational Linguistics. p. 423-30.