Özgün Araştırma / Original Article

Büyük Dil Modellerinin Prematüre Retinopatisi için Hasta Bilgilendirme Materyali Üretme Yeteneği: Okunabilirlik, Doğruluk ve Kapsamlılık Değerlendirmesi

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 Üniversitesi, Tayfur Ata Sökmen Tıp Fakültesi, Göz Hastalıkları Anabilim Dalı, Hatay, Türkiye

Öz

Amaç: Bu çalışmada, Türk Oftalmoloji Derneği (TOD) prematüre retinopatisi (ROP) rehberindeki hasta bilgilendirme materyallerinin okunabilirlik düzeyi, büyük dil modelleri (BDM) tarafından üretilen metinlerle karşılaştırıldı. GPT-4.0, GPT-4o mini ve Gemini'nin hasta eğitim materyalleri üretme becerileri, doğruluk ve kapsamlılık açısından değerlendirildi.

Gereç ve Yöntem: Çalışmada, TOD ROP rehberinde yer alan 30 soru GPT-4.0, GPT-4o mini ve Gemini'ye yöneltildi. BDM'lerin yanıtları, "Bu metni 6. sınıf eğitim seviyesine uygun şekilde düzenler misin?" (S1 formatı) ve "Bu metni daha anlaşılır hale getirir misin?" (S2 formatı) şeklinde yeniden yöneltildi. TOD ROP rehberinin ve yanıtların okunabilirliği Ateşman ve Bezirci-Yılmaz formülleriyle analiz edildi. Ayrıca yanıtlar, kapsamlılık ve doğruluk açısından ROP uzmanları tarafından değerlendirildi.

Bulgular: TOD broşürünün okuma düzeyi, literatürde önerilen seviye olan 6. sınıf okuma düzeyinin üzerinde bulundu. GPT-4.0 ve Gemini'nin ürettiği materyallerin okuma düzeyleri ise TOD broşürüne kıyasla anlamlı olarak daha düşüktü ($p<0,05$). S1 ve S2 formatlarıyla yapılan düzenlemeler, GPT-4.0'ın okuma düzeyini düşürürken, GPT-4o mini ve Gemini'de anlamlı bir fark gözlenmedi. Doğruluk ve kapsam açısından GPT-4.0 en yüksek, Gemini ise en düşük puanları aldı.

Sonuç: GPT-4.0, hasta bilgilendirme materyalleri üretiminde daha okunabilir, doğru ve kapsamlı içerikler sunma potansiyeline sahip bir model olarak öne çıkmıştır. Ancak, BDM'lerin sağlık alanında entegrasyonu yapılırken, bölgesel tıbbi farklılıklar ve verilen bilgilerin doğruluğu dikkatle değerlendirilmelidir.

Anahtar Kelimeler: Prematüre retinopatisi, büyük dil modelleri, okunabilirlik, hasta eğitimi

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

Yazışma Adresi/Address for Correspondence: Ali Dal, Mustafa Kemal Üniversitesi, Tayfur Ata Sökmen Tıp Fakültesi, Göz Hastalıkları Anabilim Dalı, Hatay, Türkiye

E-posta: alidal19@hotmail.com **ORCID-ID:** orcid.org/0000-0002-0748-6416

Geliş Tarihi/Received: 09.09.2024 **Kabul Tarihi/Accepted:** 11.11.2024

DOI: 10.4274/tjo.galenos.2024.58295

Giriş

Prematüre retinopatisi (ROP), retinanın vazoproliferatif ve multifaktöriyel bir hastalığıdır. Öncelikle preterm bebeklerde görülür, ancak yüksek düzeyde oksijen tedavisi alan term bebeklerde de ortaya çıkabilir.[1] Yenidoğan bakımındaki ilerlemeler, preterm bebeklerin sağkalım oranlarını artırmış ve bu da ROP ile daha sık karşılaşılmasına neden olmuştur. Her yıl, dünya çapında yaklaşık 15 milyon bebek erken doğmaktadır. (gebeliğin 37 haftası tamamlanmadan önce).[2] Her yıl 23.800 ila 45.600 bebekte ROP sonucu geri dönüşü olmayan görme

©Telif Hakkı 2024 Türk Oftalmoloji Derneği / Türk Oftalmoloji Dergisi, Galenos Yayınevi tarafından yayınlanmıştır.
Creative Commons Atıf-GayriTicari-Türetilemez 4.0 Uluslararası (CC BY-NC-ND 4.0) lisansı altında lisanslanmıştır.

kaybı geliştiği bildirilmektedir.[3] Özellikle düşük ve orta gelirli ülkelerde, çocukluk çağı körlüğünün %40'a varan bölümünün önlenebilir ROP olgularına bağlı olduğu ileri sürülmektedir ve Türkiye bu ülkelerden biridir.[4] Türkiye'de yapılan çok merkezli bir çalışmada, 6.115 preterm bebeğin %27'sine farklı evrelerde ROP tanısı konduğu ve %6,7'sinde şiddetli ROP geliştiği tespit edilmiştir.[5]

ROP, sürekli izlem ve hızlı tedavi ile etkili şekilde yönetilebilir.[6,7] İzlem doğumdan hemen sonra başlar ve retinal vaskülarizasyon tam olarak sağlanana kadar devam eder. İzlem sıklığı hastalığın şiddetine göre değişir; ROP'lu bebekler haftalık olarak takip edilirken, diğer bebekler uzun aralıklarla görülür. Bununla birlikte, izlemdeki aksamalar tedavi fırsatlarının kaybolmasına ve sonuçta tam körlüğe neden olabilir.[8] Hastalık ve tedavi süreci ile ilgili ailelere detaylı bilgi verilmesi, takip ve tedaviye uyumlarını büyük ölçüde artırdığı için son derece önemlidir. Önceki araştırmalar, ailelerin bilgi düzeylerinin artmasının kaygıyı azalttığını ve tedavi rejimlerine uyumu artırdığını göstermiştir.[9,10]

Türkiye'de, Türk Oftalmoloji Derneği'nin (TOD), resmi web sitesinde bir dizi hastalık için hasta eğitim kaynakları ve bilgilendirilmiş onam formları yer almaktadır. Hastaların bilgi edinme sürecini kolaylaştırmak için bu materyallerin anlaşılır olmasını sağlamak çok önemlidir.[11] Amerikan Tıp Derneği ve Ulusal Sağlık Enstitüleri'nin yönergelerine göre, hasta eğitim materyalleri 6. sınıf öğrencilerinin okuma düzeyine eşdeğer olacak şekilde hazırlanmalıdır.[12] Okunabilirliği değerlendirmek için cümle uzunluğu ve kelime yapısı gibi faktörleri analiz eden çeşitli formüller sıklıkla kullanılmaktadır.[13] Türkçe metinler için okunabilirlik genellikle Ateşman[14] ve Bezirci-Yılmaz'ın[15] okunabilirlik formülleri kullanılarak belirlenir.

Son yıllarda, çevrimiçi bilgi kaynakları, hastaların büyük ölçüde tercih ettiği hazır araçlar olarak ortaya çıkmıştır. Pew Center tarafından yapılan bir anket, Amerika Birleşik Devletleri'nde halkın %61'inin sağlık bilgilerine internet platformları aracılığıyla aktif olarak eriştiğini göstermiştir.[16] Bununla birlikte, çevrimiçi sağlık ile ilgili bilgilerin anlaşılabilmesi için genellikle eğitim düzeyinin daha yüksek olması gerektiği yaygın olarak kabul edilmektedir.[17,18,19] Büyük dil modelleri (BDM), doğal dilde metinler oluşturmak için internette bulunan içeriği kullanan eğitilmiş yapay zeka sistemleridir.[20] OpenAI tarafından geliştirilen ChatGPT ve Google tarafından geliştirilen Gemini gibi makine öğrenimi modelleri, hastaların eğitimi ve bilgilendirici içerik oluşturmak için tıp alanında kullanılmaktadır.[21,22] Ancak, bu modellerin güvenilirliği hala bir tartışma konusudur ve bu konu hakkında çalışmalar devam etmektedir.[23]

Bu çalışmada Ateşman ve Bezirci-Yılmaz formülleri kullanılarak TOD web sitesinde yer alan soru-cevap şekilde yapılandırılmış ROP hasta eğitim materyallerinin okunabilirliği değerlendirilmiştir. Bu materyallerden ileri düzey dil modelleri GPT-4.0, GPT-4o mini ve Gemini'ye otuz soru yöneltilmiş ve yanıtlar hastalar için broşür oluşturmak için kullanılmıştır. Bu broşürlerin okunabilirliği, doğruluğu ve kapsamı daha sonra modellerin hasta eğitim materyalleri üretmedeki etkinliğini araştırmak için değerlendirilmiştir.

Gereç ve Yöntem

TOD web sitesinden elde edilebilen, aileler için hazırlanmış olan ROP tedavi kılavuzları hakkındaki bilgilendirme broşürleri bu çalışmanın ana veri kaynağıdır. (https://www.todnet.org/tod-rehber/rop-tedavi-rehberi-2021.pdf, Türkçe olarak mevcuttur: Ek 1: Aileler için Bilgilendirme Broşürü: Prematüre Retinopatisi Taraması, Ek 2: Aileler için Bilgilendirme Broşürü: Prematüre Retinopatisi Tedavisi).[24] Kılavuzlar, "ROP nedir?" ve "ROP nasıl tedavi edilir?" gibi ROP ile ilgili 30 soru ve yanıtlarından oluşmaktadır. Ateşman ve Bezirci-Yılmaz okunabilirlik formülleri kullanılarak kılavuzlardaki her yanıt için bağımsız bir analiz yapılmıştır. Çalışmamızda sadece kamuya açık veriler ve literatür kullanıldığından ve herhangi bir hayvan veya insan verisinin kullanılmasını gerektirmediğinden, etik kurul onayı ve hasta onamı gerekmemiştir.

Büyük Dil Modelleri Kullanımı

Bu çalışmada, TOD ROP kılavuzlarından 30 soru ChatGPT-4.0, ChatGPT-4o mini ve Gemini modellerine yöneltildi. Çalışmada kullanılan yapay zeka araçlarına yönelik örnek sorular Tablo 1'de sunulmuştur. Her soru yeni bir sohbet oturumunda soruldu ve cevaplar kaydedildi. Ayrıca, BDM'lerin daha düşük eğitim seviyesi için metinleri basitleştirme yeteneği değerlendirildi. Bunu değerlendirmek için modellere ilk yanıtları (ilk format) verildi ve iki yeni yanıt üretmeleri istendi:[25]

Soru 1: "Aşağıdaki metni 6. sınıf okuma düzeyine getirmek için revize edebilir misiniz?" (S1 formatı).

Soru 2: "Anlaşılmasını kolaylaştırmak için aşağıdaki metni revize edebilir misiniz?' (S2 formatı).

Her yanıt Ateşman ve Bezirci-Yılmaz okunabilirlik formülleri kullanılarak ayrı ayrı analiz edildi.

Okunabilirlik Kriterleri

Ateşman Okunabilirlik Formülü: Ateşman formülü ortalama cümle ve kelime uzunluğuna göre 0 ile 100 arasında bir puan vermektedir. Ateşman analizini çevrimiçi bir program

Tablo 1. Çalışmada yapay zeka araçlarına yöneltilen örnek sorular

Sorular
ROP nedir?
ROP ne kadar yaygındır?
ROP için tarama nedir?
ROP neden oluşur?
Tarama ne zaman yapılmalıdır?
Tarama sırasında neler olur?
Muayene ağrılı mıdır?
Göz muayenesi zamanı geldiğinde bebeğim hastaysa ne olur?
ROP bulunursa ne olur?
Bebeğim eve gitmeden önce taramalar bitecek mi?

ROP: Prematüre retinopatisi

kullanarak gerçekleştirdi. Skorlama sistemi aşağıdaki gibi kategorize edilmiştir: 90-100 puan 4. sınıf veya altı, 80-89 puan 5. veya 6. sınıf düzeyi, 70-79 puan 7. veya 8. sınıf düzeyi, 60-69 puan 9. veya 10. sınıf düzeyi, 50-59 puan 11. veya 12. sınıf düzeyi, 40-49 puan ön lisans düzeyi, 30-39 puan lisans düzeyi ve 29 puan veya altı lisansüstü düzeyine karşılık gelmektedir.[14]

Bezirci-Yılmaz Okunabilirlik Formülü: Bezirci-Yılmaz formülü okunabilirliği ortalama cümle uzunluğu ve sözcüklerdeki hece sayısına göre değerlendirir. Bezirci-Yılmaz analizi özel bir yazılım kullanılarak yapıldı. Puanlama sistemi aşağıdaki gibidir: 1-8 puan ilkokul düzeyine, 9-12 puan lise düzeyine, 12-16 puan lisans düzeyine karşılık gelir; 16'nın üzerindeki puanlar akademik düzeydeki metinlerin okunabilirliğini göstermektedir.[15]

Büyük Dil Modelleri Tarafından Üretilen Hasta Hedefli Bilgilerin Kapsamı ve Doğruluğu

BDM'ler tarafından oluşturulan yanıtlar, TOD ROP kılavuzları temel alınarak kapsam ve doğruluk açısından değerlendirildi. ROP konusunda uzman ve hastalığın klinik yönetiminde deneyimli uzmanlar (S.A.P. ve A.D.) yanıtların doğruluğunu ve kapsamını değerlendirdi. Cevapların kapsamı şu şekilde derecelendirildi:[26]

- 1 puan: Kapsamı yetersiz (önemli bilgiler eksik)
- 2 puan: Biraz kapsamlı (asgari düzeyde ancak gerekli bilgiler mevcuttur)
- 3 puan: Orta derecede kapsamlı (makul düzeyde ayrıntı mevcuttur)
- 4 puan: Kapsamlı (kritik bilgileri içerir)
- 5 puan: Çok kapsamlı (ayrıntılı ve eksiksiz bilgi mevcuttur)

Yanıtların doğruluğu şu şekilde değerlendirildi:[27]

- 1 puan: Zayıf (önemli yanlışlar vardır ve hastalar için zararlı olabilir)
- 2 puan: Orta (bazı yanlışlar mevcuttur, ancak hastalar için olumsuz etkiler oluşturması olası değildir)
- 3 puan: Mükemmel (yanlış yoktur)

İstatistiksel Analiz

Veri analizinde, ortalamaların karşılaştırılmasında tek yönlü varyans analizi (ANOVA) kullanıldı ve ardından anlamlı ikili farklılıkları belirlemek için post-hoc Tukey gerçekten anlamlı farklar testinden yararlanıldı. İstatistiksel analizler SPSS yazılımı (IBM SPSS Statistics, sürüm 26.0) kullanılarak yapıldı. İstatistiksel açıdan p değerinin <0,05 olması anlamlı kabul edildi.

Bulgular

Bezirci-Yılmaz Okunabilirlik Skorları

Bezirci-Yılmaz okunabilirlik analizi, GPT-4.0 ve Gemini tarafından ilk üretilen yanıtların okuma düzeyinin TOD broşüründen anlamlı düzeyde düşük olduğunu gösterdi (sırasıyla p=0,010 ve p=0,039). GPT-4o mini ile oluşturulan materyaller ile TOD broşürü arasında istatistiksel olarak anlamlı fark bulunmadı (p=0,325). Diğer gruplar arasında yapılan karşılaştırmalarda istatistiksel olarak anlamlı bir fark yoktu (Tablo 2).

BDM'lerin (GPT-4.0, Gemini ve GPT-4o mini) ilk yanıtları ile S1 ve S2 formatlarındaki yanıtları karşılaştırıldığında, yalnızca GPT-4.0 yanıtlarında okunabilirlikte istatistiksel olarak anlamlı bir artış gözlenmiştir (sırasıyla p=0,005 ve p=0,012). Diğer gruplarda anlamlı fark bulunmadı. Ayrıca, BDM gruplarının hiçbirinde S1 ve S2 formatlarındaki yanıtlar arasında istatistiksel olarak anlamlı bir fark gözlenmedi (p>0,05) (Tablo 3).

Ateşman Okunabilirlik Skorları

Ateşman okunabilirlik puanları incelendiğinde, GPT-4.0 ve Gemini tarafından oluşturulan ilk yanıtların okuma düzeyinin, TOD broşürüne kıyasla anlamlı derecede daha düşük olduğu bulundu (sırasıyla p=0,016 ve p=0,006). GPT-4o mini ile TOD broşürü arasında anlamlı fark saptanmadı (p=0,910). Ayrıca, GPT-4.0 ve Gemini'nin yanıtlarının okuma düzeyi, GPT-4o mini'ye kıyasla anlamlı düzeyde düşüktü (sırasıyla p=0,042 ve p=0,035). Ancak GPT-4.0 ile Gemini arasında anlamlı bir fark yoktu (Tablo 2).

BDM'lerin ilk yanıtlarının hiçbiri, Ateşman okunabilirlik skorunda S1 ve S2 formatlarındaki yanıtlarıyla karşılaştırıldığında istatistiksel olarak anlamlı bir fark göstermedi. Ayrıca, modellerin hiçbirinde S1 ve S2 formatları arasında dikkati çeken bir fark yoktu (Tablo 4). GPT-4o mini tarafından üretilen

Tablo 2. TOD broşürü ve BDM'lerin ilk yanıtlarının Bezirci-Yılmaz ve Ateşman okunabilirlik skorlarının karşılaştırılması

	TOD	GPT-4.0	Gemini	GPT-4o mini	p değeri	
Bezirci-Yılmaz okunabilirlik skoru, Ortalama (SD)	12,30 (7,58)	8,30 (2,50)	9,17 (2,40)	10,72 (4,20)	TOD ve GPT-4.0: TOD ve Gemini: TOD ve GPT-4o mini: GPT-4.0 ve Gemini: GPT-4.0 ve GPT-4o mini: Gemini ve GPT-4o mini:	**0,010** **0,039** 0,325 0,838 0,209 0,525
Ateşman okunabilirlik skoru, ortalama (SD)	51,57 (21,74)	62,06 (6,86)	63,61 (7,94)	51,07 (10,57)	TOD ve GPT-4.0: TOD ve Gemini: TOD ve GPT-4o mini: GPT-4.0 ve Gemini: GPT-4.0 ve GPT-4o mini: Gemini ve GPT-4o mini:	**0,016** **0,006** **0,910** **0,682** **0,042** **0,035**

Anlamlı sonuçlar (p<0,05) koyu renkle gösterilmiştir. TOD: Türk Oftalmoloji Derneği, BDM: Büyük dil modeli, SD: Standart deviasyon

yanıtların okuma düzeyi 11.-12. sınıf seviyesindeyken diğer BDM gruplarının okuma düzeyleri 9.-10. sınıf seviyesindeydi.

Kapsamlılık Skorları

BDM'lerden gelen ilk yanıtların kapsamlılık skorları karşılaştırıldığında, GPT-4.0 tarafından üretilen yanıtların, GPT-4o mini ve Gemini'nin yanıtlarına kıyasla anlamlı derecede daha kapsamlı olduğu bulundu (sırasıyla p=0,045 ve p=0,001). Bununla birlikte, GPT-4o mini ve Gemini arasında kapsam açısından anlamlı bir fark gözlenmedi. GPT-4.0'ın S1 ve S2 formatlarındaki yanıtlarının kapsamlılık skorları GPT-4o mini ve Gemini'den daha yüksekti (Tablo 5).

Doğruluk Skorları

BDM'lerden alınan ilk yanıtların doğruluk skorları karşılaştırıldığında, GPT-4.0'ın doğruluk skoru Gemini'den istatistiksel olarak anlamlı derecede yüksek bulundu (p=0,001). Ancak, GPT-4o mini ile Gemini ve GPT-4.0 arasında doğruluk açısından anlamlı bir fark gözlenmedi. S1 ve S2 formatlarındaki yanıtların doğruluk skorları karşılaştırıldığında, GPT-4.0 Gemini'den anlamlı düzeyde daha doğruydu (sırasıyla p=0,039 ve p=0,034). Başka istatistiksel olarak anlamlı fark gözlenmedi (Tablo 5).

Tartışma

Bu çalışmada, TOD ROP tedavi kılavuzunda yer alan hasta eğitim materyallerinin okunabilirliği değerlendirilmiştir. Bezirci-Yılmaz okunabilirlik formülüne göre materyaller lise için ortalama düzeyde iken Ateşman okunabilirlik formülüne göre 11. veya 12. sınıfa karşılık geliyordu. Türkiye'de yapılan araştırmalar ortalama eğitim süresinin 6,51 yıl olduğunu ortaya koymuştur.[28] Hasta eğitim materyalleri oluşturulurken her ülkenin ortalama eğitim düzeyinin dikkate alınması önemlidir.[29] Literatürde hasta eğitim materyalleri için önerilen okuma düzeyi genellikle 6. sınıf seviyesindedir.[12] Bu seviyeyi aşan materyallerin, sağlık okuryazarlığı sınırlı olan hasta popülasyonu tarafından yorumlanması zor olabilir ve bu da tedaviye uyumu azaltabilir. Bu nedenle, TOD ROP kılavuzunun okuma düzeyi, hasta eğitim materyalleri için önerilenden daha yüksek bulunmuştur ve sonuçlar basitleştirilmesi gerektiğine işaret etmektedir. ChatGPT-4.0, ChatGPT-4o mini ve Gemini tarafından üretilen materyallerde de benzer bir sorun ile karşılaşılmıştır. Bu materyallerin okuma düzeylerinin önerilen düzeyin üzerinde ve literatürde belirtilen standartlar ile uyumlu olmadığı belirlenmiştir.[30,31]

Tablo 3. GPT-4.0, Gemini ve GPT-4o mini'nin ilk yanıtları (İY), S1 ve S2 formatındaki yanıtları arasında Bezirci-Yılmaz okunabilirlik skorlarının ve eğitim düzeylerinin karşılaştırılması

		Bezirci-Yılmaz okunabilirlik skoru, ortalama (SD)	Eğitim düzeyi	p değeri
GPT-4.0	İY	8,30 (2,50)	İlkokul	İY ve S1: **0,005** İY ve S2: **0,012** S1 ve S2: 0,974
	S1	7,04 (3,04)	İlkokul	
	S2	6,74 (3,62)	İlkokul	
Gemini	İY	9,17 (2,40)	Lise	İY ve S1: 0,970 İY ve S2: 0,942 S1 ve S2: 0,907
	S1	8,53 (1,58)	İlkokul	
	S2	8,22 (1,46)	İlkokul	
GPT-4o mini	İY	10,72 (4,20)	Lise	İY ve S1: 0,879 İY ve S2: 0,971 S1 ve S2: 0,990
	S1	9,78 (3,04)	Lise	
	S2	10,16 (3,62)	Lise	

Anlamlı sonuçlar (p<0,05) koyu renkle gösterilmiştir. SD: Standart deviasyon

Tablo 4. GPT-4.0, Gemini ve GPT-4o mini'nin ilk yanıtları (İY), S1 ve S2 formatındaki yanıtları arasında Ateşman okunabilirlik skorları ve eğitim düzeylerinin karşılaştırılması

		Ateşman okunabilirlik skoru, ortalama (SD)	Eğitim düzeyi	p değeri
GPT-4.0	İY	62,06 (6,86)	9.-10. sınıf	İY ve S1: 0,256 İY ve S2: 0,312 S1 ve S2: 0,999
	S1	68,03 (7,56)	9.-10. sınıf	
	S2	67,65 (6,90)	9.-10. sınıf	
Gemini	İY	63,61 (7,94)	9.-10. sınıf	İY ve S1: 0,484 İY ve S2: 0,219 S1 ve S2: 0,901
	S1	65,54 (6,65)	9.-10. sınıf	
	S2	67,84 (6,85)	9.-10. sınıf	
GPT-4o mini	İY	51,07 (10,57)	11.-12. sınıf	İY ve S1: 0,904 İY ve S2: 0,684 S1 ve S2: 0,793
	S1	58,12 (9,52)	11.-12. sınıf	
	S2	56,02 (9,39)	11.-12. sınıf	

SD: Standart deviasyon

Tablo 5. GPT-4.0, Gemini ve GPT-4o mini'nin kapsam ve doğruluk skorlarının karşılaştırılması

		GPT-4.0	Gemini	GPT-4o mini	p değeri	
Kapsamlılık puanı, ortalama (SD)	İY	3,83 (0,91)	2,80 (1,16)	2,83 (1,26)	GPT-4.0 ve Gemini: GPT-4.0 ve GPT-4o mini: GPT-4o mini ve Gemini:	**0,001** **0,045** 0,078
	S1	3,57 (0,90)	2,57 (0,97)	2,70 (1,18)	GPT-4.0 ve Gemini: GPT-4.0 ve GPT-4o mini: GPT-4o mini ve Gemini:	**0,004** **0,002** 0,093
	S2	3,53 (0,90)	2,50 (1,01)	2,43 (1,14)	GPT-4.0 ve Gemini: GPT-4.0 ve GPT-4o mini: GPT-4o mini ve Gemini:	**0,030** **0,013** 0,061
Doğruluk skoru, ortalama (SD)	İY	2,90 (0,31)	2,10 (0,76)	2,50 (0,57)	GPT-4.0 ve Gemini: GPT-4.0 ve GPT-4o mini: GPT-4o mini ve Gemini:	**0,001** 0,058 0,345
	S1	2,90 (0,31)	2,13 (0,73)	2,50 (0,57)	GPT-4.0 ve Gemini: GPT-4.0 ve GPT-4o mini: GPT-4o mini ve Gemini:	**0,039** 0,159 0,397
	S2	2,90 (0,31)	2,13 (0,73)	2,50 (0,57)	GPT-4.0 ve Gemini: GPT-4.0 ve GPT-4o mini: GPT-4o mini ve Gemini:	**0,034** 0,217 0,231

Anlamlı sonuçlar (p<0,05) koyu renkle gösterilmiştir. SD: Standart deviasyon

ROP tedavisindeki gecikmeler, geri dönüşü olmayan görme kaybının yanı sıra sağlık çalışanları için önemli medikolegal sorunlara yol açabilir.[32] ROP ile ilgili malpraktis olgularında en sık karşılaşılan konu zamanında tarama veya takip yapılmamasıdır.[33] Bunun temel nedenlerinden biri ailelerin ROP ve tarama süreci hakkında yeterli bilgiye sahip olmamasıdır. Literatürde yapılan çalışmalarda ebeveynler bilgilendirildiğinde ve bilinçlendirildiğinde tedaviye uyumun arttığı ve bebeklerde sonuçların daha iyi olduğu gösterilmiştir.[9,10] Bir çalışmada, çok düşük doğum ağırlıklı bebeklerin ebeveynlerinin, özellikle İngilizce'si yetersiz ve sağlık okuryazarlığı düşük ise ROP hakkında yeterince bilgi edinemediği ve bunun da tedaviyi olumsuz etkilediği bildirilmiştir.[34] Çalışma, ebeveynlerin yarısından fazlasının taburcu olduktan sonra bebeklerinin ROP durumu hakkında yeterli bilgi almadığını göstermiştir. Bu bilgi eksikliğinin bir nedeni, Amerika Birleşik Devletleri'ndeki 10 yetişkinden 1'inin sağlık okuryazarlığının düşük olmasıdır.[2]

Pediatrik oftalmoloji alanında yapılan bir analiz, çevrimiçi hasta eğitim materyallerinin ortalama 11,75±2,72 yıl süre ile eğitim almış bir popülasyon için uygun olduğunu ortaya koymuştur.[34] Eğitim materyalinin anlaşılabilirliğinin düşük olması, sağlık okuryazarlığı sınırlı olan kişilerde tedaviye uyum sorunlarına yol açabilir. Bu nedenle, bilgi düzeyi düşük bireyler için anlaşılması kolay hasta eğitim materyallerinin sağlanması gerekmektedir. Çalışmamızda toplanan verilere göre, TOD ROP kılavuzlarının okuma düzeyi kabul edilemez derecede yüksek bulunmuştur. Bu nedenle bu materyallerin anlaşılabilirliğinin artırılması gerekmektedir.

Bu çalışmada, GPT-4.0, GPT-4o mini ve Gemini tarafından hazırlanan broşürlerin okunabilirlik düzeyleri TOD broşürü ile karşılaştırıldığında, GPT-4.0 ve Gemini'nin okunabilirlik düzeylerinin TOD broşürüne kıyasla daha düşük olduğu bulunmuştur. Ayrıca, anlaşılabilirliği artırmak için tasarlanan S1 ve S2 formatlarında, GPT-4.0 tarafından oluşturulan broşürün okunabilirliğinde bir artış (Bezirci-Yılmaz skoru ile değerlendirilmiştir) gözlenirken, Gemini veya GPT-4o mini için anlamlı bir değişiklik meydana gelmemiştir. Bu bulgular literatür ile uyumludur.[27,35,36] Okunabilirlik açısından bu bulgular, GPT-4.0'ın bir Türk ROP kılavuzu oluşturmak için daha uygun bir seçenek olabileceğini göstermektedir.

BDM'ler, sağlık sektöründe yeni ve ilgi çeken araçlardır ve gelişmektedirler. Özellikle hasta konsültasyonu, tıbbi triyaj ve bilgi sağlama konusunda potansiyele sahiptirler. BDM'ler, hastalardan gelen genel tıbbi soruları yanıtlayarak ve uzak veya yeterli hizmet alamayan bölgelerdeki bireylerin sağlık hizmetlerine erişimini artırabilirler.[22,37] Ayrıca, bu modellerin idari görevler üstlendiği ve sağlık çalışanlarının hasta bakımına daha fazla zaman ayırmasına olanak sağladığı gözlemlenmiştir.[38] Bununla birlikte, BDM'lerin kullanımının belirli zorlukları vardır. BDM'ler, özellikle tıbbi ortamlarda hastalar ve aileleri için risk oluşturarak yanlış bilgiler verebilirler.[39] Bu modellerin yanıtlarını kendi kendine kontrol etme ve hataları düzeltme kapasitesi sınırlıdır. Yanıltıcı veya eksik bilgiler tıbbi hatalara yol açarak hasta güvenliği için ciddi riskler oluşturabilir.[40] BDM'leri klinik uygulamaya tam olarak entegre etmek için, doğrulama süreçlerinde iyileştirme yapılması ve bu modellerin daha sıkı gözetim altında tutulması gereklidir.

Hasta eğitim materyallerinin okunması sadece kolay olmamalı, aynı zamanda eksiksiz ve doğru olmalıdır. Çalışmamızda, BDM tarafından oluşturulan broşürlerin doğruluğunu ve kapsamını da değerlendirdik. Sonuçlar, GPT-4.0 tarafından oluşturulan materyallerin GPT-4o mini ve Gemini tarafından hazırlanan materyallerden daha eksiksiz olduğunu gösterdi. Doğruluk açısından GPT-4.0 en yüksek puanı alırken, Gemini en düşük

puanı aldı. Bu veriler, GPT-4.0'ın hasta eğitim materyalleri hazırlamak için daha güvenilir bir model olabileceğini göstermektedir. Benzer şekilde, Pushpanathan ve ark.[26], GPT-4.0'ın karmaşık oküler semptomlar ile ilgili soruları yanıtlarken doğruluk ve kapsam açısından GPT-3.5 ve Google Bard'dan daha iyi performans gösterdiğini bulmuş ve hasta eğitiminde yerleri olabileceğini bildirmişlerdir. Antaki ve ark.[21] da diğer BDM'lere kıyasla GPT-4.0'ın oftalmoloji alanında daha tutarlı ve konu ile ilgili tıbbi bilgiler sağladığını bildirmiş ve güvenilir eğitim materyalleri üretmede yararlanılabileceğini ifade etmiştir.

BDM'ler tarafından verilen tıbbi bilgilerle ilgili dikkat edilmesi gereken bir diğer nokta, verilerde coğrafi varyasyon olma olasılığıdır. ROP için tarama kriterleri ülkeye göre farklılık gösterebilir.[2] Gelişmiş ülkelerde bazı kriterler karşılanmayabilirken, az gelişmiş ülkelerde şiddetli ROP riski daha yüksektir.[39] TOD ROP kılavuzu, 34. gebelik haftasından önce doğan veya 1.700 gramdan düşük ağırlığa sahip tüm yenidoğanların taranmasını önermektedir.[5] GPT-4.0'ın bu soruya verdiği yanıt ("30 haftadan önce doğan veya 1.500 gramdan düşük ağırlığa sahip bebekler"), Birleşik Krallık'ta kullanılan tarama kriterleriyle uyumluydu, ancak Türkiye için TOD standartlarından farklıydı.[41] Bu fark, hasta yakınları için belirsizlik yaratabilir ve yanlış bilgi edinme ve tedaviye uyumun azalmasına yol açabilir.

Çalışmanın Kısıtlılıkları

Çalışmamızın en önemli kısıtlılıklarından biri, dil modellerinin farklı dillerdeki performanslarının değişken olmasıdır. Çalışmamızda Türkçe sorular sorduk ve yanıtların Türkçe olarak verilmesini istedik. Ayrıca, dil modellerinden Türkçe kaynaklardan daha anlaşılır yanıtlar üretmelerini istedik. Bununla birlikte, BDM'ler tipik olarak İngilizce veriler üzerinde eğitildiğinden, Türkçe gibi dillerdeki performansları aynı düzeyde olmayabilir. Bu fark, dil yapıları arasındaki farklılıklara ve mevcut Türk veri setlerinin sınırlı olmasına bağlanabilir.[20] Literatürde, BDM'lerin daha az yaygın olan dillerde tıbbi bilgi üretirken düşük performans gösterme eğiliminde olduğu ve bunun da klinik uygulamalarda hata riskini artırabileceği belirtilmiştir.[42] Ayrıca, TOD broşüründe yer alan sorular herhangi bir ek bilgi verilmeden ve Türkiye'deki bir kullanıcı tarafından sorulduğu belirtilmeden, doğrudan sorulmuştur. Bu nedenle, "Türkiye için soruyorum" gibi bir ifadenin modelin yanıtları üzerindeki potansiyel etkisi değerlendirilmemiştir. Bu sebeple, bu modeller Türkçe gibi dillerde kullanılmadan önce dikkatle düşünülmeli ve yerel uzmanlar tarafından yürütülen doğrulama süreçleriyle desteklenmelidir.

Sonuç

Hastalar ve ailelerini eğitmek, ROP tedavisinde kritik öneme sahiptir. TOD hasta bilgilendirme broşürlerinin okuma düzeyinin kabul edilebilir düzeyden daha yüksek olduğu belirlendi. Okunabilirlik, kapsam ve doğruluk açısından GPT-4.0 broşürleri, GPT-4o mini ve Gemini broşürlerinden daha iyi performans gösterdi. BDM'ler sağlık hizmetlerinde umut verici bir araç olsalar da, verdikleri bazı bilgilerin yanıltıcı olabileceği ve coğrafi farklılıklar nedeniyle yanlış yönlendirme risklerinin olduğu görüldü. Sonuç olarak, BDM'lerin sağlık hizmetlerine entegrasyonu kapsamlı bir şekilde test edilmeli ve önerilerle desteklenmelidir. BDM'ler tarafından üretilen bilgilerin, özellikle de temel tıbbi bilgilerin doğruluğunun dikkatle değerlendirilmesi gerekmektedir.

Etik

Etik Kurul Onayı: Gerekmez.
Hasta Onayı: Gerekmez.

Beyan

Yazarlık Katkıları

Konsept: S.A.P., Dizayn: A.D., Veri Toplama veya İşleme: S.A.P., Analiz veya Yorumlama: A.D., Literatür Arama: S.A.P., Yazan: S.A.P., A.D.

Çıkar Çatışması: Yazarlar bu makale ile ilgili olarak herhangi bir çıkar çatışması bildirmemiştir.

Finansal Destek: Çalışmamız için hiçbir kurum ya da kişiden finansal destek alınmamıştır.

Kaynaklar

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.