Readability of Pediatric Otolaryngology Information: Comparing AI-Generated Content with Google Search Results

Jonathan M. Carnino, Sanjeev Rampam, Elizabeth Puyo, Dean Kennedy, Jessica R. Levi

Department of Otolaryngology, Boston University Chobanian & Avedisian School of Medicine,
Boston Medical Center, Boston MA

Abstract

Objective: This study evaluates and compares the readability of pediatric otolaryngology patient education materials generated by ChatGPT-4o and those retrieved from Google searches. The goal is to determine whether AI-generated content improves accessibility compared to institutionally affiliated online resources.
Study Design: Cross-sectional readability analysis.
Setting: Online educational materials focused on pediatric otolaryngology topics.
Methods: Educational articles covering ten pediatric otolaryngology conditions were sourced either via Google search or generated using ChatGPT-4o. All texts were standardized by removing extraneous formatting. Readability was assessed using six validated metrics: Flesch-Kincaid Grade Level (FKGL), Flesch Reading Ease Score (FRES), Gunning-Fog Index, Simple Measure of Gobbledygook (SMOG), Coleman-Liau Index, and Automated Readability Index (ARI). Statistical comparisons were performed using paired t-tests or Wilcoxon signed-rank tests to evaluate differences in scores between sources.
Results: ChatGPT-4o-generated content demonstrated significantly higher FKGL, Gunning-Fog, ARI, and SMOG scores and lower FRES scores compared to Google-sourced materials, indicating greater complexity ($p < 0.05$). These differences were most pronounced for simpler conditions such as allergic rhinitis and otitis externa. For more complex topics like laryngomalacia and cleft lip and palate, readability scores were not significantly different between the two sources ($p > 0.05$).
Conclusion: ChatGPT-4o-generated patient education materials are generally more difficult to read than Google-sourced content, especially for less complex conditions. Given the importance of readability in patient education, AI-generated materials may require further refinement to improve accessibility without compromising accuracy. Enhancing clarity could increase the utility of AI tools for educating parents and caregivers in pediatric otolaryngology.

Introduction

- Patient education materials are essential for caregiver understanding and decision-making, but **most online resources exceed recommended readability levels**.
- The AMA and NIH recommend **a sixth-grade reading level** for health information, yet many institutional websites publish content above this threshold.
- Large language models (LLMs), such as ChatGPT, offer a novel way to generate or refine patient education materials, with the potential to **improve accessibility**.
- Prior studies show **LLMs can simplify complex medical information**, though concerns remain regarding accuracy, trustworthiness, and readability.
- Pediatric otolaryngology encompasses a range of conditions that vary in complexity, making it an important field to test AI-generated education content.
- This study compares the **readability of ChatGPT-4o–generated materials** with Google-sourced resources to assess whether AI provides a more accessible alternative.

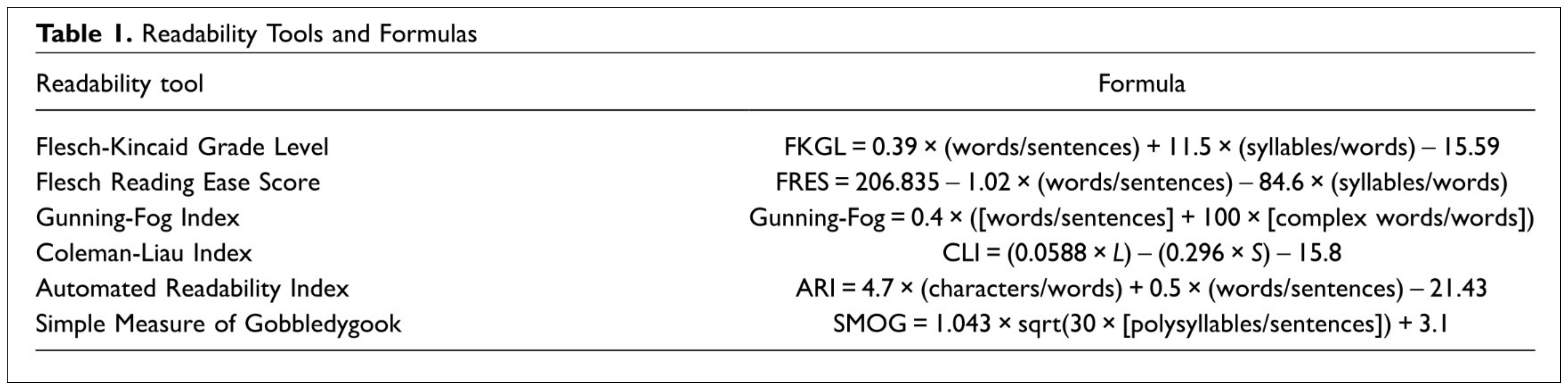

Table 1. Readability Tools and Formulas

Readability tool	Formula
Flesch-Kincaid Grade Level	FKGL = 0.39 × (words/sentences) + 11.5 × (syllables/words) – 15.59
Flesch Reading Ease Score	FRES = 206.835 – 1.02 × (words/sentences) – 84.6 × (syllables/words)
Gunning-Fog Index	Gunning-Fog = 0.4 × ([words/sentences] + 100 × [complex words/words])
Coleman-Liau Index	CLI = (0.0588 × *L*) – (0.296 × *S*) – 15.8
Automated Readability Index	ARI = 4.7 × (characters/words) + 0.5 × (words/sentences) – 21.43
Simple Measure of Gobbledygook	SMOG = 1.043 × sqrt(30 × [polysyllables/sentences]) + 3.1

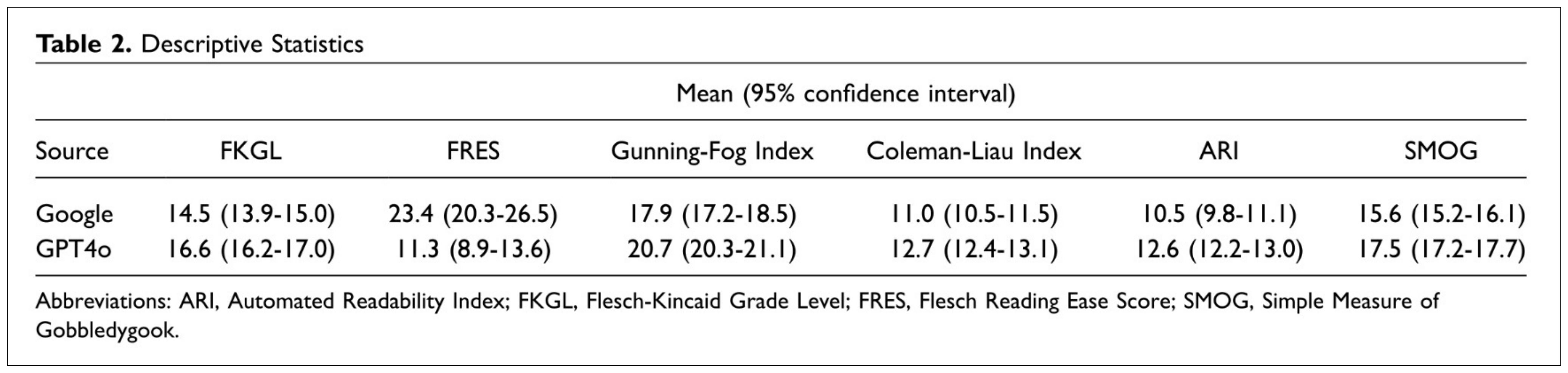

Table 2. Descriptive Statistics

	Mean (95% confidence interval)					
Source	FKGL	FRES	Gunning-Fog Index	Coleman-Liau Index	ARI	SMOG
Google	14.5 (13.9-15.0)	23.4 (20.3-26.5)	17.9 (17.2-18.5)	11.0 (10.5-11.5)	10.5 (9.8-11.1)	15.6 (15.2-16.1)
GPT4o	16.6 (16.2-17.0)	11.3 (8.9-13.6)	20.7 (20.3-21.1)	12.7 (12.4-13.1)	12.6 (12.2-13.0)	17.5 (17.2-17.7)

Abbreviations: ARI, Automated Readability Index; FKGL, Flesch-Kincaid Grade Level; FRES, Flesch Reading Ease Score; SMOG, Simple Measure of Gobbledygook.

Results

- Across all ten pediatric otolaryngology topics, ChatGPT-4o content had **higher grade-level readability scores** (FKGL, Gunning-Fog, ARI, SMOG) and **lower FRES scores**, indicating greater complexity than Google-sourced materials.
- For **simpler conditions** (e.g., allergic rhinitis, otitis externa, acute otitis media), ChatGPT-4o articles were significantly more difficult to read across multiple metrics ($p < 0.05$).
- For **moderately complex topics** (tonsillitis/adenoiditis, sinusitis), readability differences persisted, with ChatGPT-4o consistently requiring higher grade levels.
- For **more complex conditions** (laryngomalacia, choanal atresia, cleft lip/palate), readability scores were high for both sources, and differences were not statistically significant.
- Overall, AI-generated materials were **less accessible for common conditions**, while showing no readability advantage for complex topics.

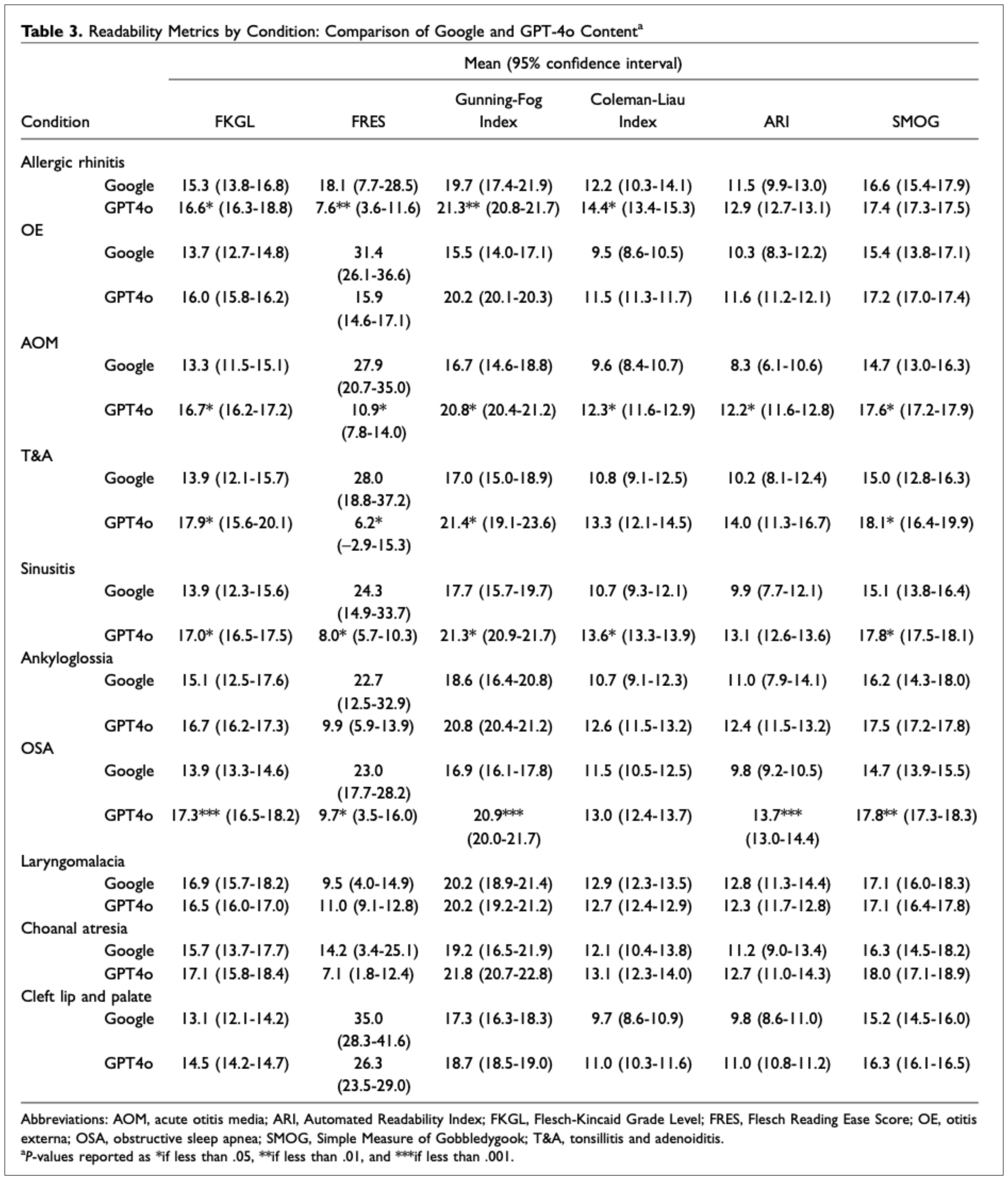

Table 3. Readability Metrics by Condition: Comparison of Google and GPT-4o Content[a]

		Mean (95% confidence interval)					
Condition		FKGL	FRES	Gunning-Fog Index	Coleman-Liau Index	ARI	SMOG
Allergic rhinitis							
	Google	15.3 (13.8-16.8)	18.1 (7.7-28.5)	19.7 (17.4-21.9)	12.2 (10.3-14.1)	11.5 (9.9-13.0)	16.6 (15.4-17.9)
	GPT4o	16.6* (16.3-18.8)	7.6** (3.6-11.6)	21.3** (20.8-21.7)	14.4* (13.4-15.3)	12.9 (12.7-13.1)	17.4 (17.3-17.5)
OE							
	Google	13.7 (12.7-14.8)	31.4 (26.1-36.6)	15.5 (14.0-17.1)	9.5 (8.6-10.5)	10.3 (8.3-12.2)	15.4 (13.8-17.1)
	GPT4o	16.0 (15.8-16.2)	15.9 (14.6-17.1)	20.2 (20.1-20.3)	11.5 (11.3-11.7)	11.6 (11.2-12.1)	17.2 (17.0-17.4)
AOM							
	Google	13.3 (11.5-15.1)	27.9 (20.7-35.0)	16.7 (14.6-18.8)	9.6 (8.4-10.7)	8.3 (6.1-10.6)	14.7 (13.0-16.3)
	GPT4o	16.7* (16.2-17.2)	10.9* (7.8-14.0)	20.8* (20.4-21.2)	12.3* (11.6-12.9)	12.2* (11.6-12.8)	17.6* (17.2-17.9)
T&A							
	Google	13.9 (12.1-15.7)	28.0 (18.8-37.2)	17.0 (15.0-18.9)	10.8 (9.1-12.5)	10.2 (8.1-12.4)	15.0 (12.8-16.3)
	GPT4o	17.9* (15.6-20.1)	6.2* (–2.9-15.3)	21.4* (19.1-23.6)	13.3 (12.1-14.5)	14.0 (11.3-16.7)	18.1* (16.4-19.9)
Sinusitis							
	Google	13.9 (12.3-15.6)	24.3 (14.9-33.7)	17.7 (15.7-19.7)	10.7 (9.3-12.1)	9.9 (7.7-12.1)	15.1 (13.8-16.4)
	GPT4o	17.0* (16.5-17.5)	8.0* (5.7-10.3)	21.3* (20.9-21.7)	13.6* (13.3-13.9)	13.1 (12.6-13.6)	17.8* (17.5-18.1)
Ankyloglossia							
	Google	15.1 (12.5-17.6)	22.7 (12.5-32.9)	18.6 (16.4-20.8)	10.7 (9.1-12.3)	11.0 (7.9-14.1)	16.2 (14.3-18.0)
	GPT4o	16.7 (16.2-17.3)	9.9 (5.9-13.9)	20.8 (20.4-21.2)	12.6 (11.5-13.2)	12.4 (11.5-13.2)	17.5 (17.2-17.8)
OSA							
	Google	13.9 (13.3-14.6)	23.0 (17.7-28.2)	16.9 (16.1-17.8)	11.5 (10.5-12.5)	9.8 (9.2-10.5)	14.7 (13.9-15.5)
	GPT4o	17.3*** (16.5-18.2)	9.7* (3.5-16.0)	20.9*** (20.0-21.7)	13.0 (12.4-13.7)	13.7*** (13.0-14.4)	17.8** (17.3-18.3)
Laryngomalacia							
	Google	16.9 (15.7-18.2)	9.5 (4.0-14.9)	20.2 (18.9-21.4)	12.9 (12.3-13.5)	12.8 (11.3-14.4)	17.1 (16.0-18.3)
	GPT4o	16.5 (16.0-17.0)	11.0 (9.1-12.8)	20.2 (19.2-21.2)	12.7 (12.4-12.9)	12.3 (11.7-12.8)	17.1 (16.4-17.8)
Choanal atresia							
	Google	15.7 (13.7-17.7)	14.2 (3.4-25.1)	19.2 (16.5-21.9)	12.1 (10.4-13.8)	11.2 (9.0-13.4)	16.3 (14.5-18.2)
	GPT4o	17.1 (15.8-18.4)	7.1 (1.8-12.4)	21.8 (20.7-22.8)	13.1 (12.3-14.0)	12.7 (11.0-14.3)	18.0 (17.1-18.9)
Cleft lip and palate							
	Google	13.1 (12.1-14.2)	35.0 (28.3-41.6)	17.3 (16.3-18.3)	9.7 (8.6-10.9)	9.8 (8.6-11.0)	15.2 (14.5-16.0)
	GPT4o	14.5 (14.2-14.7)	26.3 (23.5-29.0)	18.7 (18.5-19.0)	11.0 (10.3-11.6)	11.0 (10.8-11.2)	16.3 (16.1-16.5)

Abbreviations: AOM, acute otitis media; ARI, Automated Readability Index; FKGL, Flesch-Kincaid Grade Level; FRES, Flesch Reading Ease Score; OE, otitis externa; OSA, obstructive sleep apnea; SMOG, Simple Measure of Gobbledygook; T&A, tonsillitis and adenoiditis.

[a] *P*-values reported as *if less than .05, **if less than .01, and ***if less than .001.

Conclusions

- ChatGPT-4o generated pediatric otolaryngology education materials that were **more difficult to read** than institutionally affiliated Google sources, especially for simpler conditions.
- For **complex topics**, readability differences diminished, with both AI and Google content exceeding recommended literacy levels.
- These findings highlight a **missed opportunity**, as AI did not consistently simplify information to meet AMA/NIH sixth-grade readability standards.
- Increased complexity in AI outputs may stem from use of precise medical terminology and longer sentence structures, which **improve accuracy but reduce accessibility**.
- Optimizing AI tools through **prompt refinement, iterative revisions, and expert review** could enhance readability while maintaining medical accuracy.
- Future directions include evaluating **caregiver comprehension, language accessibility, and real-world integration of AI** into patient education workflows.

References

- Rooney MK, Santiago G, Perni S, et al. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. J Patient Exp. 2021;8:2374373521998847.
- Eltorai AE, Ghanian S, Adams CA Jr, Born CT, Daniels AH. Readability of patient education materials on the American Association for Surgery of Trauma website. Arch Trauma Res. 2014;3:e18161.
- Breneman A, Trager MH, Gordon ER, Samie FH. Readability rescue: large language models may improve readability of patient education materials. Arch Dermatol Res. 2024;316:669.
- Srinivasan N, Samaan JS, Rajeev ND, et al. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3.5, GPT-4, Bard, and online institutional resources. Surg Endosc. 2024;38:2522–2532.
- Dihan QA, Brown AD, Chauhan MZ, et al. Leveraging large language models to improve patient education on dry eye disease. Eye (Lond). 2024.
- Swisher AR, Wu AW, Liu GC, et al. Enhancing Health Literacy: Evaluating the Readability of Patient Handouts Revised by ChatGPT's Large Language Model. Otolaryngol Head Neck Surg. 2024;171:1751–1757.
- Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898.