

Evaluating the efficacy of artificial intelligence as a patient education tool in paediatric care: a cross-sectional comparative study

Ellaha Haidar ,¹ Alice Ruan²

To cite: Haidar E, Ruan A. Evaluating the efficacy of artificial intelligence as a patient education tool in paediatric care: a cross-sectional comparative study. *BMJ Paediatrics Open* 2026;**10**:e004890. doi:10.1136/bmjpo-2026-004890

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjpo-2026-004890>).

EH and AR are joint first authors.

Received 3 May 2026
Accepted 8 September 2026



© Author(s) (or their employer(s)) 2026. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

¹Imperial College Healthcare NHS Trust, London, UK
²Imperial College London, London, UK

Correspondence to
Dr Ellaha Haidar; vera.haidar3@nhs.net

Patients and caregivers are turning to artificial intelligence (AI) to understand their medical conditions and treatment options;^{1 2} however, concerns remain about the quality, accuracy and readability of AI-generated advice. In this study, we sought to evaluate the ability of AI chatbots to generate health information materials compared to National Health Service UK (NHS) leaflets.

A retrospective qualitative analysis was conducted comparing information generated by ChatGPT-3.5 and DeepSeek R1 with NHS leaflets from Moorfields Eye Hospital on four paediatric ophthalmological conditions: chalazion, blepharitis, amblyopia and strabismus.³ Prompts are detailed in online supplemental materials (appendices A-D). Readability was assessed using four metrics: Flesch-Kincaid Reading Ease, Flesch-Kincaid Grade Level (FKGL), Gunning Fog Score (GFS) and the Simple Measure of Gobbledygook (SMOG) index.⁴ Information quality was evaluated by two assessors blinded to the material's source using the Ensuring Quality Information for Patients (EQIP)⁵ tool. Statistical analysis included analysis of variance (ANOVA) for group comparisons and Cronbach's alpha to assess inter-rater reliability. Statistical significance was set as $p < 0.05$.

The FKGL indicated that AI-generated materials required a significantly higher reading level ($p = 0.0011$). As seen in [table 1](#), NHS content required the lowest reading grade level (mean 4.1, SD 0.90), compared with ChatGPT (mean 7.5, SD 1.81) and DeepSeek (mean 9.2, SD 1.02). Similarly, FRE scores demonstrated that Moorfields content was easiest to read (mean 74.8, SD 6.74), followed by ChatGPT (mean 57.7, SD 11.36) and DeepSeek (mean 48.1, SD 5.01), with a statistically significant difference ($p = 0.0038$).

A score above 65 is considered accessible for the general population, and only Moorfields exceeded this threshold. GFS and SMOG, which estimate the years of education required to understand a text, showed no statistically significant differences between sources; however, Moorfields still had the lowest average scores. EQIP scores showed NHS materials scored significantly higher for information quality ([table 2](#)): Moorfields mean 69.61, SD 4.74; ChatGPT mean 58.26, SD 3.35; DeepSeek mean 56.75, SD 4.73).

This is the first comparative study evaluating AI performance in generating patient education materials for parents of children with common ophthalmic conditions. Given that over 75% of individuals turn to the internet during health crises and approximately one in six adults use AI chatbots monthly for health information,¹ this study offers timely insight into the readability and quality of these emerging tools.

Our analysis demonstrates that human-generated Moorfields Eye Hospital leaflets outperform both AI systems. According to four readability indices, ChatGPT and DeepSeek required reading ages of 11–14 and 14–17, respectively, compared with 8–11 for NHS leaflets. Human-authored content was the only material to attain the recommended FRE benchmark score of

These findings are consistent with previous research across other medical specialties.⁶

The limited readability of AI-generated content may reflect how large language model algorithms prioritise information volume over clarity, and their lack of real-world patient education experience.

Several quality limitations were also identified. Neither AI system referenced reliable sources or provided citations, undermining

Table 1 Mean readability scores measured by FKGL, FRE, GFS and SMOG index for information provided by Moorfields, ChatGPT and DeepSeek across four paediatric ophthalmic conditions: chalazion, amblyopia, squint and blepharitis

Readability metric	Moorfields	ChatGPT	DeepSeek
Mean FKGL	4.10±0.902	7.53±1.81	9.20±1.02
Mean FRE	74.8±6.74	57.7±11.4	48.1±5.01
Mean GFS	11.1±0.622	12.5±3.23	13.2±1.39
Mean SMOG	11.6±1.48	12.6±1.98	13.2±0.847

FKGL, Flesch-Kincaid Grade Level; FRE, Flesch Reading Ease; GFS, Gunning Fog Score; SMOG, Simple Measure of Gobbledygook.

evidence-based practice and patient safety. Visual aids, shown to improve understanding and retention, were absent from AI-generated materials. Conversely, AI tools offer an advantage in timeliness. Unlike the NHS leaflets last reviewed between 2020 and 2022, ChatGPT and DeepSeek can provide up-to-date information through real-time browsing capabilities. Despite their potential, AI tools carry significant risks. As this study focused on readability and information quality, factual accuracy was not assessed as a separate outcome. However, recognised risks include ‘hallucinations’, a phenomenon where AI confidently produces fabricated or inaccurate information.⁷ Human oversight therefore remains essential as a fact-checking safeguard.

Looking ahead, future AI models with adaptable readability settings and multidisciplinary development could represent a scalable approach for personalised patient education. Until such advances are realised, trusted professional leaflets should remain the standard. Future research should evaluate AI-generated materials through real-world patient involvement and assess whether

Table 2 Ensuring Quality Information for Patients (EQIP) scores assessing the quality of information provided by Moorfields, ChatGPT and DeepSeek across four paediatric ophthalmic conditions: chalazion, amblyopia, squint and blepharitis

Quality metric	Moorfields	ChatGPT	DeepSeek
EQIP Person 1	70.3±3.58	58.3±2.25	57.0±5.78
EQIP Person 2	69.0±6.47	59.2±4.85	55.6±3.91
Mean EQIP	69.6±4.74	58.8±3.35	56.3±4.73

individually adapted prompts improve the accessibility and quality of outputs.

As the use of AI expands, this study offers insight into their current capabilities and limitations. Models like ChatGPT and DeepSeek can generate accessible, timely content, but in this sample of paediatric patient information, outputs remain less readable and lower quality than clinician-authored materials. AI should remain supplementary until standards improve, though with oversight may support patient education.

Contributors Both authors EH and AR have contributed equally to this work and share first authorship.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests No, there are no competing interests.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <https://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Ellaha Haidar <https://orcid.org/0009-0003-0739-7477>

REFERENCES

- 1 KFF health misinformation tracking poll: artificial intelligence and health information. KFF; 2025. Available: https://www.kff.org/health-information-trust/poll-finding/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/?utm_source=chatgpt.com
- 2 Yun HS, Bickmore T. Online Health Information-Seeking in the Era of Large Language Models: Cross-Sectional Web-Based Survey Study. *J Med Internet Res* 2025;27:e68560.
- 3 Leaflet library. Moorfields Eye Hospital; 2025. Available: <https://www.moorfields.nhs.uk/for-health-professionals/leaflet-library>
- 4 Wang L-W, Miller MJ, Schmitt MR, et al. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Res Social Adm Pharm* 2013;9:503–16.
- 5 Moul B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expect* 2004;7:165–75.
- 6 Shamil E, Ko TK, Fan KS, et al. Assessing the Quality and Readability of Online Patient Information: ENT UK Patient Information e-Leaflets versus Responses by a Generative Artificial Intelligence. *Facial Plast Surg* 2025;41:472–81.
- 7 Alkaissi H, McFarlane SI, Alkaissi H. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* 2023;15:e35179.