A Parallel Corpus Analysis of Text and Audio Comprehension and an Evaluation of the Effectiveness of Readability Formulas

Arif Ahmed, Gondy Leroy, David Kauchak, Prosanta Barai, Philip Harber, Steven A. Rains

Submitted to: Journal of Medical Internet Research
on: December 09, 2024

Table of Contents

A Parallel Corpus Analysis of Text and Audio Comprehension and an Evaluation of the Effectiveness of Readability Formulas

Arif Ahmed[1] MS, BSC; Gondy Leroy[1] PhD; David Kauchak[2] PhD; Prosanta Barai[1] MS, BSC; Philip Harber[1] MD; Steven A. Rains[1] PhD

[1]The university of Arizona Tucson US
[2]Pomona College 333 N College Way Claremont US

Corresponding Author:
Arif Ahmed MS, BSC
The university of Arizona
1200 E University Blvd
Tucson
US

Abstract

Background: Health literacy, the ability to understand and act on health information, is critical for patient outcomes and healthcare system efficiency. While plain language guidelines enhance text-based communication, audio-based health information remains underexplored, despite the growing use of virtual assistants and smart devices in healthcare. Traditional readability formulas, such as Flesch-Kincaid, provide limited insights into the complexity of health-related texts and fail to address challenges specific to audio formats. Factors like syntax and semantic features significantly influence comprehension and retention across modalities.

Objective: This study investigates features that affect comprehension of medical information delivered via text or audio formats. We also examine existing readability formulas and their correlation with perceived and actual difficulty of health information for both modalities.

Methods: We developed a parallel corpus of health-related information that differed in delivery format: text or audio. We used text from BMJ Lay Summary (N = 193), WebMD (N = 40), Patient Instruction (N = 40), Simple Wikipedia (N = 243), and BMJ Journal (N = 200). Participants (N = 487) read or listened to a health text and then completed a questionnaire evaluating perceived difficulty of the text measured using a 5-point Lickert scale and actual difficulty measured using multiple-choice and true-false questions (comprehension) as well as free recall of information (retention). Questions were generated by ChatGPT 4.0. Underlying syntactic, semantic, and domain-specific features, as well as common readability formulas, were evaluated for their relation to information difficulty.

Results: In general, the text versions were perceived as easier than the audio versions. The features of the underlying text were related to both perceived and actual difficulty. Longer texts were perceived to be more difficult in text than audio, while free recall decreased with longer texts in both modalities. Higher content word frequency was associated with lower perceived difficulty in audio and improved free recall results for text. Verb-heavy content was easier to comprehend, especially in audio, while noun- and adjective-heavy content increased difficulty. Finally, readability formulas are found ineffective in assessing information difficulty, except for the perceived difficulty in the text condition.

Conclusions: Text was more effective for conveying complex health information, but audio can be suitable for easier content. In addition, several textual features affect information comprehension and retention for both modalities. Finally, existing readability formulas did not explain actual difficulty. This study highlighted the importance of tailoring health information delivery to content complexity by using appropriate style and modality.

(JMIR Preprints 09/12/2024:69772)
DOI: https://doi.org/10.2196/preprints.69772

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

A Parallel Corpus Analysis of Text and Audio Comprehension and an Evaluation of the Effectiveness of Readability Formulas

[a]Arif Ahmed, [a] Gondy Leroy, [b] David Kauchak, [a] Prosanta Barai, [a]Philip Harber,

[a] Stephen A. Rains

[a] The University of Arizona, Tucson 85721, U.S.A
[b] Pomona College, 333N College Way, Claremont 91711, U.S.A

Abstract

Background: Health literacy, the ability to understand and act on health information, is critical for patient outcomes and healthcare system efficiency. While plain language guidelines enhance text-based communication, audio-based health information remains underexplored, despite the growing use of virtual assistants and smart devices in healthcare. Traditional readability formulas, such as Flesch-Kincaid, provide limited insights into the complexity of health-related texts and fail to address challenges specific to audio formats. Factors like syntax and semantic features significantly influence comprehension and retention across modalities.
Objective: This study investigates features that affect comprehension of medical information delivered via text or audio formats. We also examine existing readability formulas and their correlation with perceived and actual difficulty of health information for both modalities.
Method: We developed a parallel corpus of health-related information that differed in delivery format: text or audio. We used text from BMJ Lay Summary (N=193), WebMD (N=40), Patient Instruction (N=40), Simple Wikipedia (N=243), BMJ Journal (N=200). Participants (N = 487) read or listened to a health text and then completed a questionnaire evaluating perceived difficulty of the text measured using a 5-point Lickert scale and actual difficulty measured using multiple-choice and true-false questions (comprehension) as well as free recall of information (retention). Questions were generated by ChatGPT 4.0. Underlying syntactic, semantic, and domain specific features, as well as common readability formulas were evaluated for their relation to information difficulty.
Results: In general, the text versions were perceived as easier than the audio versions. The features of the underlying text were related to both perceived and actual difficulty. Longer texts were perceived to be more difficult in text than audio, while free recall decreased with longer texts in both modalities. Higher content word frequency was associated with lower perceived difficulty in audio and improved free recall results for text. Verb-heavy content was easier to comprehend, especially in audio, while noun and adjective-heavy content increased difficulty. Finally, readability formulas are found ineffective in assessing information difficulty, except for the perceived difficulty in the text condition.
Conclusion: Text was more effective for conveying complex health information, but audio can be suitable for easier content. In addition, several textual features affect information comprehension and retention for both modalities. Finally, existing readability formulas did not explain actual difficulty. This study highlighted the importance of tailoring health information delivery to content complexity by using appropriate style and modality.

Keywords: health literacy, parallel corpora, generative AI, user evaluation, text, audio,

comprehension, retention, actual difficulty, perceived difficulty.

1. Introduction

Clear and understandable information is vital in healthcare to enhance health literacy. Improving health literacy is a significant national goal with numerous benefits [1] such as making better-informed choices, managing chronic illnesses, averting health issues, and reducing healthcare expenses [2, 3]. The U.S. National Action Plan aims to improve health literacy, emphasizing the need for clear health-related information, while the Plain Writing Act promotes clarity in communication [4]. Although there are many plain language guidelines for delivering health information through text, audio is often ignored. Incorporating audio guidelines for providing health information could offer a substantial opportunity to improve health literacy [5, 6].

Audio formats are gaining popularity as mobile devices and smart speakers with virtual assistants (e.g., Siri, Alexa) become increasingly prevalent for accessing information. In 2022, approximately 142 million people in the United States used virtual assistants, representing nearly half of the population. By 2026, this number is projected to rise to 157.1 million users [7]. These devices are also incorporated into healthcare settings, allowing patients to converse with medical professionals and seek information [8]. Patients can pose health-related questions and receive responses, provided the information is clear and understandable. In 2019, health-related inquiries accounted for sixteen percent of all smart speaker interactions [9]. American adults' utilization of virtual assistants for healthcare queries experienced a notable increase, surging from 19 million in 2019 to 51.3 million in 2020 and further to 54.4 million in 2021 [10].

There are currently no established metrics for measuring the difficulty of health information conveyed through audio [11]. Yet, text difficulty metrics such as the Flesch-Kincaid Grade Level, Gunning Fog Index, and Dale-Chall Readability Score are prevalent. These metrics typically consider factors such as sentence length, word frequency, and syllable count to estimate the reading level or educational grade required to understand a given text. While these metrics provide useful approximations of text complexity, they have limitations and fail to account for factors like background knowledge or syntactic or semantic features of texts [12].

Our goal is to analyze how health information is comprehended when delivered via text versus audio. We aim to evaluate which method is more effective in facilitating understanding and retention of the information. In addition to this comparison, we investigate the features of underlying text and its effect in each modality to determine their respective strengths and weaknesses in conveying health-related content.

2. Background

2.1 Health Literacy/Plain Language

Health literacy is crucial to effective healthcare communication and patient education. Health literacy refers to an individual's capacity to obtain, process, and comprehend basic health information and services necessary to make appropriate health decisions. This concept extends beyond the ability to read and write, encompassing skills such as numeracy, decision-making, and critical thinking in health contexts [13]. The importance of health literacy cannot be overstated, as it directly impacts patient outcomes and healthcare system efficiency. Individuals with low health literacy often struggle to understand medical instructions, manage chronic conditions, and navigate complex healthcare systems [14]. This can increase hospitalization rates, poor medication adherence, and higher healthcare costs. Recognizing this, healthcare providers and policymakers have increasingly focused on improving health literacy through various initiatives [15].Plain language is a key strategy in addressing health literacy challenges. It involves using clear, concise, and jargon-free communication

to convey complex health information. The principles of current plain language advice include using everyday words, short sentences, active voice, and logical organization of information [16]. By implementing plain language practices, healthcare providers can significantly improve patient understanding, leading to better health outcomes and increased patient satisfaction [17].

2.2 Difficulty Measures

The readability of source texts is another critical factor in health communication. Readability refers to the ease with which a reader can understand written text. In healthcare, ensuring that patient education materials, consent forms, and medical instructions are easily readable is paramount. Various factors, including vocabulary complexity, sentence length, and overall text structure influence readability [18]. Readability formulas, such as Flesch-Kincaid, are commonly employed to evaluate text readability [19]. However, recent research indicates that Flesch-Kincaid is ineffective in evaluating text simplification metrics [20].

Recent endeavors to simplify text have concentrated on semantic and syntactic analysis to identify features that mitigate text difficulty. Qualitative elements such as linguistic conventions, clarity, depth of content, and the reader's background knowledge significantly impact perceived difficulty. Quantitative measures such as word frequency, grammar frequency, length, and sentence structure also impact understanding [21] . For example, texts that are harder to read often have fewer verbs and function words, more nouns, and more complex vocabulary [22]. Lexical chains (i.e., topics in text), their length, and how they are intertwined can also differentiate between difficult and easy texts [23]. Metrics such as specificity, ambiguity, concept density, and topic density are used to determine the difficulty of medical texts [24]. Advanced tools like Coh-Metrix analyze multiple dimensions of text, including cohesion, syntactic complexity, and word concreteness. These analyses provide a more nuanced understanding of what makes a text challenging. Such measures are particularly valuable in educational settings and in developing materials for diverse audiences [25].

Although research focusing specifically on delivering information in audio format is much more limited, there is evidence that features like audio speech rate, pause, and emphasis can play an important role in the comprehension of audio content [26, 27]. The Speech Transmission Index (STI) and Speech Intelligibility Index (SII) are examples of tools used to evaluate the clarity of spoken content [28]. Further research is needed to explore how text and audio features affect the difficulty of health-related information [29]. Audio difficulty measures are important, especially in an era where health information is increasingly disseminated through audio formats via virtual assistants. In healthcare settings, audio difficulty measures are needed to ensure that verbal instructions, telemedicine consultations, and health-related audio content are accessible to all patients, including those with hearing impairments or non-native speakers. By considering both text and audio difficulty, healthcare providers can create more inclusive and effective communication strategies. In this study, we evaluate user comprehension and retention of information delivered via text and audio. We also examine the features of the texts that determine the comprehensiveness of the information in both modalities and assess the effectiveness of existing readability formulas in evaluating the difficulty of the information.

3. Methodology

3.1 Corpus Creation

We developed a parallel corpus of health-related texts using common disease and health conditions listed in the International Classification of Diseases-10 (ICD-10) coding system [30]. We extracted health related text from five sources representing a range of difficulty and styles: BMJ Journal (200 texts), BMJ Lay Summary (193 texts), WebMD (40 texts), Patient Instruction (40 texts),

and Simple Wikipedia (243 texts). These were chosen to create a diverse corpus. WebMD and Patient Instruction have fewer texts since they did not contain text for all diseases in our list. The BMJ (British Medical Journal) is a peer-reviewed medical journal that publishes highly technical, scientific content written by medical experts and researchers. The language is largely technical, with extensive use of medical terminology and detailed scientific information. The tone is formal and objective, adhering to strict academic standards. In contrast, BMJ Lay Summaries present the key findings and implications in a more accessible, easy-to-understand manner for a general audience. The language is largely non-technical, the tone is more friendly and engaging, and the structure is more narrative-driven. WebMD is a popular online health resource for a general, non-medical audience. The articles on WebMD cover a wide range of health-related topics, using relatively simple and straightforward language with minimal technical jargon. The tone is informative and conversational, designed to educate and empower readers to make informed decisions about their health. Patient Instructions are educational materials specifically designed to provide clear, step-by-step guidance to patients on various medical procedures, treatments, or self-care practices. The language used is simple, direct, and easy to understand, focusing on providing practical, hands-on information. The tone is friendly and reassuring, and the structure is highly organized, often with numbered or bulleted steps. Simple Wikipedia is a version of the popular online Wikipedia that uses simpler language and explanations to make the content more accessible to a wider audience. The language is straightforward, with shorter sentences, simpler vocabulary, and minimal use of complex terminology or jargon. The tone is informative and objective, and the structure is typically organized with clear headings, subheadings, and bullet points to enhance readability and comprehension.

To create the parallel corpus with the texts presented in audio format, we used Microsoft Azure's text-to-speech service using the default US male voice and the default speech rate settings.

3.2 Text Features

We generated information features for each text based on our and other researchers' findings of important textual elements that have been shown to impact readability, understandability, and retention (see Table 1) [22, 23, 31-34]. Overall, there are four groups of features.

The first group of features, "Ordinariness", focuses on text difficulty using word and grammar frequencies [23]. Content word frequency, indicating vocabulary familiarity, is highest for Simple Wikipedia at 498 million, followed by WebMD (387 million), with the BMJ Journal having the lowest frequency (236 million), showing more jargon language. Grammar frequency is highest in BMJ Lay Summaries (9,882), while Patient Instructions (3,108) and BMJ Journal (4,916) are lower, reflecting their more syntactical nature.

The second group, "Healthcare Domain Specialty", comprises specificity, ambiguity, concept density, and topic density. In specialized medical content, these metrics reflect how difficult a text is based on the complexity of medical terms. Texts with more specific, ambiguous, or conceptually dense terms are harder to comprehend [24]. Specificity for WebMD and Patient Instructions are high, with values of 0.384 and 0.364, respectively, while Simple Wikipedia has lower specificity (0.122). Ambiguity is highest for WebMD (0.433), followed closely by Patient Instructions (0.422), suggesting more generalized content. BMJ Journal (0.337) and Simple Wikipedia (0.1557) show the lowest ambiguity. Concept density and topic density are highest for WebMD (0.388 and 0.505, respectively), indicating more information packed into the content, while Simple Wikipedia is the least dense, with values of 0.114 and 0.329.

The third group includes "Parts-Of-Speech Features" where a higher proportion of nouns and adjectives is linked to more difficult text, while more verbs and adverbs make text easier [24]. For parts of speech, the BMJ Journal has the highest percentage of nouns (34.28%) and adjectives (12.88%), indicating difficult content. WebMD uses more verbs (19.51%), making it easier, while Patient Instructions use the highest percentage of nouns (32.35%) [22]. Adverbs are used more

frequently in WebMD (5.86%), while BMJ Journal has the fewest (3.37%).

The fourth group, “Topic Spread”, focuses on the topics in the text, measured by lexical chains [23]. lexical chain analysis helps evaluate topic distribution and repetition within the text, with longer chains and fewer overlaps associated with simpler texts [23]. WebMD has the highest lexical chain score (0.396), suggesting greater word repetition and coherence, while BMJ Lay Summaries and Simple Wikipedia have the lowest values (0.224 and 0.123, respectively). Lexical chain length and lexical cross chains also follow a similar pattern, with WebMD showing the highest values across these metrics, while Simple Wikipedia and BMJ Lay Summaries have the lowest, making them more varied and straightforward [23].

Table1: Text readability features for the corpora.

Features	**BMJ Lay Summary (N = 193)**	**WebMD (N = 40)**	**Patient Instructions (N = 40)**	**Simple Wikipedia (N =243)**	**BMJ Journal (N= 200)**
Average Word Count	205	595	961	353	237.335
Ordinariness					
Content Word Frequency	364,959,457.80	387,640,069.50	299,361,102.70	498,830,364.50	235767137.5
Grammar Frequency	9882.89	5495.91	3108.28	6883.26	4916.36
Healthcare Domain Specialty (Averages)					
Specificity	0.210	0.384	0.364	0.122	0.220
Ambiguity	0.245	0.433	0.422	0.1557	0.337
Concept Density	0.231	0.388	0.341	0.114	0.3472
Topic Density	0.411	0.505	0.559	0.329	0.4879

Parts-Of-Speech Features (%)					
Nouns	30.90	23.31	32.35	30.91	34.28
Verbs	16.82	19.51	16.34	17.06	12.605
Adverbs	4	5.86	2.87	4.94	3.37
Adjectives	10.34	9.21	7.84	9.60	12.875
Topic Spread (Averages)					
Lexical Chains	0.224	0.396	0.326	0.123	0.2973
Lexical Chain Length	0.484	0.277	0.354	0.310	0.3911
Lexical Chain Span	0.195	0.362	0.274	0.113	0.288
Lexical Cross Chains	0.222	0.405	0.330	0.122	0.2971

3.3 Study Overview and Participant Recruitment

Two parallel studies were conducted for each text source: an audio study to assess auditory information processing and a text study to evaluate textual information processing. The procedure was the exact same for each study. Participants were presented with either audio or textual health information and then asked to complete a questionnaire to evaluate their perception and understanding of the information.

Study participants were recruited through Amazon Mechanical Turk (AMT). First, we screened Amazon Mechanical Turk workers using standard criteria, including U.S. residency and a 98% approval rating. We then collected demographic information from a thousand workers through a survey, in which we also provided three audio snippets containing a word with varied noise levels. The workers who correctly identified at least two audio snippets were further invited to participated in the study. This approach made it possible to ensure that all participants were sufficiently able to hear the audio information.

Participants received $0.50 for each completed Human Intelligence Task (HIT). While completing multiple HITs was allowed, participants were prevented from evaluating duplicate content (e.g., both the audio and printed versions of a text) to prevent potential bias from repeated exposure. Each text or its audio version represented one HIT containing the stimulus material (either text or audio) followed by a short questionnaire containing a perceived difficulty evaluation, two multiple-choice (MC) comprehension questions, two true-false (TF) comprehension questions, an attention check question (for data cleaning purposes), and a free recall task.

3.4 Measures

Our study employed a multi-faceted approach to assess information comprehension. To measure perceived difficulty, participants evaluated the audio information using a 5-point Likert scale, ranging from very easy (1) to very difficult (5).

For comprehension measurement, we used questions designed to evaluate participants' understanding of the text. Two multiple-choice (MC) questions and two true-false (TF) questions were developed for each text using ChatGPT 4.0. All questions and answers were manually evaluated by a domain expert (i.e., a medical doctor) to ensure that they were relevant and appropriate.

To evaluate information retention, participants were asked to recall as much information as possible from the presented information. Participants were presented with a text box and instructed to type everything they could remember about the text. The analysis of free recall utilized two methods of comparing participant responses to the original text: exact word match percentage and semantic similarity percentage based on word embeddings (word2vec). The latter method allowed for a more nuanced recall evaluation, accounting for semantic understanding beyond verbatim reproduction.

Finally, we included attention-check questions. Participants were asked to identify the most frequently occurring word from a list based on the information provided. The responses to the attention check questions helped us evaluate each HIT response.

4. Results

4.1 Data

To get better quality data and a better evaluation of our study result, we only included answers from workers who passed a quality control process established in previous work that prohibits copy-pasting and incoherent content [35]. HTML-based tools are used to spellcheck and flag errors or nonsensical entries, prompting workers to improve their responses for clarity and professionalism. Table 2 shows the results of our data cleaning. We followed three filtering criteria to create two datasets: strict and lenient. First, we reviewed the correct response to the attention check question. If the response was correct, we included the HIT in the strict dataset. If the response was incorrect, the HIT was categorized under the lenient dataset. Then, we removed HITs from the entire dataset whose average accuracy on MC and TF questions was below 25%, indicating random guessing. And, finally, we verified free recall responses as appropriate or not. If it was not appropriate, we removed that HIT from the entire dataset.

In this results section, we describe the strict dataset results in detail. Because the results are very similar to those of the lenient dataset, the details of the lenient** dataset are in the Multimedia Appendix.

Table 2: Data cleaning steps.

Steps	Step 1	Step 2	Step 3	Participant

	(Participants removed)			s Retained
Lenient Dataset	**Attention check overlooked**	**MC & TF accuracy <= 25%**	**Free recall ≠ Appropriate**	
Strict dataset	**Attention Check considered**	**If MC & TF accuracy <= 25%**	**Free recall ≠ Appropriate**	

(H= 459)	**Text**	**Lenient**	0	9	88	482
		Strict	112	8	10	449
WebMD (H= 120)	**Audio**	**Lenient**	0	2	6	112
		Strict	30	5	2	83
	Text	**Lenient**	0	3	16	101
		Strict	19	1	8	92
Patient Instruction (H= 120)	**Audio**	**Lenient**	0	29	22	68
		Strict	44	17	9	50
	Text	**Lenient**	0	10	26	84
		Strict	22	3	15	80
Simple Wikipedia (H= 732)	**Audio**	**Lenient**	0	12	115	602
		Strict	231	19	41	438
	Text	**Lenient**	0	15	91	626
		Strict	143	6	64	519
BMJ Journal (H= 600)	**Audio**	**Lenient**	0	29	174	397
		Strict	71	28	162	339
	Text	**Lenient**	0	23	94	483
		Strict	108	16	74	402

Table 3 presents demographic information for the AMT workers whose data was retained for the analysis, dividing participants into two groups based on their interaction with either audio or text. There were 274 participants in the text condition and 213 participants in the audio condition.

Table 3: Participant demographic information.

	Strict Dataset	
	Audio	**Text**
Characteristic	**N (%)**	**N (%)**
Total	213	274
Sex		
Male	143 (67.13)	185 (67.51)
Female	69 (32.39)	89 (32.48)
Other	0 (0)	0 (0)
Age		
Younger than 30 years old	71 (33.33)	79 (28.83)
30 to 39 years old	103 (48.35)	151 (55.10)
40 to 49 years old	19 (8.92)	25 (9.12)
50 to 59 years old	13 (6.10)	12 (4.37)
60 to 69 years old	6 (2.81)	6 (2.18)
70 to 79 years old	1 (0.46)	1 (0.36)
Race		

Asian	6 (2.81)	3 (1.09)
American Indian/ Native Alaskan	1 (0.46)	1 (0.36)
Black or African American	1 (0.46)	2 (0.72)
Native Hawaiian or other Pacific Islander	0 (0)	0 (0)
White	199 (94.83)	263 (95.98)
Asian & White	6 (2.81)	5 (1.82)
Ethnicity		
Hispanic or Latino	60 (28.16)	72 (26.27)
Not Hispanic or Latino	153 (71.83)	202 (73.72)
Education (Highest Degree Achieved)		
Less Than High School	0 (0)	0 (0)
High School	2 (0.93)	2 (0.72)
Associate's degree	2 (0.93)	1 (0.36)
Bachelor's degree	165	214 (78.10)
Master's Degree	43 (77.46)	56 (20.43)
Doctorate Degree	0 (0)	0 (0)
Other Professional Degree	0 (0)	0 (0)
English Speaking		
Never English at Home	6 (2.81)	2 (0.72)
Rarely English at Home	7 (3.28)	6 (2.18)
Half of the time English at Home	6 (2.81)	7 (2.55)
Mostly English at Home	21 (9.85)	41 (14.96)
Only English at Home	173 (81.22)	218 (79.56)

Both conditions had a similar distribution of males (~67%) and females (~32%). Most participants were between 30 and 39 years old (48% in the audio condition and 55% in the text condition). About a third of the participants were younger than 30, and a small percentage were 40 or older. The racial makeup was predominantly White (more than 94% in both groups), and 28% and 26% identified as Hispanic or Latino in the audio and text conditions respectively.

Most participants had a bachelor's degree. A smaller percentage had a master's degree, and no participants held a doctorate. English proficiency was also strong, with over 79% of participants in both conditions reporting that they spoke only English at home.

4.2 Perceived Difficulty

Overall, information was perceived as easier to understand when presented as text compared to audio. Our t-tests show that the difference is statistically significant for BMJ Lay Summary and BMJ Journal. For BMJ Lay summary, text achieves a better perceived difficulty score (1.76) compared to audio (2.1), and for the original BML Journal the text achieves a better score (2.59) compared to audio (2.83). In contrast, for the WebMD and Patient Instruction sources, which contain easier-to-understand content, the differences are not statistically significant.

Table 4: Perceived Difficulty. (A lower value means perceived as easier)

	Perceived Difficulty (Strict Dataset)	
Test Sources (N)	**Audio (SD)**	**Text (SD)**

BMJ Lay Summary (193 texts) ****	2.1 (1.1)	1.76 (1.2)
WebMD (40 texts)	1.96 (1.0)	2.06 (1.0)
Patient Instructions (40 texts)	1.84 (1.0)	1.96 (1.1)
Simple Wikipedia (243 texts)	2.09 (1.2)	2.1 (1.2)
BMJ Journal (200 texts) *	2.83 (0.9)	2.59 (1.1)
Overall *	2.28 (1.1)	2.135 (1.1)

(Significance, * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001)

4.3 Actual Difficulty: Comprehension and Retention

4.3.1 Information Comprehension

Table 5 shows the average accuracy of answering the MC and TF questions. Text consistently yields higher comprehension performance (i.e., questions answered correctly) than audio. The difference is statistically significant for all sources. For example, in the BMJ Lay Summary, text achieves 72% accuracy compared to audio's 69%, while WebMD shows a more pronounced difference, with text reaching 75% and audio trailing at 55%. The largest gap is observed in Patient Instructions, where text comprehension soars to 86% versus audio's 66%. Similarly, for BMJ Journal, which features more difficult content, text achieves 76%, significantly higher than audio's 58%. These results highlight that text is more effective than audio in facilitating comprehension, especially for difficult health information.

Table 5: Accuracy (%) for MC and TF questions. (A higher value means greater comprehension)

	Actual Difficulty (Strict Dataset)	
Test Sources (N)	**Audio (SD)**	**Text (SD)**
BMJ Lay Summary (193 texts) **	69.38 (36)	72.27 (33)
WebMD (40 texts) ****	55.2 (38)	75 (31)
Patient Instructions (40 texts) ***	66.12 (33)	85.52 (29)
Simple Wikipedia (243 texts) ***	69.69 (40)	75.09 (35)
BMJ Journal (200 texts) ***	58.04 (36)	75.68 (32)
Overall ***	65.34 (37)	75.07 (34)

(Significance, * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001)

4.3.2 Information Retention

Overall, we found that information retention measured using exact word matching (See Table 6) outperforms information retention with text versus audio information. However, the difference is only significant for Patient Instructions and BMJ Journal.

Table 6: Exact matching word (%) for Free Recall. (A higher value means better retention)

	Number of exact words recalled (Strict Dataset)	
Test Sources (N)	**Audio (SD)**	**Text (SD)**
BMJ Lay Summary (193 texts)	9.02 (4)	9.29 (5)
WebMD (40 texts)	8.01 (5)	6.53 (4)
Patient Instructions (40 texts) *	4.09 (7)	6.43 (6)
Simple Wikipedia (243 texts)	9.2 (3)	11.17 (3)
BMJ Journal (200 texts) ***	2.67 (6)	6.91 (5)

Overall *	6.98 (5)	8.95 (4)

(Significance, * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001)

When using similar word results for free recall (See Table 7), which allows for more flexibility in recall accuracy, the differences are similar but smaller. There is only a significant difference for BMJ Journal. Texts still outperforms audio except for the WebMD.

Table 7: Similar word (%) for Free Recall. (A higher value means better retention).

	Number of similar words recalled (Strict Dataset)	
Test Sources (N)	**Audio**	**Text**
BMJ Lay Summary (193 texts)	10.8 (5)	10.62 (6)
WebMD (40 texts)	9.48 (6)	7.71 (6)
Patient instruction (40 texts)	5.87 (10)	6.86 (8)
Simple Wikipedia (243 texts)	10.54 (4)	12.13 (5)
BMJ Journal (200 texts) ***	3.47 (3)	7.91 (4)
Overall	8.32 (6)	10.00 (5)

(Significance, * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001)

4.4 Content Feature Analysis

We analyzed the content features of texts and the correlation with perceived and actual difficulty measures for both modalities. Tables 8 and 9 show the four groups of content features and their correlation with text or audio comprehension and retention.

For the text condition (Table 8), the first group of features, “Ordinaries” reveals that higher content word frequency correlates with lower perceived difficulty (-0.27) and better free recall (0.23 and 0.21). Grammar frequency has a smaller negative effect on perceived difficulty (-0.14) and a negligible impact on recall (0.02 and 0.013).

The second group of features, “Healthcare Domain Specialty,” shows that specificity increases perceived difficulty (0.13) while reducing free recall (-0.22 and -0.23). Ambiguity, concept density, and topic density increase perceived difficulty (0.31, 0.32, 0.33, respectively) and negatively impact free recall (-0.31, -0.32, -0.42, respectively), making densely packed and ambiguous texts difficult to process and remember.

The third group of features, “Parts-of-Speech Features,” indicates that a higher percentage of nouns correlates with greater perceived difficulty (0.20), while verbs reduce perceived difficulty (-0.28). Adverbs slightly decrease difficulty (-0.08), but none of these features significantly affect recall.

The fourth group of features, “Topic Spread,” shows that texts with more lexical chains are perceived as more difficult (0.21) and result in lower free recall (-0.26 and -0.27). Longer chains further increase perceived difficulty (0.10) and reduce recall (-0.34 and -0.36). Overlapping topics, indicated by lexical cross chains, follow the same pattern, increasing perceived difficulty (0.21) and decreasing recall (-0.26 and -0.27).

Table 8: Correlation of text features with the dependent variables for the strict dataset: information presented as text.

Features	**Perceived**	Actual Difficulty

	Difficulty	MC and TF	Free Recall	
			Percentage of Exact Matching Words	Percentage of Similar Words
Average Word Count	0.17****	0.01	-0.25****	-0.25****
Ordinariness				
Content Word Frequency	-0.27****	0.03	0.23****	0.21****
Grammar Frequency	-0.14****	0.04	0.02	0.013
Healthcare Domain Specialty (Averages)				
Specificity	0.13****	0.01	-0.22****	-0.23****
Ambiguity	0.31****	-0.01	-0.31****	-0.31****
Concept Density	0.32****	-0.01	-0.32****	-0.32****
Topic Density	0.33****	-0.03	-0.42****	-0.42****
Parts-Of-Speech Features (%)				
Nouns	0.20 ****	-0.04	0.01	0.01
Verbs	-0.28****	0.11****	0.05	0.02
Adverbs	-0.08**	-0.04	0.03	0.02
Adjectives	0.19****	0.02	-0.04	-0.03
Topic Spread (Averages)				
Lexical Chains	0.21****	-0.02	-0.26****	-0.27****
Lexical Chain Length	0.10***	0.06	-0.34****	-0.36****
Lexical Chain Span	0.17****	-0.07	-0.24****	-0.25****
Lexical Cross Chains	0.21****	-0.05	-0.26****	-0.27****

(Significance, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$)

Table 9 shows the results for the audio condition. Average word count shows a strong positive correlation with perceived difficulty, with correlations of 0.11 for MC and a negative correlation with free recall (-0.23 for exact word matching and -0.25 for similar word matching), indicating that shorter texts are perceived as easier.

When looking at the first group of features, the content word frequency has a significant negative correlation (-0.20) with perceived difficulty (more common words are seen as easier) and higher free recall. Content word frequency significantly also affects free recall results (0.14 and 0.13) with higher frequency leading to better recall. Grammar frequency slightly lowers perceived difficulty (-0.08) but has no significant effect on free recall.

Table 9: Correlation of text features with the dependent variables for strict dataset: information presented as audio.

Features	Perceived Difficulty	Actual Difficulty		
			Free Recall	
		MC and TF	Percentage of Exact Matching Words	Percentage of Similar Words
Average Word Count	0.11****	0.03	-0.23****	-0.25****
Ordinariness				
Content Word Frequency	-0.20****	0.01	0.14****	0.13****
Grammar Frequency	-0.08**	0.03	-0.01	-0.01
Healthcare Domain Specialty (Averages)				
Specificity	0.06*	0.02	-0.21****	-0.22****
Ambiguity	0.25****	0.08**	-0.30****	-0.32 ****

Concept Density	0.27****	0.07**	-0.30****	-0.33****
Topic Density	0.27****	0.07**	-0.39****	-0.42****
Parts-Of-Speech Features (%)				
Nouns	0.20****	0.11***	0.07*	0.04
Verbs	-0.29****	-0.11****	0.02	0.03
Adverbs	-0.13****	-0.03	0.01	0.01
Adjectives	0.18****	0.11****	-0.04	-0.04
Topic Spread (Averages)				
Lexical Chains	0.15****	0.06*	-0.25****	-0.27****
Lexical Chain Length	-0.03	0.01	-0.05	-0.06*
Lexical Chain Span	0.12****	0.03	-0.24****	-0.25****
Lexical Cross Chains	0.15****	0.06*	-0.25****	-0.27****

(Significance, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$)

The second group of features show that specificity, which represents the use of precise medical terms, increases perceived difficulty (0.06) and reduces recall performance (-0.21 and -0.22). Ambiguity mirrors this pattern, indicating that unclear or generalized terms also hinder comprehension and recall. Concept density and topic density, which measure the amount of information packed into an audio, correlate positively with perceived difficulty (0.27) and negatively impact recall, with values of -0.30 and -0.33 for concept density and -0.39 and -0.42 for topic density.

The third group of features shows that a higher percentage of nouns increases perceived difficulty (0.20), although the effect on recall is minimal (0.07 for exact word matching). Verbs, on the other hand, decrease the perceived difficulty of the audio (-0.29), though the impact on recall is negligible (0.02). Adjectives slightly increase perceived difficulty (0.18), while adverbs reduce it (-0.13), but neither significantly affects free recall.

The fourth group of features shows that audio with more lexical chains (indicating distinct topics) correlates with increased perceived difficulty (0.15) and reduced recall of information (-0.25 and -0.27). Lexical chain span, representing the extent of topic coverage, similarly increases perceived difficulty (0.12) and negatively impacts recall (-0.24 and -0.25). Lexical cross chains, which measure topic overlap, follow the same trend, correlating positively with perceived difficulty (0.15) and negatively with free recall, while slightly benefiting MC performance (0.06).

Overall, the findings suggest that the impact of text features on difficulty varies somewhat between text and audio delivery. Participants perceive dense, healthcare-specific content as more difficult in text form, but the effects on actual difficulty and recall are mixed across both modalities.

5. Follow up analysis

5.1 Education level

We further analyzed the results for participants with different education levels. We show the actual difficulty (MC and TF) results in Fig 1. The complete dataset is shown in Appendix H and Appendix I. Data was not always available for each group for this analysis. The most complete data was found for simple Wikipedia, BMJ Journal, and BMJ Lay summary.

Figure 1 presents average actual difficulty results by education level across the sources, with 1 being the lowest level (high school) and 5 the highest (doctorate). Overall, the data shows decreased

accuracy for higher educated participants across the corpora in both conditions. The higher-educated participant may consider the information already known and pay less attention to the content.

The BMJ Lay Summary text condition improves from 50% for levels 1-2 to 66.7% at level 3, peaking at 74.5% at higher levels, whereas its audio condition demonstrates low accuracy with

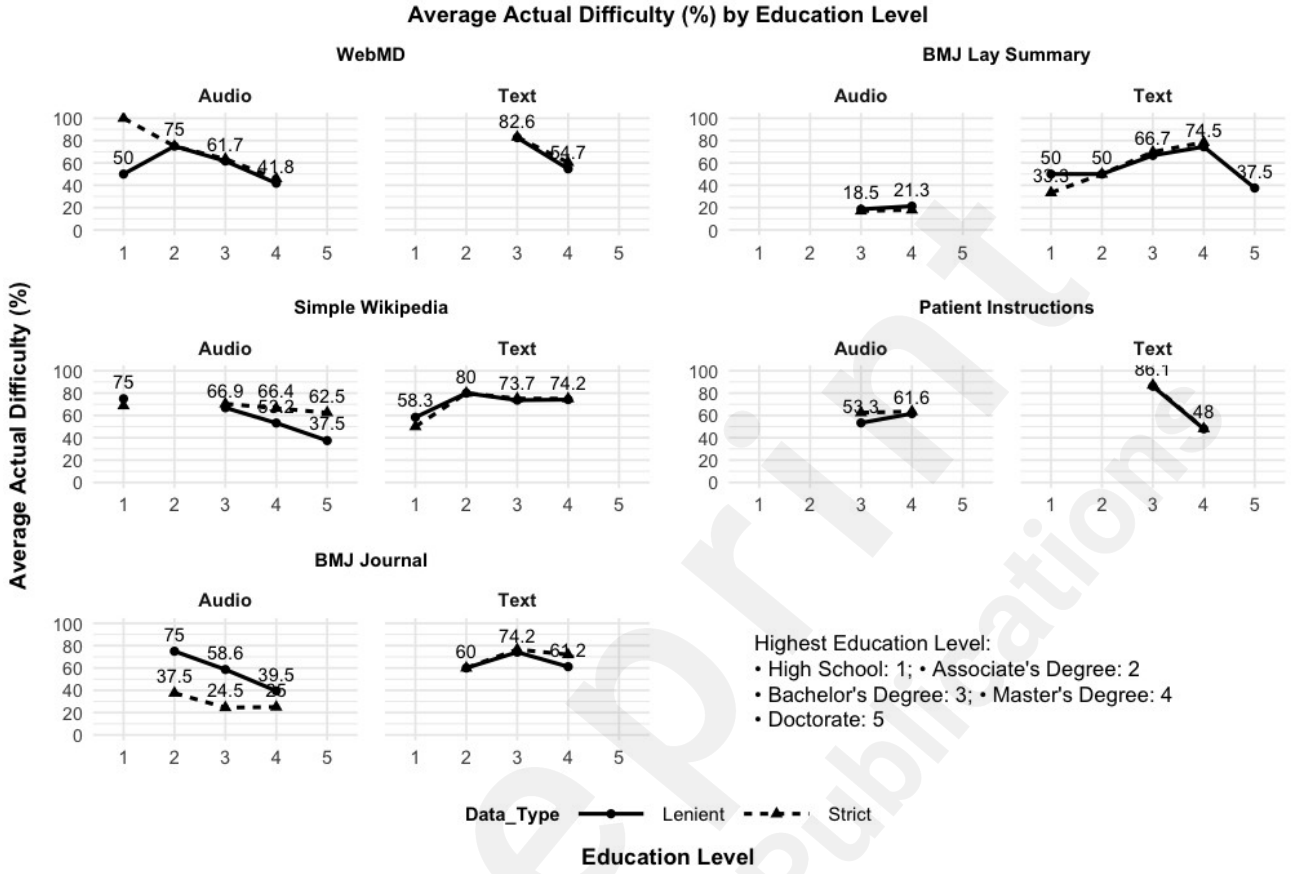

Fig. 1. Average actual difficulty (MC and TF) results by education level.

18.5% at level 2 and 21.3% at level 3. Simple Wikipedia's audio condition begins at 75% accuracy for level 1 but decreases to 66.9% at level 3 and 62.5% at level 5. The BMJ Journal's text format records 60% accuracy for level 1, while its audio format starts at 37.5% at level 1, dips to 24.5% at level 2, and stabilizes between 30-35% for levels 3-4.

5.2 English spoken at home

We conducted a parallel analysis for participants with different English language levels. Figure 2 presents the average actual difficulty results by English spoken at home across the sources, with 1 being never speaking English at home (Never English) to 5 the most frequent (Only English). The most complete data was found for simple Wikipedia, BMJ Journal, and BMJ Lay summary.

Overall, the data suggests varied accuracy patterns, with some showing increased accuracy for non-native English speakers and others showing more complex relationships between English speaking frequency at home and comprehension accuracy.

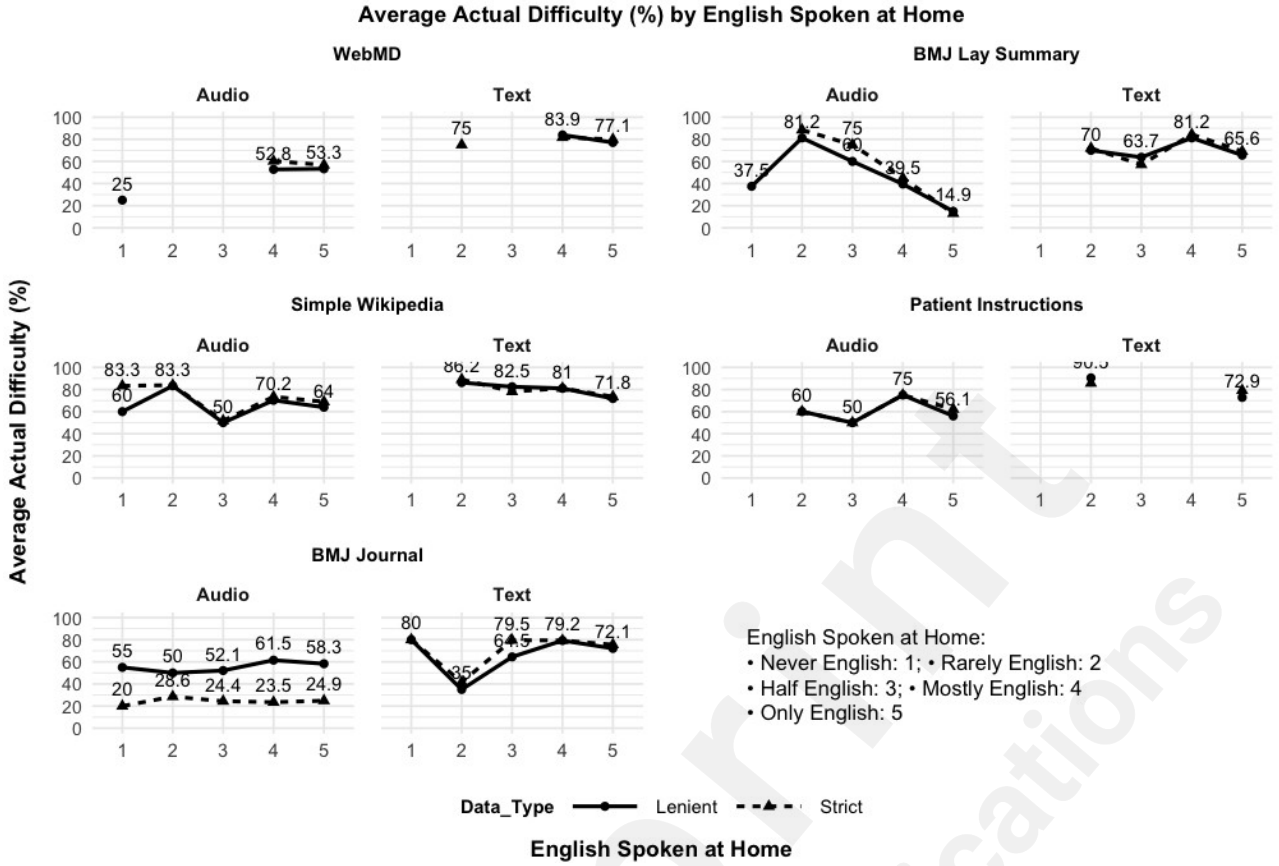

Fig. 2. Average actual difficulty (%) by English spoken at home. A higher value means better result.

The BMJ Lay Summary's text condition shows fluctuations, ranging from 63.7% to 81.2% across different levels. Its audio condition demonstrates a varied pattern, starting at 37.5% accuracy for level 1, peaking at 81.2% for level 2, and then steadily decreasing to 14.9% by level 5. The text condition for Simple Wikipedia remains relatively stable between 71.8% and 82.5% across all levels. And Simple Wikipedia's audio condition exhibits high accuracy for levels 1 and 2 at 83.3%. The BMJ Journal's text condition demonstrates high accuracy across all levels, ranging from 72.1% to 80% while audio condition shows relatively low accuracy, ranging from 20% to 28.6% for levels 1-3, with a notable increase to 61.5% and 58.3% for levels 4 and 5.

5.3 Readability formulas

We have analyzed the relationship between existing readability formulas, Flesch reading ease, Gunning Fog Index, Smog Index, Dale-Chall readability Score, and perceived and actual difficulty (Tables 10 and 11). We apply a Bonferroni correction since these were not a priori posed hypotheses. Results with p<=0.03 are considered significant after the correction.

For the text condition (Table 10), the readability formulas are aligned with the perceived difficulty results. The Flesch Reading Ease has a significant negative correlation (-0.292), indicating that as texts become simpler, they are perceived as less difficult, while the Smog Index shows a strong positive correlation (0.334), indicating that higher difficulty increases perceived difficulty. The Gunning Fog Index (0.201) and Dale-Chall Readability Score (0.196) also positively correlate with perceived difficulty but to a lesser degree.

Table 10. Correlation of readability metrics with the dependent variables for strict dataset: information presented as text.

Readability	Perceived Difficulty	Actual Difficulty		
		MC and TF	Free Recall	
			Percentage of Matching Words	Percentage of Similar Words
Flesch Reading Ease	-0.29 ****	-0.03	0.15 ****	0.13 ****
Gunning Fog Index	0.20 ****	0.05	-0.05	-0.03
Smog Index	0.33****	0.01	-0.34 ****	-0.34****
Dale Chall Readability Score	0.20 ****	0.02	0.016	0.05

(Significance, * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001)

For actual difficulty, the Smog Index provides significant results for free recall. There are no significant results for the correlation with MC and TF. The Flesch Reading Ease shows a significant positive correlation with free recall (0.15 and 0.126 of information. Conversely, the Smog Index has strong negative correlations with free recall (-0.336 and -0.337) of information.

For the audio condition (Table 11), we see only significant correlations for the Smog Index and free recall. It has strong negative correlations with free recall (-0.12 and -0.14).

Table 11. Correlation of readability metrics with the dependent variables for strict dataset: information presented as audio.

Readability	Perceived Difficulty	Actual Difficulty		
		MC and TF	Free Recall	
			Percentage of Matching Words	Percentage of Similar Words
Flesch Reading Ease	0.08	0.07	0.07	0.07
Gunning Fog Index	-0.09	-0.07	-0.06	-0.05
Smog Index	0.04	-0.05	-0.12 ***	-0.14 ****
Dale Chall Readability Score	-0.07	-0.05	-0.01	0.00

(Significance, * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001)

6. Discussion

Our study reveals significant differences in the effectiveness of text and audio modalities for delivering health information. Consistently, information comprehension was higher with text than audio. Better performance with text was particularly pronounced in difficult content sources such as the BMJ Journal. This suggests that text remains the more effective for ensuring accurate understanding and retention of intricate or difficult health information.

For easier content, e.g., sources such as WebMD, audio showed comparable or even slightly better performance in perceived difficulty and free recall tasks. This indicates that audio can be an effective medium for delivering more conversational content, possibly due to its similarity to natural speech patterns and potential to engage auditory learners more effectively.

A dichotomy emerged between perceived difficulty and actual difficulty results. While text was generally perceived as easier to understand, this perception did not always translate to better comprehension. This discrepancy highlights the importance of considering both subjective experiences and objective measures when evaluating the effectiveness of health communication methods.

The stronger negative correlation between content word frequency and perceived difficulty in the audio study suggests that a familiar vocabulary (high frequency words) might be essential for how

information is perceived and considered accessible [26]. The differential impact of parts of speech, particularly the facilitating effect of verbs on comprehension in both modalities (with a more substantial impact in audio), suggests that verb-rich content is easier to understand. This replicates findings in earlier work [32, 36]. The stronger correlation between healthcare-specific features (like concept density and topic density) and perceived difficulty in the text highlights the challenges of presenting specialized medical information in written form. Correlation analyses of existing readability formulas indicate that these tools are effective in estimating perceived difficulty for text-based materials but ineffective in assessing actual difficulty. Moreover, when applied to information presented in audio formats, readability formulas fail to reliably assess both perceived and actual difficulty.

7. Future Directions and Conclusion

Our study demonstrates the general superiority of text for conveying difficult health information compared to audio formats. Yet, it also reveals nuances that suggest opportunities for optimizing health information delivery through careful consideration of content complexity, target audience, and delivery modality is possible but requires more than readability formulas. While our study provides valuable insights, it also has limitations. Further exploration of education, prior health knowledge, and learning style preferences may influence the effectiveness of text versus audio modalities and could lead to more personalized health communication strategies. Additionally, while AMT workers reflect the general population, the results may vary if patients with a personal stake and prior knowledge of the information read the text or listen to the audio.

Acknowledgements

The research reported in this paper was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would also like to acknowledge the contributions of the AMT workers, who served as the primary participants in this study.

Authors Contributions

Arif Ahmed: Writing – original draft & editing, Validation, Methodology, Conceptualization, Data curation, formal analysis. **Gondy Leroy:** Writing – review & editing, Supervision, Funding acquisition, Validation, Methodology. **Prosanta Barai:** Methodology, Data curation, Formal analysis. **David Kauchak:** Writing – review, Validation, Methodology. **Philip Harber:** Writing – review, Validation, Methodology, Conceptualization. **Stephen A. Rains:** Writing – Methodology, review, Conceptualization.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Eichler, K., S. Wieser, and U. Brügger, *The costs of limited health literacy: a systematic review.* International journal of public health, 2009. **54**: p. 313-324.
2. Hart, T.L., et al., *Development of multimedia informational tools for breast cancer patients*

with low levels of health literacy. Patient education and counseling, 2015. **98**(3): p. 370-377.
3. Koh, H.K., et al., *New federal policy initiatives to boost health literacy can help the nation move beyond the cycle of costly 'crisis care'.* Health Affairs, 2012. **31**(2): p. 434-443.
4. digital.gov and plainlanguage.gov. *Plain Language Summit 2019: A Day of talks and conversations on how to advance plain language in government communications.* 2019 [cited 2023 April 29]; Available from: https://digital.gov/event/2019/09/05/plain-language-summit-2019/20198.
5. Grene, M., Y. Cleary, and A. Marcus-Quinn, *Use of plain-language guidelines to promote health literacy.* IEEE Transactions on Professional Communication, 2017. **60**(4): p. 384-400.
6. Choi, S., et al., *Bridging the gap in health literacy research: The inclusion of individuals with visual impairments.* Patient Education and Counseling, 2023. **116**: p. 107932.
7. Thormundsson, B. *U.S.: Voice assistant users 2026.* 2023, December 5 [cited 2024 October,13]; Available from: https://www.statista.com/statistics/1299985/voice-assistant-users-us/.
8. Leibler, S. *Cedars-Sinai Taps Alexa for Smart Hospital Room Pilot*. 2019 [cited 2023 April 29]; Available from: https://www.cedars-sinai.org/newsroom/cedars-sinai-taps-alexa-for-smart-hospital-room-pilot/2019.
9. Yoo, T.K., et al., *Deep learning-based smart speaker to confirm surgical sites for cataract surgeries: A pilot study.* PloS one, 2020. **15**(4): p. e0231322.
10. Modev Staff Writers. *Voice Tech in Healthcare: Transformation and Growth*. 2022 6 June 2023]; Available from: https://www.modev.com/blog/voice-tech-in-healthcare-transformation-and-growth#:~:text=The%20report%20tells%20us%20that,2019%20to%2021%25%20in%202021.
11. Sun, W., et al. *A deep learning based no-reference quality assessment model for ugc videos*. in *Proceedings of the 30th ACM International Conference on Multimedia*. 2022.
12. DuBay, W., *The principles of readability.* Impact Information, 2004.
13. Berkman, N.D., et al., *Low health literacy and health outcomes: an updated systematic review.* Annals of internal medicine, 2011. **155**(2): p. 97-107.
14. Mackert, M., et al., *Health literacy and health information technology adoption: the potential for a new digital divide.* Journal of medical Internet research, 2016. **18**(10): p. e264.
15. Nutbeam, D., *The evolving concept of health literacy.* Social science & medicine, 2008. **67**(12): p. 2072-2078.
16. Rudd, R.E., *The evolving concept of health literacy: new directions for health literacy studies*. 2015, Taylor & Francis. p. 7-9.
17. Mayer, G.G. and M. Michael Villaire, *Health literacy in primary care: A clinician's guide*. Vol. 130. 2007: Springer Publishing Company.
18. Badarudeen, S. and S. Sabharwal, *Assessing readability of patient education materials: current role in orthopaedics.* Clinical Orthopaedics and Related Research®, 2010. **468**(10): p. 2572-2580.
19. Readable. *Flesch reading ease and the Flesch Kincaid grade level.* . 2021 [cited 2023 April 29]; Available from: https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/
20. Leroy, G., et al. *Evaluating online health information: Beyond readability formulas*. in *AMIA Annual Symposium Proceedings*. 2008. American Medical Informatics Association.
21. Davidson, M.M., *Reading comprehension in school-age children with autism spectrum disorder: Examining the many components that may contribute.* Language, Speech, and Hearing Services in Schools, 2021. **52**(1): p. 181-196.
22. Kauchak, D., et al. *Text simplification tools: Using machine learning to discover features that identify difficult text*. in *2014 47th Hawaii international conference on system sciences*. 2014. IEEE.

23. Mukherjee, P., G. Leroy, and D. Kauchak, *Using lexical chains to identify text difficulty: a corpus statistics and classification study.* IEEE journal of biomedical and health informatics, 2018. **23**(5): p. 2164-2173.
24. Leroy, G., et al., *Text and Audio Simplification: Human vs. ChatGPT.* AMIA Summits on Translational Science Proceedings, 2024. **2024**: p. 295.
25. McNamara, D.S., et al., *Automated evaluation of text and discourse with Coh-Metrix*. 2014: Cambridge University Press.
26. Ahmed, A., et al., *Influence of Audio Speech Rate and Source Text Difficulty on Health Information Comprehension and Retention.* 2024.
27. Ahmed, A., et al., *Effects of Added Emphasis and Pause in Audio Delivery of Health Information.* AMIA Summits on Translational Science Proceedings, 2024. **2024**: p. 54.
28. Kang, O., R.I. Thomson, and M. Moran, *Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension.* Language Learning, 2018. **68**(1): p. 115-146.
29. Sulem, E., O. Abend, and A. Rappoport, *Semantic structural evaluation for text simplification.* arXiv preprint arXiv:1810.05022, 2018.
30. Brämer, G.R., *International statistical classification of diseases and related health problems. Tenth revision.* World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales, 1988. **41**(1): p. 32-36.
31. Leroy, G., et al., *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, 2013. **15**(7): p. e2569.
32. Leroy, G. and J.E. Endicott. *Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty*. in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. 2012.
33. Kauchak, D., G. Leroy, and A. Hogue, *Measuring text difficulty using parse-tree frequency.* Journal of the Association for Information Science and Technology, 2017. **68**(9): p. 2088-2100.
34. Leroy, G., et al., *Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty.* JAMIA open, 2022. **5**(2): p. ooac044.
35. Barai, P., et al., *Crowdsourcing with Enhanced Data Quality Assurance: An Efficient Approach to Mitigate Resource Scarcity Challenges in Training Large Language Models for Healthcare.* AMIA Summits on Translational Science Proceedings, 2024. **2024**: p. 75.
36. Ahmed, A., et al., *Audio delivery of health information: An NLP study of information difficulty and bias in listeners.* Procedia computer science, 2023. **219**: p. 1509-1517.

Supplementary Files