Original Article

The Ability of Large Language Models to Generate Patient Information Materials for Retinopathy of Prematurity: Evaluation of Readability, Accuracy, and Comprehensiveness

Sevinç Arzu Postacı, Ali Dal

Mustafa Kemal University, Tayfur Sökmen Faculty of Medicine, Department of Ophthalmology, Hatay, Türkiye

Abstract

Objectives: This study compared the readability of patient education materials from the Turkish Ophthalmological Association (TOA) retinopathy of prematurity (ROP) guidelines with those generated by large language models (LLMs). The ability of GPT-4.0, GPT-4o mini, and Gemini to produce patient education materials was evaluated in terms of accuracy and comprehensiveness.

Materials and Methods: Thirty questions from the TOA ROP guidelines were posed to GPT-4.0, GPT-4o mini, and Gemini. Their responses were then reformulated using the prompts "Can you revise this text to be understandable at a 6th-grade reading level?" (P1 format) and "Can you make this text easier to understand?" (P2 format). The readability of the TOA ROP guidelines and the LLM-generated responses was analyzed using the Ateşman and Bezirci-Yılmaz formulas. Additionally, ROP specialists evaluated the comprehensiveness and accuracy of the responses.

Results: The TOA brochure was found to have a reading level above the 6th-grade level recommended in the literature. Materials generated by GPT-4.0 and Gemini had significantly greater readability than the TOA brochure ($p<0.05$). Adjustments made in the P1 and P2 formats improved readability for GPT-4.0, while no significant change was observed for GPT-4o mini and Gemini. GPT-4.0 had the highest scores for accuracy and comprehensiveness, while Gemini had the lowest.

Conclusion: GPT-4.0 appeared to have greater potential for generating more readable, accurate, and comprehensive patient education materials. However, when integrating LLMs into the healthcare field, regional medical differences and the accuracy of the provided information must be carefully assessed.

Keywords: Retinopathy of prematurity, large language models, readability, patient education

Cite this article as: Postacı SA, Dal A. The Ability of Large Language Models to Generate Patient Information Materials for Retinopathy of Prematurity: Evaluation of Readability, Accuracy, and Comprehensiveness. Turk J Ophthalmol. 2024;54:330-336

Address for Correspondence: Ali Dal, Mustafa Kemal University, Tayfur Sökmen Faculty of Medicine, Department of Ophthalmology, Hatay, Türkiye
E-mail: alidal19@hotmail.com **ORCID-ID:** orcid.org/0000-0002-0748-6416
Received: 09.09.2024 **Accepted:** 11.11.2024

DOI: 10.4274/tjo.galenos.2024.58295

Introduction

Retinopathy of prematurity (ROP) is a vasoproliferative and multifactorial disease of the retina. It is primarily observed in preterm infants but can also occur in full-term infants who have received high levels of oxygen therapy.[1] Advances in neonatal care have increased survival rates for preterm infants, which has resulted in more frequent encounters with conditions such as ROP. Annually, approximately 15 million babies worldwide are born prematurely (before 37 completed weeks of gestation).[2] Each year, between 23,800 and 45,600 infants are reported to suffer from irreversible vision loss as a result of ROP.[3] Particularly in low- and middle-income countries, up to 40% of childhood blindness is attributed to preventable ROP cases, and Türkiye is one of these countries.[4] A multicenter study conducted in Türkiye revealed that among 6,115 preterm infants, 27% were diagnosed with some stage of ROP, and 6.7% developed severe ROP.[5]

ROP can be effectively managed with consistent monitoring and prompt therapy.[6,7] Monitoring commences soon after delivery and continues until retinal vascularization is fully established. The follow-up frequency is modified according to the severity of the disease; infants with severe ROP are followed on a weekly basis, while others are seen at extended intervals. However, delays in follow-up might lead to lost treatment opportunities and ultimately result in complete blindness.[8]

The dissemination of comprehensive information regarding the disease and treatment process to families is of utmost importance, as it greatly enhances their compliance with follow-up and treatment. Previous research has demonstrated that increased levels of knowledge within families are correlated with less anxiety and improved adherence to treatment regimens.[9,10]

In Türkiye, the Turkish Ophthalmological Association (TOA) offers patient education resources and informed consent forms for a range of disorders on its official website. It is crucial to ensure that these materials are comprehensible to facilitate patients' information-gathering process.[11] Per the guidelines of the American Medical Association and the National Institutes of Health, patient education materials should be produced at a reading level equivalent to that of a 6th-grade student.[12] Various formulas which analyze factors such as sentence length and word structure are frequently employed to evaluate readability.[13] For Turkish texts, readability is commonly determined using the Ateşman[14] and Bezirci and Yılmaz[15] readability formulas.

Over the past few years, online information sources have emerged as readily available tools that patients often favor greatly. A survey conducted by the Pew Center reveals that 61% of persons in the United States actively access health information through internet platforms.[16] Nevertheless, it is widely recognized that the comprehensibility of online health information generally necessitates a greater degree of education.[17,18,19] Large language models (LLMs) are artificial intelligence systems trained using content available on the internet to generate texts in natural language.[20] Machine-learning models such as OpenAI's ChatGPT and Google's Gemini are being employed in the medical domain to provide patient education and create informative content.[21,22] Nevertheless, the dependability of these models is still a topic of contention, and further investigation is now being conducted.[23]

This research examined the readability levels of ROP patient education materials, structured in a question-and-answer format, available on the TOA website using the Ateşman and Bezirci-Yılmaz formulas. Thirty questions from these materials were posed to the advanced language models GPT-4.0, GPT-4o mini, and Gemini, and the responses were used to generate patient brochures. The readability, accuracy, and comprehensiveness of these brochures were then evaluated to assess the models' effectiveness in producing patient education materials.

Materials and Methods

The main data source for this study consisted of informational brochures created for families regarding the treatment guidelines for ROP, which can be obtained from the TOA website (https://www.todnet.org/tod-rehber/rop-tedavi-rehberi-2021.pdf, available in Turkish: Appendix 1: Informational Brochure for Families: Retinopathy of Prematurity Screening, Appendix 2: Informational Brochure for Families: Retinopathy of Prematurity Treatment).[24] The guidelines comprise 30 questions pertaining to ROP, such as "What is ROP?" and "How is ROP treated?", along with their accompanying responses. An independent analysis was conducted on each response from the guidelines using the Ateşman and Bezirci-Yılmaz readability formulas. Since our study used only publicly available data and literature and did not entail the use of any animal or human data, ethics committee approval and patient consent were not required.

Use of Large Language Models

In this study, 30 questions from the TOA ROP guidelines were posed to the ChatGPT-4.0, ChatGPT-4o mini, and Gemini models. Table 1 presents sample questions directed to the artificial intelligence tools used in this study. Each question was asked in a new chat session, and the responses were recorded. Additionally, the ability of LLMs to simplify texts for lower educational levels was evaluated. To assess this, the models were given their initial responses (initial format) with prompts to generate two new responses:[25]

Prompt 1: "Can you revise the following text to make it understandable at a 6th-grade reading level?" (P1 format).

Prompt 2: "Can you revise the following text to make it easier to understand?" (P2 format).

Each response was analyzed individually using the Ateşman and Bezirci-Yılmaz readability formulas.

Readability Criteria

Ateşman Readability Formula: The Ateşman formula provides a score between 0 and 100 based on average sentence and word length. We conducted the Ateşman analysis using an online program. The scoring system is categorized as follows: 90-100 points correspond to a 4th-grade level or below, 80-89 points to a 5th- or 6th-grade level, 70-79 points to a 7th- or 8th-grade level, 60-69 points to a 9th- or 10th-grade level, 50-59 points to an 11th- or 12th-grade level, 40-49 points to an associate-degree level, 30-39 points to an undergraduate-degree level, and 29 points or below to a postgraduate-degree level.[14]

Bezirci-Yılmaz Readability Formula: The Bezirci-Yılmaz formula evaluates readability based on average sentence length and the number of syllables in words. The Bezirci-Yılmaz analysis was conducted using a specialized software tool. The scoring system is as follows: 1-8 points correspond to the primary-school level, 9-12 points to the high-school level, and

Table 1. Sample questions directed to artificial intelligence tools in the study

Questions
What is ROP?
How common is ROP?
What is screening for ROP?
What causes ROP?
When should screening be done?
What happens during screening?
Is the examination painful?
What happens if my baby is sick when it's time for the eye exam?
What happens if ROP is found?
Will the screenings be finished before my baby goes home?

ROP: Retinopathy of prematurity

12-16 points to the undergraduate level; scores above 16 indicate readability appropriate for academic-level texts.[15]

Comprehensiveness and Accuracy of Patient-Targeted Information Produced by Large Language Models

The responses generated by LLMs were evaluated for comprehensiveness and accuracy based on the TOA ROP guidelines. Experts specialized in ROP and experienced in its clinical management (S.A.P. and A.D.) assessed the accuracy and comprehensiveness of the responses. The comprehensiveness of the answers was rated as follows:[26]

- 1 point: Insufficiently comprehensive (misses crucial information)
- 2 points: Somewhat comprehensive (contains minimal but necessary information)
- 3 points: Moderately comprehensive (provides a reasonable level of detail)
- 4 points: Comprehensive (includes critical information)
- 5 points: Very comprehensive (provides detailed and complete information)

The responses were evaluated for accuracy as follows:[27]

- 1 point: Poor (includes substantial inaccuracies and may be detrimental to patients)
- 2 points: Moderate (some inaccuracies but not likely to pose adverse effects for patients)
- 3 points: Excellent (free of errors)

Statistical Analysis

In the data analysis, one-way analysis of variance (ANOVA) was used for comparison of multiple means, followed by post-hoc Tukey's honestly significant difference test to identify significant pairwise differences. Statistical analyses were conducted using SPSS software (IBM SPSS Statistics, Version 26.0). A p value of <0.05 was considered statistically significant.

Results

Bezirci-Yılmaz Readability Scores

The Bezirci-Yılmaz readability analysis revealed that the texts initially produced by GPT-4.0 and Gemini had a notably lower reading level than those in the TOA brochure (p=0.010 and p=0.039, respectively). No statistically significant difference was found between the materials generated by GPT-4o mini and the TOA brochure (p=0.325). No statistically significant differences were found in the comparisons among the other groups (Table 2).

When comparing the initial responses of the LLMs (GPT-4.0, Gemini, and GPT-4o mini) with their responses in the P1 and P2 formats, a statistically significant increase in readability was observed only in the responses of GPT-4.0 (p=0.005 and p=0.012, respectively). No significant differences were found in the other groups. Additionally, no statistically significant differences were observed between the responses in the P1 and P2 formats within any of the LLM groups (p>0.05) (Table 3).

Ateşman Readability Scores

When examining the Ateşman readability scores, the initial responses generated by GPT-4.0 and Gemini were found to have significantly lower reading levels compared to the TOA brochure (p=0.016 and p=0.006, respectively). No significant difference was found between GPT-4o mini and the TOA brochure (p=0.910). Additionally, GPT-4.0 and Gemini showed significantly lower reading levels compared to GPT-4o mini (p=0.042 and p=0.035, respectively). However, no significant difference was observed between GPT-4.0 and Gemini (Table 2).

None of the LLMs' initial responses showed any statistically significant difference in Ateşman readability score when compared to their responses in the P1 and P2 formats. Furthermore, there were no notable disparities noted between the P1 and P2 formats for any of the models (Table 4). The reading level of the other LLMs groups was assessed to be at the 9th- to 10th-grade level, whereas the responses produced by GPT-4o mini were determined to be at the 11th- to 12th-grade level.

Comprehensiveness Scores

When comparing the comprehensiveness scores of the initial responses from the LLMs, the responses generated by GPT-4.0 were found to have a significantly higher level of comprehensiveness compared to those from GPT-4o mini

Table 2. Comparison of Bezirci-Yılmaz and Ateşman readability scores between the TOA brochure and LLM initial responses

	TOA	GPT-4.0	Gemini	GPT-4o mini	p value	
Bezirci-Yılmaz readability score, mean (SD)	12.30 (7.58)	8.30 (2.50)	9.17 (2.40)	10.72 (4.20)	TOA vs. GPT 4.0: TOA vs. Gemini: TOA vs. GPT 4o mini: GPT 4.0 vs. Gemini: GPT 4.0 vs. GPT 4o mini: Gemini vs. GPT 4o mini:	**0.010** **0.039** 0.325 0.838 0.209 0.525
Ateşman readability score, mean (SD)	51.57 (21.74)	62.06 (6.86)	63.61 (7.94)	51.07 (10.57)	TOA vs. GPT 4.0: TOA vs. Gemini: TOA vs. GPT 4o mini: GPT 4.0 vs. Gemini: GPT 4.0 vs. GPT 4o mini: Gemini vs. GPT 4o mini:	**0.016** **0.006** 0.910 0.682 **0.042** **0.035**

Significant results (p<0.05) shown in bold. TOA: Turkish Ophthalmological Association, LLM: Large language model, SD: Standard deviation

and Gemini (p=0.045 and p=0.001, respectively). However, no significant difference in comprehensiveness was observed between GPT-4o mini and Gemini. The comprehensiveness scores of GPT-4.0's responses in the P1 and P2 formats were higher than those of GPT-4o mini and Gemini (Table 5).

Accuracy Scores

When comparing the accuracy scores of the initial responses from the LLMs, GPT-4.0's accuracy scores were found to be statistically significantly higher than those of Gemini (p=0.001). However, no significant difference in accuracy was observed between GPT-4o mini and Gemini or GPT-4.0. When comparing the accuracy scores of responses in the P1 and P2 formats, GPT-4.0 was significantly more accurate than Gemini (p=0.039 and p=0.034, respectively). No other statistically significant differences were observed (Table 5).

Discussion

In this study, the readability of patient education materials in the TOA ROP treatment guidelines was assessed. According to the Bezirci-Yılmaz readability formula, the materials were at an average high-school level, whereas the Ateşman readability formula placed them at 11th or 12th grade. Research conducted in Türkiye revealed the average education level to be 6.51 years.[28] When creating patient education materials, it is important to consider the average education level of each country.[29] In the literature, the recommended reading level for patient education materials is often at the 6th-grade level.[12] Materials that exceed this level may be difficult to interpret for patient populations with limited health literacy, potentially reducing treatment adherence. Therefore, the reading level of the TOA ROP guidelines is higher than suggested for patient education materials, indicating that they should be simplified. A similar problem occurred with the materials produced by ChatGPT-4.0, ChatGPT-4o mini, and Gemini. The reading levels of these materials were determined to be above the recommended level, not aligned with the norms stated in the literature.[30,31]

Delays in the treatment of ROP can lead to irreversible vision loss as well as significant medicolegal issues for healthcare professionals.[32] The most common issue in malpractice cases related to ROP is the failure to perform timely screening or follow-up.[33] One of the main reasons for this is that families do not have sufficient knowledge about ROP and the screening process. Studies in the literature have shown that when parents

Table 3. Comparison of Bezirci-Yılmaz readability scores and education levels between the initial (IF), P1, and P2 format responses from GPT-4.0, Gemini, and GPT-4o mini

		Bezirci-Yılmaz readability score, mean (SD)	Education level	p value
GPT-4.0	IF	8.30 (2.50)	Primary school	IF vs. P1: **0.005** IF vs. P2: **0.012** P1 vs. P2: 0.974
	P1	7.04 (3.04)	Primary school	
	P2	6.74 (3.62)	Primary school	
Gemini	IF	9.17 (2.40)	High school	IF vs. P1: 0.970 IF vs. P2: 0.942 P1 vs. P2: 0.907
	P1	8.53 (1.58)	Primary school	
	P2	8.22 (1.46)	Primary school	
GPT-4o mini	IF	10.72 (4.20)	High school	IF vs. P1: 0.879 IF vs. P2: 0.971 P1 vs. P2: 0.990
	P1	9.78 (3.04)	High school	
	P2	10.16 (3.62)	High school	

Significant results (p<0.05) shown in bold. SD: Standard deviation

Table 4. Comparison of Ateşman readability scores and education levels between the initial (IF), P1, and P2 format responses from GPT-4.0, Gemini, and GPT-4o mini

		Ateşman readability score, mean (SD)	Education level	p value
GPT-4.0	IF	62.06 (6.86)	9th-10th grade	IF vs. P1: 0.256 IF vs. P2: 0.312 P1 vs. P2: 0.999
	P1	68.03 (7.56)	9th-10th grade	
	P2	67.65 (6.90)	9th-10th grade	
Gemini	IF	63.61 (7.94)	9th-10th grade	IF vs. P1: 0.484 IF vs. P2: 0.219 P1 vs. P2: 0.901
	P1	65.54 (6.65)	9th-10th grade	
	P2	67.84 (6.85)	9th-10th grade	
GPT-4o mini	IF	51.07 (10.57)	11th-12th grade	IF vs. P1: 0.904 IF vs. P2: 0.684 P1 vs. P2: 0.793
	P1	58.12 (9.52)	11th-12th grade	
	P2	56.02 (9.39)	11th-12th grade	

SD: Standard deviation

Table 5. Comparison of comprehensiveness and accuracy scores of GPT-4.0, Gemini, and GPT-4o mini

		GPT-4.0	Gemini	GPT-4o mini	p value	
Comprehensiveness score, mean (SD)	IF	3.83 (0.91)	2.80 (1.16)	2.83 (1.26)	GPT 4.0 vs. Gemini: GPT 4.0 vs. GPT 4o mini: GPT 4o mini vs. Gemini:	**0.001** **0.045** 0.078
	P1	3.57 (0.90)	2.57 (0.97)	2.70 (1.18)	GPT 4.0 vs. Gemini: GPT 4.0 vs. GPT 4o mini: GPT 4o mini vs. Gemini:	0.004 **0.002** 0.093
	P2	3.53 (0.90)	2.50 (1.01)	2.43 (1.14)	GPT 4.0 vs. Gemini: GPT 4.0 vs. GPT 4o mini: GPT 4o mini vs. Gemini:	**0.030** **0.013** 0.061
Accuracy score, mean (SD)	IF	2.90 (0.31)	2.10 (0.76)	2.50 (0.57)	GPT 4.0 vs. Gemini: GPT 4.0 vs. GPT 4o mini: GPT 4o mini vs. Gemini:	**0.001** 0.058 0.345
	P1	2.90 (0.31)	2.13 (0.73)	2.50 (0.57)	GPT 4.0 vs. Gemini: GPT 4.0 vs. GPT 4o mini: GPT 4o mini vs. Gemini:	**0.039** 0.159 0.397
	P2	2.90 (0.31)	2.13 (0.73)	2.50 (0.57)	GPT 4.0 vs. Gemini: GPT 4.0 vs. GPT 4o mini: GPT 4o mini vs. Gemini:	**0.034** 0.217 0.231

Significant results ($p<0.05$) shown in bold. SD: Standard deviation

are informed and made aware, adherence to treatment improves and their infants have better outcomes.[9,10] In one study, it was reported that the parents of very low birth weight infants, especially those with limited English proficiency and poor health literacy, were not adequately informed about ROP, which negatively impacted treatment.[34] The study showed that more than half of parents did not receive adequate information about their infant's ROP condition upon discharge. One reason for this information gap is that 1 in 10 adults in the United States has low health literacy.[2]

An analysis conducted in the domain of pediatric ophthalmology revealed that online patient education materials were suitable for an audience with an average educational attainment of 11.75±2.72 years.[34] Insufficient comprehensibility of this educational material may result in inadequate compliance with therapy among persons with limited health literacy. Hence, it is imperative to provide patient education materials that are easily understandable for individuals with lower knowledge levels. According to the data collected in our study, the TOA guidelines for ROP are written at an unacceptably high reading level. Therefore, it is necessary to enhance the comprehensibility of these materials.

In this study, when comparing the readability levels of the brochures generated by GPT-4.0, GPT-4o mini, and Gemini with the TOA brochure, GPT-4.0 and Gemini were found to have lower readability levels compared to the TOA brochure. Additionally, in the P1 and P2 formats, which were designed to improve comprehensibility, an increase in readability (as assessed by Bezirci-Yılmaz score) was observed for the brochure created by GPT-4.0, while no significant changes were observed for Gemini or GPT-4o mini. These findings are consistent with the literature.[27,35,36] In terms of readability, these findings indicate that GPT-4.0 may be a more appropriate choice for creating a Turkish ROP guide.

LLMs are developing as new and intriguing instruments in the healthcare sector. They show potential particularly in patient consultation, medical triage, and providing information. LLMs can enhance access to healthcare by answering common medical questions from patients and improving care for individuals in remote or underserved areas.[22,37] Furthermore, these models have been observed to take on administrative tasks, allowing healthcare professionals to dedicate more time to patient care.[38] However, the use of LLMs presents certain challenges. LLMs may provide inaccurate information, posing a risk to patients and their families, particularly in medical settings.[39] These models have limited capacity for self-checking their responses and correcting errors. Misleading or incomplete information could lead to medical errors, posing serious risks to patient safety.[40] In order to fully integrate LLMs into clinical practice, further improvements in validation processes and stricter oversight of these models are essential.

Patient education materials must not just be easy to read, they must also be thorough and accurate. In our study, we also looked at the accuracy and comprehensiveness of the LLM-generated brochures. The results showed that the GPT-4.0 materials were more complete than the GPT-4o mini and Gemini materials. In terms of accuracy, GPT-4.0 scored highest, while Gemini received the lowest scores. These data indicate that GPT-4.0 could be a more trustworthy model for creating patient education materials. Similarly, Pushpanathan et al.[26] found that GPT-4.0 outperformed both GPT-3.5 and Google Bard in terms of accuracy and comprehensiveness when answering complex ocular symptom queries, highlighting its potential in patient education. Antaki et al.[21] also reported that GPT-4.0 provided

more consistent and relevant medical information compared to other LLMs in ophthalmology, underscoring its utility in generating reliable educational materials.

Another concern about the medical information offered by LLMs is the possibility of geographic variations in the data. Screening criteria for ROP may differ by country.[2] While some criteria may not be met in developed nations, the risk of severe ROP is higher in less developed countries.[39] The TOA ROP guidelines recommend screening all newborns delivered before 34 weeks of gestation or weighing less than 1,700 grams.[5] GPT-4.0's response for this question ("infants born before 30 weeks or weighing less than 1,500 grams") was comparable to the screening criteria employed in the United Kingdom but not with the TOA standards for Türkiye.[41] This disparity may generate uncertainty among patient relatives, potentially leading to misinformation and lower adherence to therapy.

Study Limitations

One of the major limitations of our study is the variability in the performance of language models across different languages. In our study, we asked questions in Turkish and requested that the responses be provided in Turkish as well. Additionally, we asked the language models to produce responses that were more understandable than those from Turkish sources. However, since LLMs are typically trained on English data, they may not perform as effectively in languages like Turkish. This discrepancy can be attributed to differences in linguistic structures and the limited availability of Turkish datasets.[20] It has also been noted in the literature that LLMs tend to show reduced performance when generating medical information in less-represented languages, which can increase the risk of errors in clinical applications.[42] Furthermore, the questions were posed as they appear in the TOA brochure, without the additional context of being asked from the perspective of a user in Türkiye. As such, the potential impact of including a phrase like "I am asking for Türkiye" on the model's responses was not evaluated. Therefore, the use of these models in languages such as Turkish requires careful consideration and should be supported by validation processes conducted by local experts.

Conclusion

Educating patients and their families is critical in the management of ROP. The reading level of TOA patient information pamphlets was determined to be higher than the acceptable level. In terms of readability, comprehensiveness, and accuracy, GPT-4.0 brochures outperformed GPT-4o mini and Gemini brochures. While LLMs are a promising tool in healthcare, it has been discovered that some information may be misleading, and there is a risk of misdirection owing to geographical variations. As a result, the integration of LLMs into healthcare should be thoroughly tested and supported by relevant recommendations. It has been determined that the accuracy of information generated by LLMs, particularly essential medical information, must be carefully assessed.

Ethics

Ethics Committee Approval: Not required.
Informed Consent: Not required.

Declarations

Authorship Contributions

Concept: S.A.P., Design: A.D., Data Collection or Processing: S.A.P., Analysis or Interpretation: A.D., Literature Search: S.A.P., Writing: S.A.P., A.D.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Dammann O, Hartnett ME, Stahl A. Retinopathy of prematurity. Dev Med Child Neurol. 2023;65:625-631.
2. Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller AB, Narwal R, Adler A, Vera Garcia C, Rohde S, Say L, Lawn JE. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. Lancet. 2012;379:2162-2172.
3. Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. Pediatr Res. 2013;74(Suppl 1):35-49.
4. Quinn GE. Retinopathy of prematurity blindness worldwide: phenotypes in the third epidemic. Eye Brain. 2016;8:31-36.
5. Bas AY, Demirel N, Koc E, Ulubas Isik Di, Hirfanoglu IM, Tunc T. Incidence, risk factors and severity of retinopathy of prematurity in Turkey (TR-ROP study): a prospective, multicentre study in 69 neonatal intensive care units. Br J Ophthalmol. 2018;102:1711-1716.
6. Hartnett ME. Retinopathy of prematurity: evolving treatment with anti-vascular endothelial growth factor. Am J Ophthalmol. 2020;218:208-213.
7. Kong L, Fry M, Al-Samarraie M, Gilbert C, Steinkuller PG. An update on progress and the changing epidemiology of causes of childhood blindness worldwide. J AAPOS. 2012;16:501-507.
8. Dogra MR, Katoch D, Dogra M. An update on retinopathy of prematurity (ROP). Indian J Pediatr. 2017;84:930-936.
9. Salehnezhad A, Zendetalab H, Naser S, Voshni HB, Abrishami M, Astaneh MA, Sani BT, Moghadam ZE. The effect of education based on the health belief model in improving anxiety among mothers of infants with retinopathy of prematurity. J Educ Health Promot. 2022;11:424.
10. McCahon H, Chen V, Paz EF, Steger R, Alexander J, Williams K, Pharr C, Tutnauer J, Easter L, Levin MR. Improving follow-up rates by optimizing patient educational materials in retinopathy of prematurity. J AAPOS. 2023;27:134.
11. Papadakos C, Papadakos J, Catton P, Houston P, McKernan P, Friedman AJ. From theory to pamphlet: the 3Ws and an H process for the development of meaningful patient education resources. J Cancer Educ. 2014;29:304-310.
12. Weiss BD, Schwartzberg JG, Davis TC, Parker RM, Williams MV, Wang CC. Health literacy a manual for clinicians with contributions from. 2008. http://lib.ncfh.org/pdfs/6617.pdf
13. Crossley SA, Allen DB, Danielle McNamara JS. Text readability and intuitive simplification: a comparison of readability formulas. 2011;23:84-101.
14. Ateşman E. Türkçede okunabilirliğin ölçülmesi. Dil Dergisi. 1997;58:71-74.
15. Bezirci B, Yılmaz AE. A software library for measurement of readability of texts and a new readability metric for Turkısh. DEÜ Mühendislik Fakültesi Fen Bilimleri Dergisi. 2010;3;49-62.
16. The Social Life of Health Information. Pew Research Center. https://www.pewresearch.org/internet/2009/06/11/the-social-life-of-health-information/

17. Williams AM, Muir KW, Rosdahl JA. Readability of patient education materials in ophthalmology: a single-institution study and systematic review. BMC Ophthalmol. 2016;16:133.
18. Rouhi AD, Ghanem YK, Hoeltzel GD, Yi WS, Collins JL, Prout EP, Williams NN, Dumon KR. Quality and readability of online patient ınformation on adolescent bariatric surgery. Obes Surg. 2023;33:397-399.
19. Lee KC, Berg ET, Jazayeri HE, Chuang SK, Eisig SB. Online patient education materials for orthognathic surgery fail to meet readability and quality standards. J Oral Maxillofac Surg. 2019;77:180.
20. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930-1940.
21. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci. 2023;3:100324.
22. Song H, Xia Y, Luo Z, Liu H, Song Y, Zeng X, Li T, Zhong G, Li J, Chen M, Zhang G, Xiao B. Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. J Med Syst. 2023;47:125.
23. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, Berkowitz ST, Finn AP, Jahangir E, Scoville EA, Reese TS, Friedman DL, Bastarache JA, van der Heijden YF, Wright JJ, Ye F, Carter N, Alexander MR, Choe JH, Chastain CA, Zic JA, Horst SN, Turker I, Agarwal R, Osmundson E, Idrees K, Kiernan CM, Padmanabhan C, Bailey CE, Schlegel CE, Chambless LB, Gibson MK, Osterman TJ, Wheless LE, Johnson DB. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6:2336483.
24. Koç E, Yağmur A, Prof B, Özdek Ş, Ovalı F. Türk Neonatoloji Derneği, Türk Oftalmoloji Derneği, Türkiye Prematüre Retinopatisi Rehberi 2021. 2021. https://neonatology.org.tr/uploads/content/tan%C4%B1-tedavi/7_min_min.pdf
25. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. Ophthalmol Retina. 2024;8:195-201.
26. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, Wong WM, Wang X, Jin Tan MC, Chang Koh VT, Tham YC. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience. 2023;26:108163.
27. Srinivasan N, Samaan JS, Rajeev ND, Kanu MU, Yeo YH, Samakar K. 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.
28. Yeşilyurt ME, Karadeniz O, Gülel FE, Çağlar A, Kangallı Uyar GK. Mean and expected years of schooling for provinces in Turkey. PJESS. 2016;3:1-7.
29. Ay IE, Doğan M. An evaluation of the comprehensibility levels of ophthalmology surgical consent forms. Cureus. 2021;13:16639.
30. Yılmaz FH, Tutar MS, Arslan D, Çeri A. Readability, understandability, and quality of retinopathy of prematurity information on the web. Birth Defects Res. 2021;113:901-910.
31. Huang G, Fang CH, Agarwal N, Bhagat N, Eloy JA, Langer PD. Assessment of online patient education materials from major ophthalmologic associations. JAMA Ophthalmol. 2015;133:449-454.
32. Vinekar A, Gangwe A, Agarwal S, Kulkarni S, Azad R. Improving retinopathy of prematurity care: a medico-legal perspective. Asia Pac J Ophthalmol (Phila). 2021;10:437-441.
33. Moshfeghi DM. Top five legal pitfalls in retinopathy of prematurity. Curr Opin Ophthalmol. 2018;29:206-209.
34. John AM, John ES, Hansberry DR, Thomas PJ, Guo S. Analysis of online patient education materials in pediatric ophthalmology. J AAPOS. 2015;19:430-434.
35. Rouhi AD, Ghanem YK, Yolchieva L, Saleh Z, Joshi H, Moccia MC, Suarez-Pierre A, Han JJ. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? A pilot study. Cardiol Ther. 2024;13:137-147.
36. Lambert R, Choo ZY, Gradwohl K, Schroedl L, Ruiz De Luzuriaga A. Assessing the application of large language models in generating dermatologic patient education materials according to reading level: qualitative study. JMIR Dermatol. 2024;7:55898.
37. Srivastav S, Chandrakar R, Gupta S, Babhulkar V, Agrawal S, Jaiswal A, Prasad R, Wanjari MB. ChatGPT in radiology: the advantages and limitations of artificial intelligence for medical imaging diagnosis. Cureus. 2023;15:41435.
38. Loh E. ChatGPT and generative AI chatbots: challenges and opportunities for science, medicine and medical leaders. BMJ Lead. 2023;000797.
39. Karakas C, Brock D, Lakhotia A. Leveraging ChatGPT in the pediatric neurology clinic: practical considerations for use to improve efficiency and outcomes. Pediatr Neurol. 2023;148:157-163.
40. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine. 2023;90:104512.
41. Fierson WM; American Academy of Pediatrics Section on Ophthalmology; American Academy of Ophthalmology; American Association for Pediatric Ophthalmology and Strabismus; American Association of Certified Orthoptists. Screening examination of premature infants for retinopathy of prematurity. Pediatrics. 2018;142:e20183061. Erratum in: Pediatrics. 2019;143:e20183810.
42. Ahn S. The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions. Korean J Physiol Pharmacol. 2024;28:393-401.