Review

Large Language Models for Generating Patient Education Materials: Scoping Review

Chengjin Li[1,2*]; Juan Li[1*]; Xuezhu Liu[1*]; Zefang Gao[3]; Yuxiao Xu[4]; Xiuze Xu[1]; Xiaoyu Liu[1]; Tong Zhao[5]; Yuanyuan Cheng[1*]; Ming Li[6]

[1]Department of Human Resource, First Affiliated Hospital of Soochow University, Soochow, Jiangsu, China

[2]Shanghai Engineering Research Center of Intelligence Pediatrics, Shanghai Children's Medical Center, Shanghai Jiao Tong University, Shanghai, Shanghai, China

[3]Clinical Teaching Office, First Affiliated Hospital of Soochow University, Soochow, Jiangsu, China

[4]School of Public Health, Shanghai Jiao Tong University, Shanghai, Shanghai, China

[5]School of Political Science and Public Administration, Wuhan University, Wuhan, Hubei, China

[6]Department of Nephrology, First Affiliated Hospital of Soochow University, Soochow, Jiangsu, China

[*]these authors contributed equally

Corresponding Author:
Ming Li
Department of Nephrology
First Affiliated Hospital of Soochow University
899 Pinghai Road
Soochow, Jiangsu, 215006
China
Phone: 86 13656237709
Email: liminglm@suda.edu.cn

Abstract

Background: Patient education materials (PEMs) are essential for improving health literacy and clinical outcomes. Although large language models (LLMs) hold promise for optimizing content, their performance in generating new PEMs remains unclear.

Objective: This study aimed to summarize the current evidence on LLMs in generating new PEMs, evaluate readability, quality, and usability in various health care settings, and guide future research and practice.

Methods: This review followed PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. A comprehensive literature search was conducted in PubMed, Embase, Cochrane Library, Web of Science, and Scopus from January 1, 2000, to January 1, 2026. Eligible studies were English-language original research studies that evaluated new PEMs generated by LLMs in health care using information assessment tools. Two reviewers independently screened records and extracted data. We conducted a quantitative analysis of the characteristics of the included literature. Furthermore, we classified the PEMs based on core educational tasks and clinical scenarios, and compared their readability, quality, and usability.

Results: Thirty studies were included, including 11 LLMs and more than 30 diseases; most compared multiple LLMs and used a single prompt. PEMs were classified into 5 categories: perioperative patient education (n=8), chronic disease management education (n=6), specialty disease diagnostic and therapeutic education (n=9), pharmacotherapy safety and adherence education (n=3), and public health and prevention education (n=4). In terms of readability, only 3 records achieved a sixth-grade reading level. Specialty disease diagnostic and therapeutic education performed relatively well, while chronic disease management education had poor readability. The DeepSeek series demonstrated the most consistent performance. In terms of quality, most records were rated as good or higher. Specialty disease diagnostic and therapeutic education showed a polarized distribution, with both high and low quality ratings. Pharmacotherapy safety and adherence education and public health and prevention education had no poor-quality ratings. The ChatGPT series performed relatively well. In terms of usability, most records performed well in understandability but poor in actionability. Perioperative patient education showed a relatively balanced performance.

Conclusions: LLMs offer a transformative opportunity for developing effective and high-quality PEMs, but they currently fail to meet health literacy standards, particularly in readability and actionability. Future efforts must shift from performance evaluation

XSL•FO
RenderX

to implementation strategies. Only through rigorous validation and equitable deployment can LLMs truly improve patient education.

(J Med Internet Res 2026;28:e98084) doi: 10.2196/98084

KEYWORDS

large language models; generative language model; patient education; readability; quality; usability; scoping review

Introduction

Background

Patient education materials (PEMs) are key tools for improving patients' health literacy, enhancing treatment adherence, and improving clinical outcomes [1-3]. However, the effectiveness of PEMs is largely limited by their readability [4-9]. The National Institutes of Health (NIH) recommends that medical information be written at a sixth-grade reading level to accommodate the health literacy of the general population [10]. Unfortunately, due to the overuse of technical terminology, traditional manually written materials often fail to meet this standard and are costly, time-consuming, and labor-intensive, which limits access to information and exacerbates health inequalities [11-13].

In recent years, breakthroughs in generative AI (GenAI) have opened up new opportunities to address this challenge [11,14-17]. Technologies such as large language models (LLMs), with their powerful natural language processing capabilities, can efficiently generate structured, personalized medical content [18,19]. As tools like ChatGPT and Google Gemini become increasingly widely adopted by the public, more and more patients and clinicians are beginning to use LLMs to obtain health information or assist in creating educational materials [15,16,20]. Evidence suggests that LLMs have great potential in simplifying text and explaining complex treatment regimens, and are expected to alleviate the limitations of traditional materials such as high costs and long update cycles [21-23].

Despite rapid development, the efficacy and safety of LLMs in generating PEMs remain controversial [24-27]. Most existing studies focus on a single disease or a specific model, with inconsistent results [28-30]. Some studies indicate that model-generated content performs well in terms of technical accuracy, but often has problems such as poor readability, insufficient actionable guidance, and the potential risk of hallucinations [28,31-33]. Patients generally have low digital health literacy and are more accustomed to directly accessing the internet or using LLMs to search for health knowledge, rather than actively consulting evidence-based clinical guidelines [28,31-33]. Furthermore, there is currently a lack of systematic reviews of different disease areas, model performance, and the use of evaluation tools, making it difficult to reach a unified consensus on clinical application [34-36]. Notably, current evaluations usually rely mainly on expert judgment or generic metrics, while patient involvement is still largely lacking [32,37].

Objective

Although prior reviews have explored the broader application of LLMs in patient education, they have not systematically focused on the generation of new PEMs. The emerging evidence base for their performance in different clinical scenarios remains fragmented. Given the heterogeneity of the literature in terms of study designs, evaluation methods, and reporting depth, we conducted a scoping review to map the landscape rather than to estimate pooled effect sizes.

The primary objective of this review was to synthesize the published literature on assessing the performance of new PEMs generated by LLMs. The specific objectives were to (1) characterize the clinical scenarios or core educational tasks addressed by these materials; (2) identify the information assessment tools used to evaluate their readability, quality, and usability; and (3) synthesize the findings reported across these evaluative dimensions and compare the performance of LLMs. These findings are expected to provide a basis for improving the design, evaluation, and implementation of LLMs, and for promoting more consistent and comparable assessment across future studies.

Methods

Protocol and Registration

This scoping review was designed to systematically map and synthesize evidence on the readability, quality, and usability of LLM-generated new PEMs across clinical settings. The review followed the methodological framework proposed by Arksey and O'Malley [38] and was reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [39] guidelines, and a completed checklist is provided in Multimedia Appendix 1. The review protocol was registered in the Open Science Framework (OSF registration number: EN425).

Information Sources

We conducted a systematic search in PubMed, Embase, the Cochrane Library, Web of Science, and Scopus. The searches were limited to peer-reviewed papers published between January 1, 2000, and January 1, 2026. We also conducted backward and forward citation tracking for the included studies. The final search was completed on January 25, 2026. Additionally, grey literature databases were not searched because they may lack complete information on study design or data collection, and are not necessarily peer-reviewed, which could compromise the standardization of data extraction and the reliability of comparisons across studies.

XSL•FO
RenderX

Search

The search strategy was developed and refined by 2 researchers (CL and YC) with the assistance of a research librarian. It was structured around 2 core concepts: LLMs and PEMs. Controlled vocabulary and free-text terms were combined using Boolean operators. The search syntax and concept-block structure were adapted to the indexing system and functionality of each database. Apart from publication-date limits, no filters were applied for language, age, or publication type at the search stage. The full search strategy is provided in Multimedia Appendix 2.

Eligibility Criteria

Eligibility criteria were defined a priori using the population, concept, and context (PCC) framework [40] recommended for scoping reviews (Textbox 1). In this review, the population referred to implemented LLMs relevant to health care, the concept referred to new PEMs generated by LLMs and its evaluation, and the context referred to any health care setting. Original studies published in English that evaluated new PEMs generated by LLMs in health care using information assessment tools were included, provided that the full text was available. The exclusion of literature other than English is due to the significant differences in syllabic structure, orthography, and syntax between different languages, which may affect the consistency of evaluation criteria and the comparability of evaluation results.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria:

- Use large language models (LLMs) to generate text for patient education.
- Generate text for patients with any disease or undergoing any treatment process.
- Create new patient education materials (PEMs).
- Evaluate the text generated by the model using information assessment tools.

Exclusion criteria:

- Duplicate literature;
- LLMs were not used to generate PEMs.
- The generated text was not patient-centered health education.
- Improved existing PEMs rather than generating new ones.
- Review, grey literature, protocols, conference papers, letters, and other literature without full text.
- Studies published in languages other than English.

Selection of Sources of Evidence

One researcher (CL) conducted the literature search. All retrieved records were imported into NoteExpress, and duplicate entries were removed. Two researchers (CL and XL) independently screened the titles and abstracts to assess whether they met the inclusion criteria. Records judged potentially eligible or unclear by either researcher proceeded to full-text assessment. The same 2 researchers (CL and XL) then independently conducted a detailed assessment based on the eligibility criteria. Reasons for exclusion at the full-text stage were documented and reported in the PRISMA flowchart. Disagreements at any stage of the screening process were resolved by consultation with a senior researcher (ML).

Data Charting Process

Two independent reviewers (CL and YC) developed a standardized data extraction form in Microsoft Excel. The form was piloted on 3 randomly selected included studies and refined to ensure consistent interpretation of variables and coding rules. One researcher (CL) initially was responsible for data charting, and another researcher (JL) independently verified the extracted data for accuracy and consistency. Any discrepancies were resolved through discussion, and a third researcher was consulted when necessary (ML).

Data Items

The data extracted from each included study included title, publication year, journal, first author, name and version of the LLMs, input instructions of LLMs, diseases, evaluation indicators, evaluator characteristics, data collection time, and main conclusions. We also extracted the information assessment tools used in each study such as DISCERN [41]. Although the assessment tools varied across studies, most evaluated at least one of the following domains: (1) readability, including language difficulty and reading fluency; (2) quality, including the reliability, accuracy, and completeness of the generated content; and (3) usability, including the understandability and actionability of PEMs. All tools used to evaluate these domains and their corresponding results in each study were extracted completely.

Critical Appraisal of Individual Sources of Evidence

In accordance with scoping review methodology, we did not conduct a formal methodological quality appraisal or risk-of-bias assessment. This study was designed to map the range and characteristics of the available evidence rather than to evaluate intervention effects or the strength of evidence outcomes. Although no formal appraisal was performed, this limitation was acknowledged and considered when interpreting the findings.

XSL•FO
RenderX

Synthesis of Results

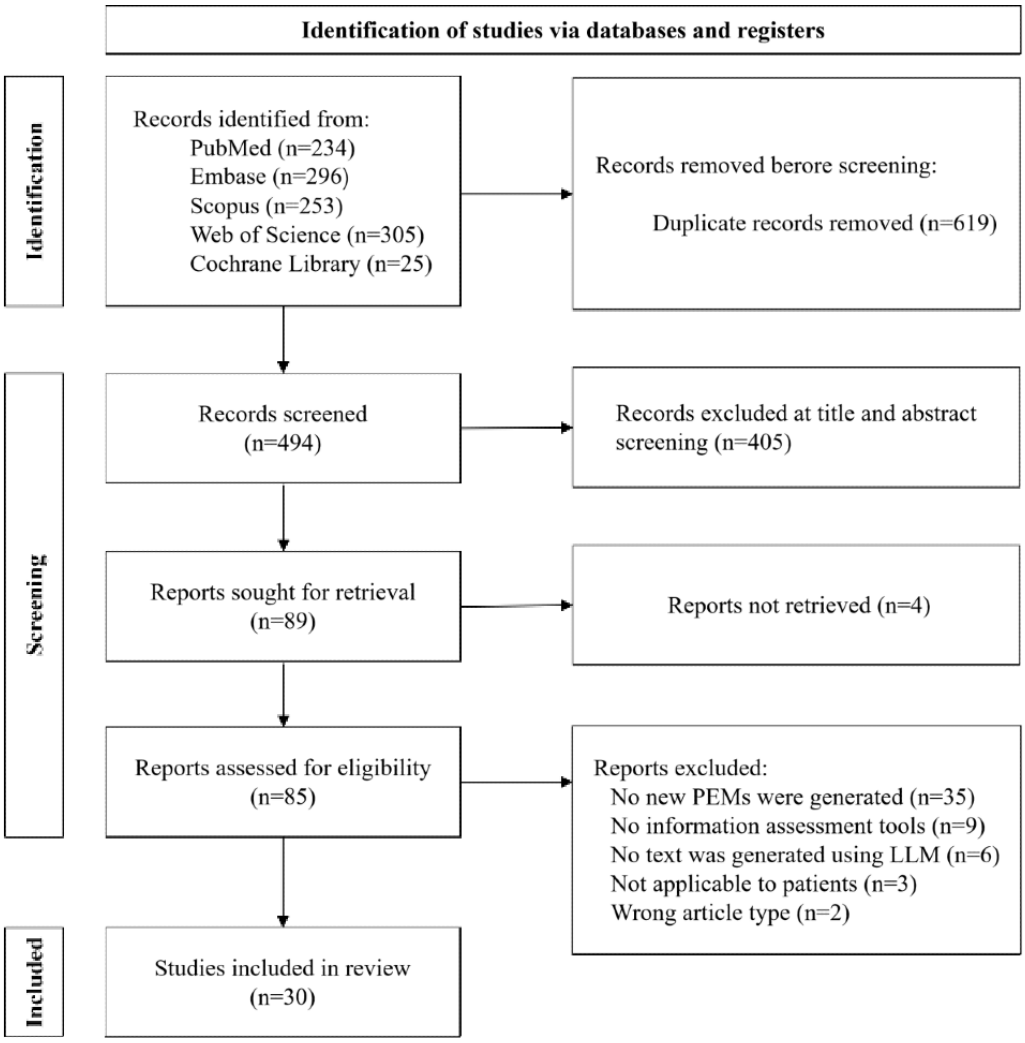
All charted data were synthesized using descriptive and narrative methods. The calculation of frequencies and percentages was performed using Microsoft Excel. To visually present evaluation gaps, we classified the PEMs evaluated in each included study and compared their readability, quality, and usability. Findings were summarized using tables and figures to support interpretability and transparency.

Results

Overview

The study screening process is summarized in the PRISMA-ScR flow diagram (Figure 1). A total of 1113 records were retrieved from the 5 databases. After removing 619 duplicates, 494 records were screened for titles and abstracts, of which 405 were excluded. The remaining 89 reports were sought for retrieval, and 4 were not retrieved. Following full-text review, 30 studies [42-71] met the inclusion criteria.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram. LLM: large language model; PEM: patient education material.



Data Summary

As shown in Table 1, a total of 5 studies [52-54,67,68] evaluated a single model, while 25 studies [42-51,55-66,69-71] compared multiple models. Included studies covered 11 different LLM families. All studies included models from the ChatGPT family; the Google Gemini family appeared in 14 studies [42,45-47,49-51,59-62,64,65,69], and the DeepSeek family appeared in 6 studies [43,45,56-58,71]. Other evaluated models included Google Bard (n=3), Anthropic Claude (n=2), Meta Llama (n=2), and Microsoft Copilot (n=2).

XSL•FO
RenderX

Table 1. Characteristics of included articles.

Author, year	LLMs[a]	Instruction	Sample size	Collection data	Disease	Main conclusion
Aayush et al, 2025 [42]	ChatGPT 3.5, Gemini 1.5.2	Prompt	Unknown	February 14-22, 2024	3 Neurological diseases: Parkinson, Alzheimer, migraine	Gemini generated shorter, more original content, yet neither demonstrated clear superiority, and PEMs need to continue optimizing.
Mohammed et al, 2025 [43]	DeepSeek, ChatGPT 4	Prompt	Unknown	March 1-8, 2025	Imaging diagnosis of stroke	ChatGPT and DeepSeek performed similarly in understanding and readability, and are both suitable for PEMs[b].
Qais et al, 2025 [44]	ChatGPT 3.5, ChatGPT 4, Google Bard	Prompt (20 repetitions)	2 researchers	November-December 2023	Idiopathic intracranial hypertension (IIH)	LLMs, particularly ChatGPT-4, can produce high-quality, readable PEMs on IIH.
Jamila et al, 2025 [45]	ChatGPT, Gemini, DeepSeek, RCS England	Prompt	Unknown	Unknown	Carpal tunnel release (CTR), total hip arthroplasty (THR), total knee replacement (TKR)	Despite their potential, LLM-generated leaflets were less readable than professional ones.
Ali et al, 2025 [46]	ChatGPT-4-turbo, Gemini 2.5 Pro, ENT UK	5 questions	6 surgeons	May 2025	6 ear, nose, and throat (ENT) conditions: septal perforation, juvenile nasopharyngeal angiofibroma, otosclerosis, vestibular migraine, tracheomalacia in children, and thyroglossal cyst	GPT and Gemini matched professional content in quality but required higher reading levels, failing to meet recommended literacy standards.
Nuran et al, 2025 [47]	ChatGPT, Gemini	32 questions	8 experts, diabetes nurses	May 8, 2024	Gestational diabetes mellitus (GDM)	ChatGPT and Gemini demonstrated high readability, comprehensibility, and actionability on GDM.
Polly et al, 2025 [48]	ChatGPT 3.5, Pi, Google Bard, Perplexity, Bing AI, ChatSonic, Moorfields	Prompt	3 researchers	October 2023	Cataract surgery	ChatGPT 3.5 scored the most poorly, ChatSonic indicated promise in assisting information giving in ophthalmology.
Gowri et al, 2025 [49]	Chat GPT4.0, Claude 3.5 Sonnet, Gemini 1.5 Flash, Llama 3.1-405B	Prompt	5 dentists	Unknown	Postoperative guidance after tooth extraction (PGTE), tooth extraction (TE), brushing techniques (BT), oral cancer screening (OCS)	GPT-4 excelled in understandability; Llama 3.1-405B led in reliability. LLM outputs required refinement to meet PEMs standards.
Christoph et al, 2025 [50]	ChatGPT 3.5, Microsoft Copilot, Gemini 1.0 Pro, AI Version 24.0	10 questions	2 researchers	August 2024	Chemotherapy-induced cardiotoxicity	LLMs produced high-quality PEMs but with poor readability, still usable but not substitutes for doctors. The best LLMs lack consensus.
Jing et al, 2025 [51]	ChatGPT 3.5, Gemini	Prompt	2 vascular surgeons	February 2025	Varicose veins and deep vein thrombosis of the lower extremities	ChatGPT and Gemini were equally effective in creating PEMs in 2 diseases.
Kirk et al, 2025 [52]	ChatGPT 4, Clinical writing	Prompt (5 repetitions)	2 orthopedic resident physicians	Unknown	Total knee arthroplasty (TKA)	ChatGPT-4 produced high-quality TKA PEMs as good as human-generated resources, but limited by readability. Future LLM developers should work with physicians.

XSL•FO
RenderX

Author, year	LLMs[a]	Instruction	Sample size	Collection data	Disease	Main conclusion
Shinya et al, 2025 [53]	ChatGPT-4o, Web page	Prompt (10 repetitions)	1 epidemiological researcher, 1 doctor	August 20-30, 2024	Monkeypox	LLMs were more readable than web pages, which offered better visuals and guidance. Combining both could enhance PEMs. Further research is needed on integrating LLMs into public health communication.
Hatice et al, 2025 [54]	ChatGPT 4	Prompt	10 researchers	Unknown	Ostomy Care	LLMs produced accurate and operable PEMs, but was at a high reading level for patients.
Allison et al, 2025 [55]	AAO, ChatGPT-4o, Microsoft Copilot, Llama-3.1-70B-Instruct	Prompt	Unknown	September 2024	15 Ophthalmic diseases	LLMs produced readable PEMs approaching AAO[c] standards, yet often exceed the 6th-grade level, and show promise in increasing accessible PEMs availability.
Jaziya et al, 2025 [56]	ChatGPT-4o, DeepSeek V3	Prompt	2 researchers	April 10-17, 2025	4 Chronic diseases: epilepsy, heart failure, chronic obstructive pulmonary disease, chronic kidney disease	ChatGPT and DeepSeek were similar but fall short in readability, originality, and actionability, requiring human oversight for clinical use.
Mi et al., 2025 [57]	DeepSeek V3, DeepSeek R1, ChatGPT-4o, ChatGPT-o3-mini	Prompt	2 researchers	February 1, 2025	Lumbar discectomy (LD), spinal fusion (SF), decompressive laminectomy (DL)	All LLMs scored "fair," requiring better practices and personalization. DeepSeek R1 and ChatGPT-4o produced more readable PEMs, both were prioritized for assisting patients.
Shuai et al, 2025 [58]	ChatGPT 4, DeepSeek R1	30 questions	3 surgeons	January-April 2025	Bariatric surgery	ChatGPT responses were more reliable; DeepSeek's more readable, both lacked attention to psychosocial and cultural aspects of patient care.
Nithin et al, 2025 [59]	ChatGPT 3.5, Gemin	Prompt	Unknown	June 2024	3 new oral hypoglycemic agents: Semaglutide, Tirzepatide, Liraglutide	Gemini provided more readable content; ChatGPT offered greater detail. Both effectively generated reliable PEMs for glucagon-like peptide-1.
Qais et al, 2025 [60]	ChatGPT 3.5, ChatGPT 4, Gemini	Prompt (20 repetitions)	3 researchers	December 2023-January, 2024	Retinopathy of prematurity (ROP)	LLMs, particularly ChatGPT-4, can generate readable and high-quality PEMs for parents on ROP.
Mohammad et al, 2025 [61]	Gemini, ChatGPT 3.5, ChatGPT-4o (o1 Preview)	Prompt (20 repetitions)	2 researchers	October-December 2024	Childhood myopia	ChatGPT-4o produced PEMs with high accuracy, quality, and understandability.
Yuwan et al, 2025 [62]	ChatGPT-4o, Claude 3.5, Gemini 1.5	Prompt (20 repetitions)	2 researchers	October 10, 2024	Thyroid eye disease	LLMs demonstrated strong potential in generating high-quality, understandable, accurate, and empathetic content, though readability required further enhancement.
George et al, 2025 [63]	ChatGPT 3.5, ChatGPT 4	Prompt	Unknown	Unknown	50 Common hand pathologies	ChatGPT 4 proved more effective than ChatGPT 3.5 in producing sixth-grade-level PEMs for common hand conditions.

XSL•FO
RenderX

Author, year	LLMs[a]	Instruction	Sample size	Collection data	Disease	Main conclusion
Qais et al, 2025 [64]	ChatGPT 3.5, ChatGPT 4, Gemini Advanced	Prompt (20 repetitions)	2 researchers	January 13-March 9, 2024	Dry eye disease (DED)	LLMs, particularly ChatGPT-4, generated high-quality, accurate, and readable PEMs on DED.
Salahaldin et al, 2025 [65]	ChatGPT, Gemini, Web page	7 questions	2 researchers	Unknown	Meniere disease	Web pages and LLMs failed to meet AMA[d] readability standards and were poor quality and reliability. Future LLMs must enhance readability while ensuring reliability.
Maha et al, 2025 [66]	ChatGPT-4o, ChatGPT-4o (Fine-tuned as 'ACI Guide')	22 questions	2 researchers	February 1-2, 2025	Autologous chondrocyte implantation (ACI)	ChatGPT-4o produced poor-quality, excessively complex responses. Fine-tuning improved readability to 8th-9th grade levels.
Gülseren et al, 2025 [67]	ChatGPT 4	12 questions	10 researchers	Unknown	Lymphedema after breast cancer surgery	PEMs were reasonable regarding understandability and actionability. The Turkish readability level was also reasonable and easily understandable.
Elizabeth et al, 2025 [68]	ChatGPT, Lexi-Comp®	Prompt	2 researchers	Unknown	8 common drugs: Albuterol, Apixaban, Atorvastatin, Hydrocodone, Insulin Glargine, Levofloxacin, Omeprazole, Valsartan	ChatGPT failed to meet standards with inconsistent accuracy; not recommended for medication education.
Xuewei et al, 2025 [69]	ChatGPT 3.5, ChatGPT 4, ChatGPT-4o, Gemini	35 questions	5 researchers	June-August 2024	Myopia	LLMs, particularly ChatGPT-4o, delivered useful myopia education but needed refinement for complex queries.
Qais Dihan et al, 2025 [70]	ChatGPT 3.5, ChatGPT 4, Google Bard	Prompt (20 repetitions)	3 researchers	October-November 2023	Childhood glaucoma	LLMs serve as supplementary tools for generating high-quality, accurate PEMs and enhancing readability.
Amruth A Alluri et al, 2025 [71]	ChatGPT 4, DeepSeek V3	Prompt	Unknown	March 20-27, 2025	Systemic Lupus Erythematosus (SLE), Systemic Sclerosis, Dermatomyositis	ChatGPT and DeepSeek are valuable supplements to, but not replacements for, expert-verified educational resources in rheumatology.

[a]LLM: large language model.

[b]PEM: patient education material.

[c]AAO: American Academy of Ophthalmology.

[d]AMA: American Medical Association.

Regarding the instruction, 22 studies (73%) [42-45,48,49,51-57,59-64,68,70,71] used a single prompt, such as "Generate patient education materials for a specific disease," while 8 studies (27%) [46,47,50,58,65-67,69] used a predefined list of questions, ranging from 5 to 35 questions. Overall, 23 studies (77%) [44,46-54,56-58,60-62,64-70] reported the number of researchers, ranging from 1 to 12, and no patient involvement. A total of 22 studies (73%) [42-44,46-48,50,51,53,55-62,64,66,69-71] provided data collection dates, ranging from October 2023 to May 2025.

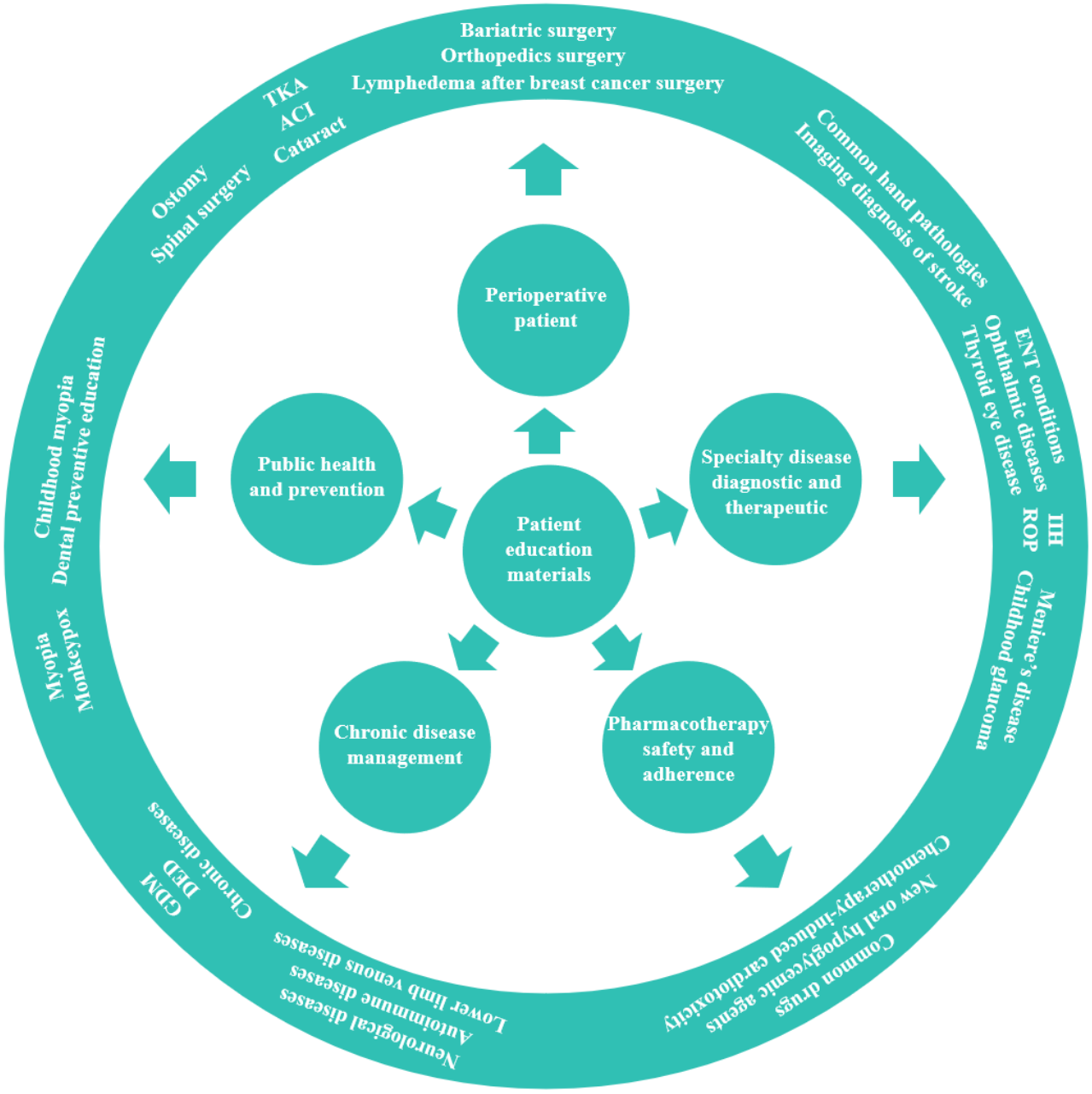
The range of diseases covered included more than 30 conditions such as ophthalmology (n=8), orthopedics (n=5), neurological disorders (n=4), and metabolic disorders (n=4). The main findings indicated that most studies reported that LLM-generated content generally exceeded the recommended sixth-grade reading level. All studies emphasized that LLMs must be used in conjunction with professional review and cannot fully replace medical resources.

Classification of PEMs

PEMs designed for different scenarios vary in core functions, content structure, knowledge complexity, and communication goals. We adhered to 2 principles, homogeneity and distinctiveness, meaning that the core educational tasks within the same category are consistent, while the boundaries between application scenarios of different categories are clear. We categorized all studies into 5 categories, as shown in Figure 2.

XSL•FO
RenderX

Figure 2. Categories of large language models. ACI: autologous chondrocyte implantation; DED: dry eye disease; ENT: ear, nose, and throat; GDM: gestational diabetes mellitus; IIH: idiopathic intracranial hypertension; ROP: retinopathy of prematurity; TKA: total knee arthroplasty.



Perioperative Patient Education

Eight studies [45,48,52,54,57,58,66,67] were included. Its core task is to provide guidance closely related to the entire perioperative period for specific surgical operations or invasive procedures. The content is highly time-consuming and actionable, and aims to ensure that patients can safely navigate the perioperative, intraoperative, and postoperative phases. Representative studies include general orthopedic surgery, cataract surgery, etc.

Chronic Disease Management Education

Six studies [42,47,51,56,64,71] were included. Its core task is to provide supportive education for patients with chronic diseases who need long-term management, focusing on disease awareness, self-monitoring, lifestyle adjustment, and promotion of treatment adherence, and aims to help patients translate medical advice into sustainable daily practices. Representative studies include cardiovascular diseases, autoimmune diseases, etc.

Specialty Disease Diagnostic and Therapeutic Education

Nine studies [43,44,46,55,60,62,63,65,70] were included. Its core task is to explain specialized disease knowledge with complex diagnosis and highly specialized treatment regimens to patients, focusing on clarifying the pathological mechanism of disease, specialized diagnosis methods, or complex treatment plans, so as to assist informed decision-making. Such studies cover many specialty areas such as neurology, otolaryngology, ophthalmology, etc.

Pharmacotherapy Safety and Adherence Education

Three studies [50,59,68] were included. Its core task is to focus on specific medications or treatment regimens and provide

XSL•FO
RenderX

accurate information about correct use, potential side effects, risk avoidance, and promotion of treatment adherence.

Public Health and Prevention Education

Four studies [49,53,61,69] were included. Its core task is to publicize disease prevention knowledge and advocate health behaviors to the public or specific risk groups. The goal is forward-looking and universal. The content includes prevention guidance and self-examination of infectious diseases (such as monkeypox), common health problems (such as children's myopia), etc.

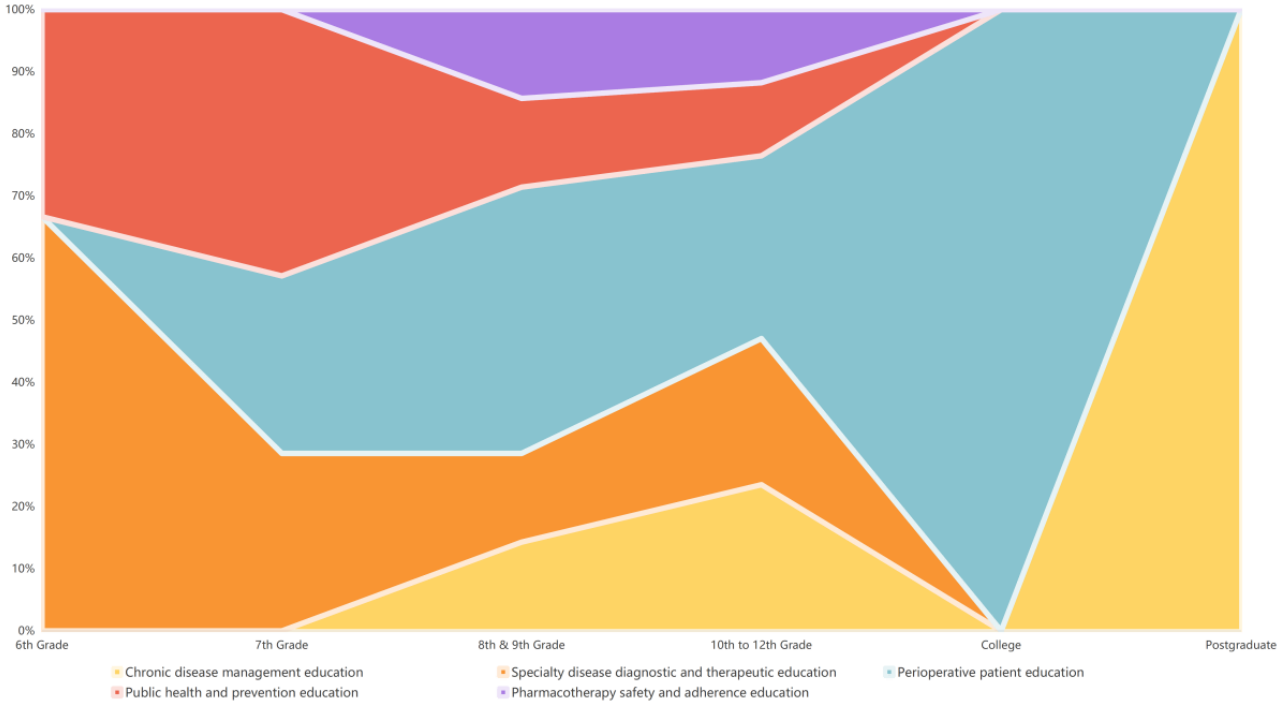
Study Outcomes

Readability

In the included studies, PEMs may have been generated by multiple LLMs and evaluated using multiple readability metrics. Given the high heterogeneity across studies, we performed an extreme-case analysis (specifically, a best-case analysis) by extracting the lowest grade level (ie, the best readability) for each PEM. Rather than representing typical readability performance, we evaluated the potential readability under the best observed condition. This better reflects the comprehension threshold that PEMs can achieve under the most favorable conditions—that is, the likelihood of being understood by individuals with low health literacy. This approach aligns with established methodological recommendations for handling heterogeneity in systematic reviews [72].

We first extracted all readability metric evaluation data, and then extracted the lowest grade level for each PEM. To address potential biases, we also reported all readability results and the range of grade levels (if available). The complete dataset is provided in Multimedia Appendix 3. Finally, based on the classification of PEMs, we presented the readability levels of 5 major categories of PEMs, as shown in Figure 3.

Figure 3. Readability based on five major disease categories.



As shown in Table 2, we extracted a total of 37 records from the 30 included studies. The excess was due to 3 studies [45,49,57] that evaluated PEMs for more than a single disease. In terms of best readability, only 3 records met the NIH-recommended 6th-grade reading level [62,63,69], while in terms of worst readability, 25 records reached the reading level of college or above. Specifically, Specialty disease diagnostic and therapeutic education performed best overall; it was the only category to reach the 6th-grade level, with most records falling within the 7th-grade to high school range and none at the college level or higher. Public health and prevention education ranked second, with a relatively balanced distribution of reading levels and a higher proportion in the lower grade ranges. Perioperative patient education had the largest number of entries and the widest range of reading levels, with high school level dominating. Chronic disease management education presents the greatest reading difficulty, with most records at the high school level, and it was the only category to include graduate-level [47]. Pharmacotherapy safety and adherence education had a small sample size, with readability generally concentrated at the high school level and grades 8-9.

XSL•FO
RenderX

Table 2. The best and worst readability of each study.

Author, year	Classification of PEMs[a]	The lowest grade levels	The highest grade levels
Aayush et al, 2025 [42]	Chronic disease management education	10th to 12th Grade	Postgraduate
Mohammed et al, 2025 [43]	Specialty disease diagnostic and therapeutic education	10th to 12th Grade	College
Qais et al, 2025 [44]	Specialty disease diagnostic and therapeutic education	7th Grade	10th to 12th Grade
Jamila et al, 2025 [45]	Perioperative patient education	• CTR[b]: 10th to 12th Grade • THA[c]: 10th to 12th Grade • TKR[d]: 10th to 12th Grade	• CTR: Postgraduate • THA: Postgraduate • TKR: Postgraduate
Ali et al, 2025 [46]	Specialty disease diagnostic and therapeutic education	10th to 12th Grade	College
Nuran et al, 2025 [47]	Chronic disease management education	Postgraduate	Postgraduate
Polly et al, 2025 [48]	Perioperative patient education	7th Grade	College
Gowri et al, 2025 [49]	Public health and prevention education	• PGTE[e]: 10th to 12th Grade • TE[f]: 7th Grade • BT[g]: 7th Grade • OCS[h]: 8th and 9th Grade	• PGTE: 10th to 12th Grade • TE: 10th to 12th Grade • BT: College • OCS: College
Christoph et al, 2025 [50]	Pharmacotherapy safety and adherence education	10th to 12th Grade	Postgraduate
Jing et al, 2025 [51]	Chronic disease management education	10th to 12th Grade	College
Kirk et al, 2025 [52]	Perioperative patient education	College	College
Shinya et al, 2025 [53]	Public health and prevention education	10th to 12th Grade	10th to 12th Grade
Hatice et al, 2025 [54]	Perioperative patient education	8th and 9th Grade	8th and 9th Grade
Allison et al, 2025 [55]	Specialty disease diagnostic and therapeutic education	10th to 12th Grade	Postgraduate
Jaziya et al, 2025 [56]	Chronic disease management education	10th to 12th Grade	College
Mi et al, 2025 [57]	Perioperative patient education	• LD[i]: 8th and 9th Grade • SF[j]: 7th Grade • DL[k]: 8th and 9th Grade	• LD: Postgraduate • SF: Postgraduate • DL: Postgraduate
Shuai et al, 2025 [58]	Perioperative patient education	College	College
Nithin et al, 2025 [59]	Pharmacotherapy safety and adherence education	10th to 12th Grade	College
Qais et al, 2025 [60]	Specialty disease diagnostic and therapeutic education	8th and 9th Grade	10th to 12th Grade
Mohammad et al, 2025 [61]	Public health and prevention education	7th Grade	10th to 12th Grade
Yuwan et al, 2025 [62]	Specialty disease diagnostic and therapeutic education	6th Grade	10th to 12th Grade
George et al, 2025 [63]	Specialty disease diagnostic and therapeutic education	6th Grade	College
Qais et al, 2025 [64]	Chronic disease management education	8th and 9th Grade	10th to 12th Grade
Salahaldin et al, 2025 [65]	Specialty disease diagnostic and therapeutic education	10th to 12th Grade	Postgraduate

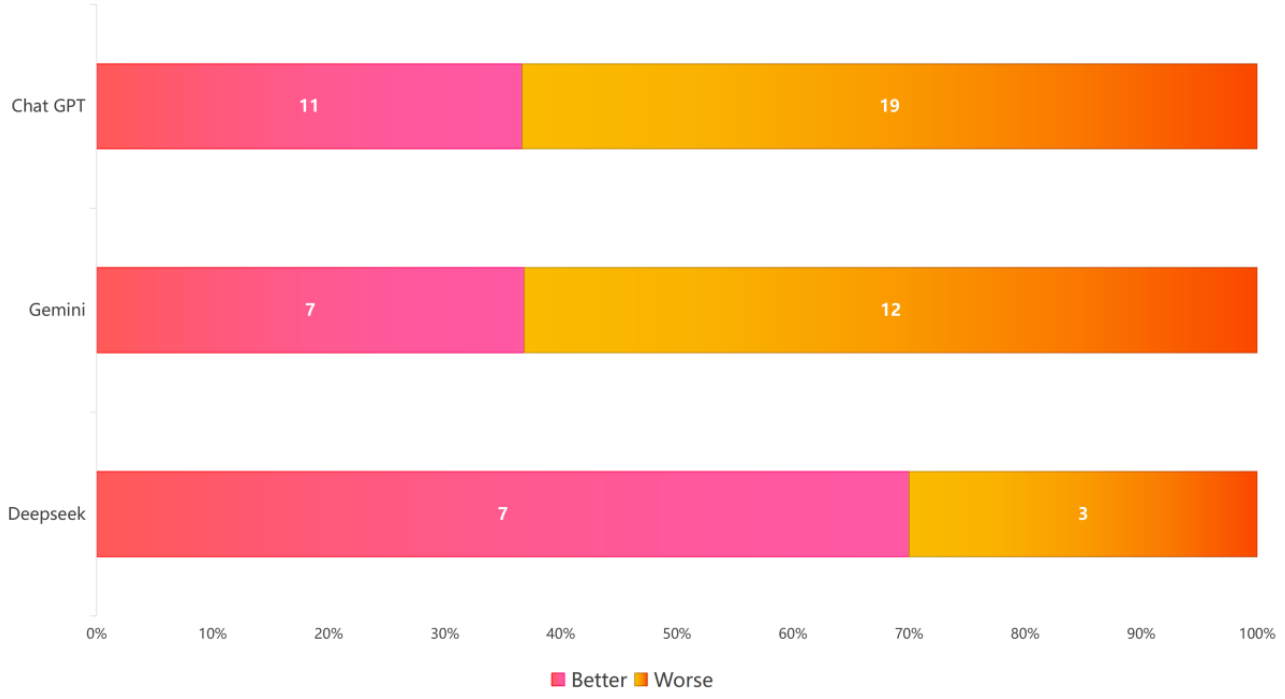
XSL•FO
RenderX

Author, year	Classification of PEMs[a]	The lowest grade levels	The highest grade levels
Maha et al, 2025 [66]	Perioperative patient education	10th to 12th Grade	College
Gülseren et al, 2025 [67]	Perioperative patient education	10th to 12th Grade	College
Elizabeth et al, 2025 [68]	Pharmacotherapy safety and adherence education	8th and 9th Grade	8th and 9th Grade
Xuewei et al, 2025 [69]	Public health and prevention education	6th Grade	8th and 9th Grade
Qais Dihan et al, 2025 [70]	Specialty disease diagnostic and therapeutic education	7th Grade	10th to 12th Grade
Amruth A Alluri et al, 2025 [71]	Chronic disease management education	10th to 12th Grade	Postgraduate

[a]PEM: patient education material.
[b]CTR: carpal tunnel release.
[c]THA: total hip arthroplasty.
[d]TKA: total knee replacement.
[e]PGTE: postoperative guidance for tooth extraction.
[f]TE: tooth extraction.
[g]BT: brushing technique.
[h]OCS: oral cancer screening.
[i]LD: lumbar discectomy.
[j]SF: spinal fusion.
[k]DL: decompressive laminectomy.

In addition, for studies evaluating multiple LLMs, we compared the performance of LLMs within the same record, as shown in Figure 4. The DeepSeek series demonstrated the most consistent performance and has outperformed other models in most comparisons. ChatGPT and Gemini, on the other hand, performed poorly in most records, with readability levels generally being too high.

Figure 4. Comparison of readability between models.



Quality

Similarly, we extracted all data on quality assessment metrics in Multimedia Appendix 4, with a total of 22 records included in quality assessments [42-44,46,48,50-52,56-62,64-66,70,71], all of which used the DISCERN tool or a modified version. Based on the classification of PEMs, we present the quality levels of 5 major categories of PEMs, as shown in Figure 5.

XSL•FO
RenderX

Overall, the quality of PEMs was generally good, with most records rated as good or higher. Specifically, the highest proportion of high-quality records was found in specialty disease diagnostic and therapeutic education, where the quality distribution was relatively balanced. Perioperative patient education had the largest number of entries, with 5 records rated as good. In chronic disease management education, the quality of records was predominantly good. The sample size for pharmacotherapy safety and adherence education was small, but both records were rated as excellent [59,68], demonstrating outstanding performance. In addition, we compared the performance of LLMs: the ChatGPT series demonstrated a significant advantage, the Gemini series performed consistently, and the DeepSeek series lagged slightly behind, as shown in Figure 6.

Figure 5. Quality based on five major disease categories.

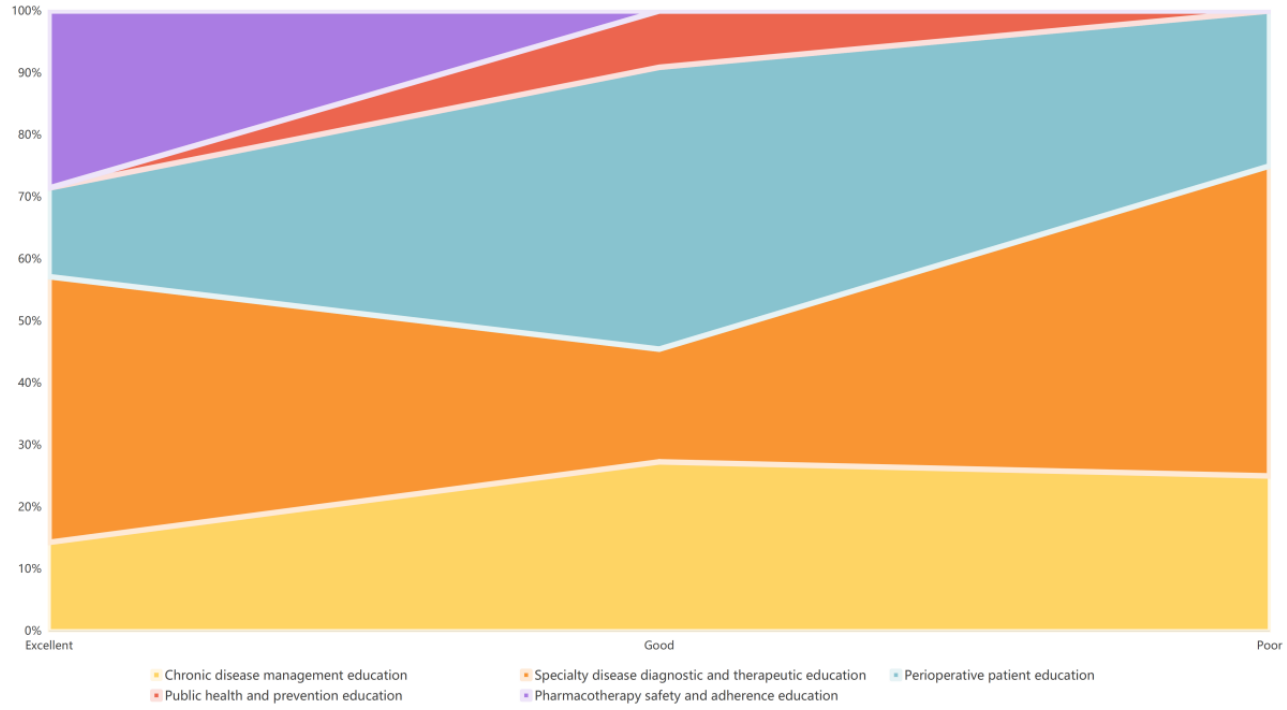
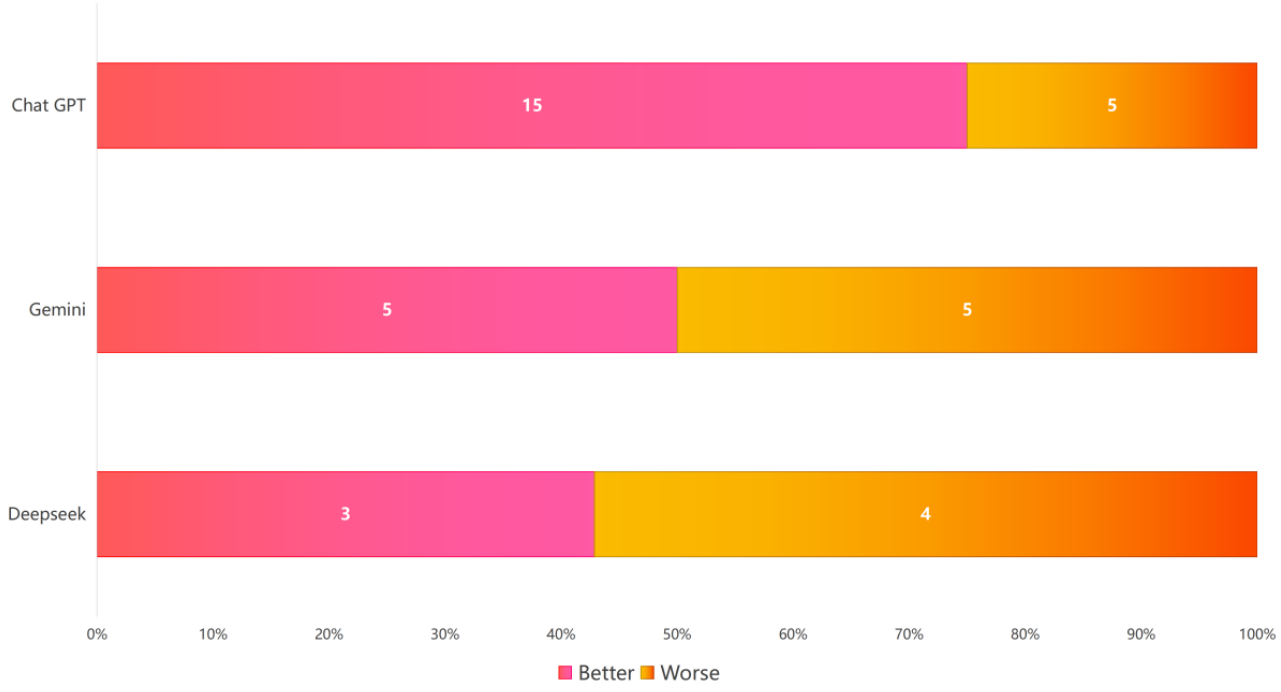


Figure 6. Quality based on five major disease categories.



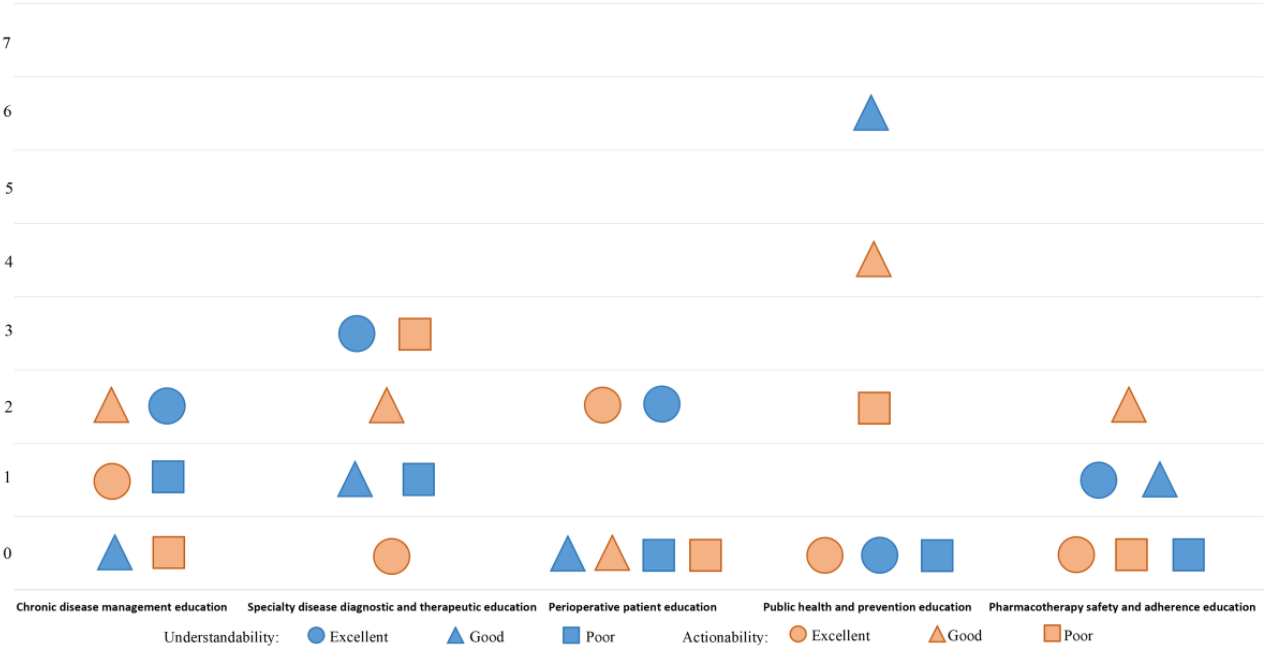
XSL•FO
RenderX

Usability

Similarly, we extracted all data on usability assessment metrics in Multimedia Appendix 5; a total of 18 records included usability assessments [44,46,47,49,50,53,54,56,60-62,64,67], all of which used the Patient Education Materials Assessment Tool (PEMAT) or Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P) scales. Based on the classification of PEMs, we present the usability levels of 5 major categories of PEMs, as shown in Figure 7.

Overall, most PEMs performed well in terms of understandability, but there were widespread shortcomings in terms of actionability. Specifically, perioperative patient education performed best, with 2 records rated as excellent for both understandability and actionability [54,67]. Specialty disease diagnostic and therapeutic education was primarily rated as excellent for understandability, but rated as poor in actionability in many cases. Chronic disease management education generally scored higher on understandability than on actionability, with most records rated as excellent in understandability. Pharmacotherapy safety and adherence education showed balanced performance, with all records rated as good in actionability. For public health and preventive education, both understandability and actionability were mostly rated as good, though 2 records were rated as poor in actionability [53,61].

Figure 7. Usability based on five major disease categories.



Discussion

Principal Findings

In this study, we conducted a comprehensive assessment of the current state of LLMs in generating new PEMs. We found that the LLMs in this field are rapidly expanding, driven primarily by the ChatGPT family of models. Although LLMs have demonstrated the ability to generate well-structured content, there is a significant disparity between the quality of their outputs and established health literacy standards. Specifically, only 10% of the included studies reached the reading level recommended by the NIH, and materials in most disease categories consistently lacked actionability. Furthermore, only half of the studies used standardized usability assessment tools, highlighting the fragmentation of current PEMs evaluation systems.

Our study suggests that while LLMs improve the efficiency of developing PEMs, without optimization, direct application of LLMs may inadvertently exacerbate health inequities, because the generated materials are often too complex for patients to understand or translate into healthy behaviors.

Gap Between Readability and Health Literacy

Aligning the readability of PEMs with patients' health literacy is the foundation to achieve health benefits [73]. Although the NIH recommends a sixth-grade reading level, we found that the vast majority of LLM-generated materials exceeded this threshold [28,29,31,36]. Notably, we observed significant heterogeneity across different disease categories. Specialty disease diagnostic and therapeutic education demonstrated the best readability, with most studies falling within the 7th-grade to high school level, while chronic disease management education had poor readability. This discrepancy may stem from the inherent complexity of chronic disease management, which involves subtle lifestyle changes and multifactorial decision-making, making it difficult for LLMs to develop PEMs without losing critical context [74-76]. Ma et al [77] presented a similar conclusion.

We believe that the readability dilemma stems less from limitations in the model capabilities and more from a lack of scientifically sound implementation strategies [78,79]. Excessively high reading difficulty not only hinders access to information but may also exacerbate health inequalities among groups with lower socioeconomic status [16,17]. Future

XSL•FO
RenderX

developments should not focus solely on improving model performance or relying on default outputs; instead, they must use targeted prompt engineering to explicitly specify a sixth-grade reading level and conduct rigorous manual verification after generation to improve the readability of information [4,22,78,80]. Given that primary health care settings have a better understanding of patients' health literacy and digital preferences, they should play an important role in the development and evaluation process of PEMs, proactively enhance patients' digital literacy, and guide them to participate in the design [37,81,82].

Disconnect Between Knowledge and Action

We found that although most studies scored highly on "understandability" in terms of information delivery, they were generally deficient in "actionability," particularly in the fields of specialized disease diagnostic and therapeutic education and public health and prevention education. This finding highlights a deeper problem: current LLMs are essentially good at retrieving and reorganizing static knowledge, but have difficulty dynamically translating it into health behaviors tailored to specific contexts [28,36,83,84]. This limitation is particularly detrimental in chronic disease management, where the classic knowledge-attitude-practice (KAP) theory emphasizes that patient self-efficacy and behavior change are crucial [85].

This lack of adaptability stems from the model's lack of contextual awareness (perceiving the patient's real-life situations), which prevents LLMs from generating personalized PEMs, resulting in information that is theoretically correct but practically ineffective [28,86]. Many studies suggest that materials generated by LLMs should be treated as "drafts" that must be customized based on the patient's medical history, lifestyle, and living environment to be truly useful [17,87-89]. Engaging patients in the co-design of PEMs can help bridge this gap [81,90-92]. Specifically, through cognitive interviews, user testing, or iterative feedback, patients can identify comprehension and behavioral obstacles, cultural and linguistic preferences based on their own lived contexts, thereby helping PEMs move from being theoretically correct to contextually actionable [90,93].

Standardization of Evaluation System

The high heterogeneity among the included studies hinders the consistency of our conclusions. This review found that the included studies used as many as 12 different readability assessment tools, and half did not use quality assessment tools (such as PEMAT), making it extremely difficult to compare LLMs' performance across studies and to reach a unified clinical consensus [32,36]. Although LLMs generally outperform social media in terms of reliability, they still cannot match official guidelines in complex clinical decision-making such as surgical management, and carry a risk of information lag [45,46,68,94].

Furthermore, performance disparities between different model versions have raised fairness concerns. If high-quality and readable PEMs require a paid subscription, this could widen the digital divide rather than bridge it, thereby harming the interests of vulnerable groups [95,96]. Much evidence suggests that the most promising application of LLMs in patient education remains the human-machine collaboration model, rather than the deployment of autonomous AI [88,97-99].

Implications for Practice and Policy

Based on this, we propose the following 3 recommendations. First, clinical workflows should be restructured. Physicians should adopt the principle of treating LLMs as "drafting assistants" rather than "final authors," and involve patients in the design and evaluation of PEMs. Second, promote both standardized evaluation and technical optimization. Establish a unified evaluation framework covering readability, actionability, and accuracy. At the same time, institutions should promote the integration of prompt engineering and retrieval-augmented generation (RAG) technology to address information lag. Finally, prioritize equity and accessibility. Model developers and clinicians should recognize that model performance varies depending on disease complexity. Materials related to chronic diseases may require more rigorous optimization than those for specialized conditions.

Limitations

This study has several limitations. First, we did not conduct a formal quality assessment of the included studies, and the findings should not be interpreted as conclusions about the strength of specific LLMs or specific clinical scenarios. Second, inclusion was restricted to English-language publications and English-language databases, which may systematically exclude relevant studies from non-English–speaking regions and limit the applicability of our findings in different health care systems and cultural backgrounds. Third, we did not consider grey literature and conference papers. Given the rapid iteration of LLMs and the long delay in the official publication of the literature, this time lag could underestimate the current scope of LLM applications in patient education. Fourth, although we provided the readability range, the conclusion is based on best-case readability, which reflects an upper limit under ideal conditions, and may overestimate the readability of LLMs in the real world. Therefore, the "best potential" cannot be regarded as "conventional performance" and generalized as a general conclusion. Fifth, the rapid evolution of LLMs limits the long-term validity of our conclusions regarding comparisons between specific models [100].

Conclusions

LLMs offer a transformative opportunity to produce effective and high-quality PEMs. However, current applications often fail to meet established health literacy standards, particularly in terms of readability and actionability. To realize their full potential, future efforts must shift from mere capability assessment to optimizing implementation—focusing on prompt engineering, patient participation, human supervision, and retrieval-based accuracy validation. Only through rigorous validation and equitable deployment can LLMs safely bridge the gap between complex medical knowledge and patient understanding, thereby truly empowering patients in their health management.

XSL•FO
RenderX

Acknowledgments

We thank the Soochow University School of Medicine Library for guiding us in developing an accurate search strategy.

Funding

The research was supported by the National Natural Science Fund of China under Grant (82171834); Research Project on Hospital Management Innovation of Jiangsu Hospital Association under Grant (JSYGY-3-2024-262).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Generative AI Use

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), GenAI tools were used under full human supervision for typo checking and language polishing. The GenAI tools used was DeepSeek (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co, Ltd). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Authors' Contributions

Conceptualization: YC, ML
Data curation: CL, JL, XL
Formal analysis: CL, YC
Funding acquisition: ML
Investigation: CL, JL, XL
Methodology: CL, YX, YC
Project administration: ZG, YX
Resources: XX, XL
Software: CL, XL
Supervision: ZG, XX
Validation: YC, ML
Visualization: CL, JL
Writing – original draft: CL
Writing – review & editing: CL, JL, YC, ML

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-ScR checklist.
[DOCX File , 67 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search strategy.
[DOCX File , 26 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Extraction data of readability of included literature.
[DOCX File , 35 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Extraction data of quality of included literature.
[DOCX File , 19 KB-Multimedia Appendix 4]

XSL•FO
RenderX

Multimedia Appendix 5

Extraction data of usability of included literature.
[DOCX File , 22 KB-Multimedia Appendix 5]

References

1. Slatore CG, Kulkarni HS, Corn J, Sockrider M. Improving health literacy. The new American Thoracic Society guidelines for patient education materials. Ann Am Thorac Soc. 2016;13(8):1208-1211. [FREE Full text] [doi: 10.1513/AnnalsATS.201605-337OT] [Medline: 27509148]
2. Prabhu AV, Gupta R, Kim C, Kashkoush A, Hansberry DR, Agarwal N, et al. Patient education materials in dermatology: addressing the health literacy needs of patients. JAMA Dermatol. 2016;152(8):946-947. [doi: 10.1001/jamadermatol.2016.1135] [Medline: 27191054]
3. Hadden K, Prince LY, Schnaekel A, Couch CG, Stephenson JM, Wyrick TO. Readability of patient education materials in hand surgery and health literacy best practices for improvement. The Journal of Hand Surgery. 2016;41(8):825-832. [doi: 10.1016/j.jhsa.2016.05.006]
4. Gu JZ, Baird GL, Escamilla Guevara A, Sohn Y, Lydston M, Doyle C, et al. A systematic review and meta-analysis of English language online patient education materials in breast cancer: is readability the only story? Breast. 2024;75:103722. [FREE Full text] [doi: 10.1016/j.breast.2024.103722] [Medline: 38603836]
5. Bhatt C, Lin E, Ferreira-Legere LE, Jackevicius CA, Ko DT, Lee DS, et al. Evaluating readability, inderstandability, and actionability of online printable patient education materials for cholesterol management: a systematic review. J Am Heart Assoc. 2024;13(8):e030140. [FREE Full text] [doi: 10.1161/JAHA.123.030140] [Medline: 38567668]
6. Karimi AH, Shah AK, Hecht CJ, Burkhart RJ, Acuña AJ, Kamath AF. Readability of online patient education materials for total joint arthroplasty: a systematic review. J Arthroplasty. 2023;38(7):1392-1399. [doi: 10.1016/j.arth.2023.01.032] [Medline: 36716898]
7. Armache M, Assi S, Wu R, Jain A, Lu J, Gordon L, et al. Readability of patient education materials in head and neck cancer: a systematic review. JAMA Otolaryngol Head Neck Surg. 2024;150(8):713-724. [doi: 10.1001/jamaoto.2024.1569] [Medline: 38900443]
8. Badarudeen S, Sabharwal S. Assessing readability of patient education materials: current role in orthopaedics. Clin Orthop Relat Res. 2010;468(10):2572-2580. [FREE Full text] [doi: 10.1007/s11999-010-1380-y] [Medline: 20496023]
9. Padilla G, Awshah S, Mhaskar RS, Diab AF, Sujka JA, DuCoin C, et al. Spanish-language bariatric surgery patient education materials fail to meet healthcare literacy standards of readability. Surg Endosc. 2023;37(8):6417-6428. [doi: 10.1007/s00464-023-10088-9] [Medline: 37129638]
10. Hanci V, Otlu B, Biyiko?lu A. Assessment of the readability of the online patient education materials of intensive and critical care societies. Crit Care Med. 2024;52(2):e47-e57. [doi: 10.1097/ccm.0000000000006121]
11. Jiang X, Liu J, Dai S, Mao X, Cha R, Wu L. Evaluating the effectiveness of generative AI for the creation of patient education materials on coronary heart disease: a comparative study. JMIR Form Res. 2025;9:e78816-e78816. [FREE Full text] [doi: 10.2196/78816] [Medline: 41270264]
12. Shah YB, Ghosh A, Hochberg AR, Rapoport E, Lallas CD, Shah MS, et al. Comparison of ChatGPT and traditional patient education materials for men's health. Urol Pract. 2024;11(1):87-94. [doi: 10.1097/upj.0000000000000490]
13. Chen L, Orosa D, Ahn G, Putnam J, Layoun H, Saddiqui HU, et al. Are adult thoracic patient education materials designed with patients in mind? J Thorac Cardiovasc Surg. 2024;167(5):1654-1656.e5. [FREE Full text] [doi: 10.1016/j.jtcvs.2023.12.029] [Medline: 38199291]
14. Singh A, Gupta N, Chien DL, Singh R, Sachdeva A, Danasekaran K, et al. ChatGPT-4 in neurosurgery: improving patient education materials. Neurosurgery. 2026;98(1):147-160. [doi: 10.1227/neu.0000000000003606] [Medline: 40704789]
15. Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? Clin Orthop Relat Res. 2023;481(11):2260-2267. [doi: 10.1097/CORR.0000000000002668] [Medline: 37116006]
16. Rouhi AD, Ghanem YK, Yolchieva L, Saleh Z, Joshi H, Moccia MC, et al. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? A pilot study. Cardiol Ther. 2024;13(1):137-147. [FREE Full text] [doi: 10.1007/s40119-023-00347-0] [Medline: 38194058]
17. Daram NR, Maxwell RA, D'Amato J, Massengill JC. Can artificial intelligence improve the readability of patient education information in gynecology? Am J Obstet Gynecol. 2025;233(6):640.e1-640.e9. [FREE Full text] [doi: 10.1016/j.ajog.2025.06.047] [Medline: 40578780]
18. Betzler BK, Chen H, Cheng C, Lee CS, Ning G, Song SJ, et al. Large language models and their impact in ophthalmology. The Lancet Digital Health. 2023;5(12):e917-e924. [doi: 10.1016/s2589-7500(23)00201-7]
19. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
20. Rodler S, Cei F, Ganjavi C, Checcucci E, De Backer P, Rivero Belenchon I, et al. GPT-4 generates accurate and readable patient education materials aligned with current oncological guidelines: a randomized assessment. PLoS One. 2025;20(6):e0324175. [FREE Full text] [doi: 10.1371/journal.pone.0324175] [Medline: 40465696]

XSL•FO
RenderX

21. Tucker CA. Promoting personal health literacy through readability, understandability, and actionability of online patient education materials. JAHA. 2024;13(8):e33916. [doi: 10.1161/jaha.124.033916]
22. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res. 2025;27:e69955. [FREE Full text] [doi: 10.2196/69955] [Medline: 40465378]
23. Falade IO, Wilson JC, Mehari ME, Soroudi D, Song S, Kim EA. The complexity of online patient education materials for wound care strategies: a readability analysis. Surgery. 2024;176(2):324-330. [FREE Full text] [doi: 10.1016/j.surg.2024.04.014] [Medline: 38769036]
24. Matsubara S. Patient education materials by ChatGPT: real intention and interpretation. Acta Obstet Gynecol Scand. 2026;105(8):1636-1637. [doi: 10.1111/aogs.70123] [Medline: 41386744]
25. Çamur E, Cesur T, Güneş YC. Large language models in radiology reporting: bridging semantics, education, and safety. Eur J Radiol. 2025;192:112417. [doi: 10.1016/j.ejrad.2025.112417] [Medline: 40957119]
26. Flores-Gouyonnet J, Cuéllar-Gutiérrez MC, Figueroa-Parra G, Kimbrough B, Joerns EK, Navarro-Mendoza E, et al. Performance of large language models in rheumatology board-like questions: accuracy, quality, and safety. The Lancet Rheumatology. 2025;7(3):e152-e154. [doi: 10.1016/s2665-9913(24)00400-4]
27. Heston TF. Safety of large language models in addressing depression. Cureus. 2023;15(12):e50729. [FREE Full text] [doi: 10.7759/cureus.50729] [Medline: 38111813]
28. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res. 2024;26:e22769. [FREE Full text] [doi: 10.2196/22769] [Medline: 39509695]
29. Zhang Z, Zhang H, Pan Z, Bi Z, Wan Y, Song X, et al. Evaluating large language models in ophthalmology: systematic review. J Med Internet Res. 2025;27:e76947. [FREE Full text] [doi: 10.2196/76947] [Medline: 41144954]
30. Zhong R, Chen S, Li Z, Gao T, Su Y, Zhang W, et al. Large language models in lung cancer: systematic review. J Med Internet Res. 2025;27:e74177-e74177. [FREE Full text] [doi: 10.2196/74177] [Medline: 41026980]
31. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. 2024;11:e57400. [FREE Full text] [doi: 10.2196/57400] [Medline: 39423368]
32. Ferreira SJ, Ladeiras-Lopes R, Leite F, Dores H. Applications of large language models in cardiovascular disease: a systematic review. Eur Heart J Digit Health. 2025;6(4):540-553. [FREE Full text] [doi: 10.1093/ehjdh/ztaf028] [Medline: 40703130]
33. Mansoor I, Abdullah M, Rizwan MD, Fraz MM. Reasoning with large language models in medicine: a systematic review of techniques, challenges and clinical integration. Health Inf Sci Syst. 2026;14(1):6. [doi: 10.1007/s13755-025-00403-0] [Medline: 41323158]
34. 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. [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
35. Patwardhan V, Balchander D, Fussell D, Joseph J, Joshi A, Troutt H, et al. Leveraging large language models to enhance radiology report readability: a systematic review. J Am Coll Radiol. 2026;23(3):354-361. [doi: 10.1016/j.jacr.2025.09.004] [Medline: 40945554]
36. Woo BFY, Cato K, Cho H, You SB, Song J. The use of large language models in clinical documentation: a scoping review. Int J Nurs Stud. 2026;176:105322. [FREE Full text] [doi: 10.1016/j.ijnurstu.2025.105322] [Medline: 41512731]
37. Ryan P, Davoren O, Elwyn G. Fact-checking large language model responses to a health care prompt: comparative study. JMIR Form Res. 2026;10:e68223-e68223. [FREE Full text] [doi: 10.2196/68223] [Medline: 41985066]
38. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
39. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467-473. [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
40. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
41. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105-111. [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
42. Chaulagain A, Aujla S, Priyadarsini A, Godavarthi A, Yaqoob U. A cross-sectional comparison of patient information guides generated by ChatGPT versus google gemini for alzheimer's disease, parkinsonism, and migraine. Cureus. 2025;17(5):e84507. [doi: 10.7759/cureus.84507] [Medline: 40546497]
43. Hussain M, Hassan M, Taj T, Shah V. A cross-sectional study assessing the suitability of ChatGPT and DeepSeek AI for generating patient education guides on imaging modalities in stroke. Cureus. 2025;17(9):e92066. [doi: 10.7759/cureus.92066] [Medline: 40980783]

XSL•FO
RenderX

44. Dihan QA, Brown AD, Zaldivar AT, Chauhan MZ, Eleiwa TK, Hassan AK, et al. Advancing patient education in idiopathic intracranial hypertension: the promise of large language models. Neurol Clin Pract. 2025;15(1):e200366. [doi: 10.1212/CPJ.0000000000200366] [Medline: 39399571]
45. Jido JT, Wizni A. Are artificial intelligence-generated patient leaflets ready for clinical use? A readability comparison across common orthopaedic procedures. J Orthop Case Rep. 2025;15(10):295-301. [doi: 10.13107/jocr.2025.v15.i10.6240] [Medline: 41078673]
46. Alabdalhussein A, Singhania N, Nadeem S, Talib M, Al-Domaidat D, Jimoh I, et al. Artificial intelligence versus professional standards: a cross-sectional comparative study of GPT, Gemini, and ENT UK in delivering patient information on ENT conditions. Diseases. 2025;13(9):286. [FREE Full text] [doi: 10.3390/diseases13090286] [Medline: 41002722]
47. Aypar AN. Assessing artificial intelligence-generated patient educational material on gestational diabetes mellitus: content and quality evaluation. J Perinat Neonatal Nurs. 2025;39(3):210-217. [doi: 10.1097/jpn.0000000000000905]
48. Thompson P, Thornton R, Ramsden CM. Assessing chatbots ability to produce leaflets on cataract surgery: Bing AI, ChatGPT 3.5, ChatGPT 4o, ChatSonic, Google Bard, Perplexity, and Pi. J Cataract Refract Surg. 2025;51(5):371-375. [doi: 10.1097/j.jcrs.0000000000001622] [Medline: 39885649]
49. Sivaramakrishnan G, Almuqahwi M, Ansari S, Lubbad M, Alagamawy E, Sridharan K. Assessing the power of AI: a comparative evaluation of large language models in generating patient education materials in dentistry. BDJ Open. 2025;11(1):59. [FREE Full text] [doi: 10.1038/s41405-025-00349-1] [Medline: 40533491]
50. Stephenson-Moe CA, Behers BJ, Gibons RM, Behers BM, Jesus Herrera LD, Anneaud D, et al. Assessing the quality and readability of patient education materials on chemotherapy cardiotoxicity from artificial intelligence chatbots: an observational cross-sectional study. Medicine (Baltimore). 2025;104(15):e42135. [FREE Full text] [doi: 10.1097/MD.0000000000042135] [Medline: 40228277]
51. Huang J, Wu Z. Comparative analysis of ChatGPT and google gemini in creating patient education materials for lower extremity varicose veins and deep vein thrombosis. Phlebology. 2026;41(5):451-454. [doi: 10.1177/02683555251375253] [Medline: 40882985]
52. Lower K, Lin J, Jenkin D, Campbell C, Seth I, Morris M, et al. Comparing the quality and readability of ChatGPT-4-generated vs. human-generated patient education materials for total knee arthroplasty. Cureus. 2025;17(6):e86491. [doi: 10.7759/cureus.86491] [Medline: 40693057]
53. Ito S, Furukawa E, Okuhara T, Okada H, Kiuchi T. Comparison of Japanese mpox (Monkeypox) health education materials and texts created by artificial intelligence: cross-sectional quantitative content analysis study. JMIR AI. 2025;4:e70604. [FREE Full text] [doi: 10.2196/70604] [Medline: 41105953]
54. Yüceler Kaçmaz H, Kahraman H, Akutay S, Dağdelen D. Development and validation of an artificial intelligence-assisted patient education material for ostomy patients: a methodological study. J Adv Nurs. 2025;81(7):3859-3867. [doi: 10.1111/jan.16542] [Medline: 39422196]
55. Kufta AY, Djalilian AR. Enhancing patient education with AI: a readability analysis of AI-generated versus American Academy of Ophthalmology online patient education materials. J Clin Med. 2025;14(19). [FREE Full text] [doi: 10.3390/jcm14196968] [Medline: 41096048]
56. Jabeen J, Saji J. Evaluating AI-Generated patient education guides: a comparative study of ChatGPT and Deepseek. Cureus. 2025;17(6):e85277. [doi: 10.7759/cureus.85277] [Medline: 40612853]
57. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. Int J Med Inform. 2025;198:105871. [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.105871] [Medline: 40107040]
58. Guo S, Yang C, Lin X, Jiang M, Chen J, Tuo K, et al. Evaluating artificial intelligence-generated patient education materials for bariatric surgery: comparative analysis of response quality, reliability, and readability across ChatGPT and deepSeek models. Obes Surg. 2025;35(11):4628-4638. [doi: 10.1007/s11695-025-08249-x] [Medline: 41014443]
59. Karnan N, Nair S, Fidai F, Gurrala S, Salim J, Gomma A. Evaluating the efficacy of ChatGPT vs. Google Gemini in generating patient education materials for GLP-1 receptor agonists (semaglutide, liraglutide, tirzepatide): a cross-sectional study. Cureus. 2025;17(4):e81993. [doi: 10.7759/cureus.81993] [Medline: 40351930]
60. Dihan QA, Brown AD, Zaldivar AT, Montgomery KE, Chauhan MZ, Abdelnaem SE, et al. Implementing generative AI to enhance patient education on retinopathy of prematurity. J Pediatr Ophthalmol Strabismus. 2025;62(6):443-452. [doi: 10.3928/01913913-20250515-01] [Medline: 40568824]
61. Delsoz M, Hassan A, Nabavi A, Rahdar A, Fowler B, Kerr NC, et al. Large language models: pioneering new educational frontiers in childhood myopia. Ophthalmol Ther. 2025;14(6):1281-1295. [doi: 10.1007/s40123-025-01142-x] [Medline: 40257570]
62. Gao Y, Xu Q, Zhang O, Wang H, Wang Y, Wang J, et al. Large language models: unlocking new potential in patient education for thyroid eye disease. Endocrine. 2025;90(2):689-698. [doi: 10.1007/s12020-025-04339-z] [Medline: 40682758]
63. Abdelmalek G, Uppal H, Garcia D, Farshchian J, Emami A, McGinniss A. Leveraging ChatGPT to produce patient education materials for common hand conditions. J Hand Surg Glob Online. 2025;7(1):37-40. [FREE Full text] [doi: 10.1016/j.jhsg.2024.10.002] [Medline: 39991597]

XSL•FO
RenderX

64. Dihan QA, Brown AD, Chauhan MZ, Alzein AF, Abdelnaem SE, Kelso SD, et al. Leveraging large language models to improve patient education on dry eye disease. Eye (Lond). 2025;39(6):1115-1122. [doi: 10.1038/s41433-024-03476-5] [Medline: 39681711]
65. Alamleh S, Mavedatnia D, Francis G, Le T, Davies J, Lin V, et al. Readability, reliability, and quality analysis of internet-based patient education materials and large language models on meniere's disease. J Otolaryngol Head Neck Surg. 2025;54:19160216251360651. [FREE Full text] [doi: 10.1177/19160216251360651] [Medline: 40776601]
66. Alsadaan M, Fahy S, Dan Milinkovic D, Bartek B, Winkler T, Jung T, et al. Refining patient education on autologous chondrocyte implantation for chondral lesions of the knee: a fine-tuned ChatGPT-4o model improves readability and quality. J Exp Orthop. 2025;12(4):e70445. [FREE Full text] [doi: 10.1002/jeo2.70445] [Medline: 41018948]
67. Maraş Baydoğan G, Sürme Y, Topan H. Understandability and actionability of artificial intelligence-assisted lymphedema education material in patients undergoing breast cancer surgery: expert evaluation. J Clin Nurs. 2026;35(10):4295-4304. [doi: 10.1111/jocn.70123] [Medline: 41077951]
68. Covington EW, Watts Alexander CS, Sewell J, Hutchison AM, Kay J, Tocco L, et al. Unlocking the future of patient education: ChatGPT vs. LexiComp® as sources of patient education materials. J Am Pharm Assoc (2003). 2025;65(1):102119. [doi: 10.1016/j.japh.2024.102119] [Medline: 38729608]
69. Li X, Zhang Y, Zheng T, Deng Y, Lu Y, Hu J, et al. Using large language models to generate child-friendly education materials on myopia. Digit Health. 2025;11:20552076251362338. [FREE Full text] [doi: 10.1177/20552076251362338] [Medline: 40755959]
70. Dihan Q, Chauhan MZ, Eleiwa TK, Hassan AK, Sallam AB, Khouri AS, et al. Using large language models to generate educational materials on childhood glaucoma. Am J Ophthalmol. 2024;265:28-38. [doi: 10.1016/j.ajo.2024.04.004] [Medline: 38614196]
71. Alluri AA, Khan Z, Krithika V, Jalil M, Dantu G, Annapureddy K, et al. Assessing the suitability of ChatGPT and DeepSeek AI for patient education on common rheumatological disorders. Cureus. 2025;17(8):e90600. [doi: 10.7759/cureus.90600] [Medline: 40984935]
72. Barbateskovic M, Larsen LK, Oxenbøll-Collet M, Jakobsen JC, Perner A, Wetterslev J. Pharmacological interventions for delirium in intensive care patients: a protocol for an overview of reviews. Syst Rev. 2016;5(1):211. [FREE Full text] [doi: 10.1186/s13643-016-0391-5] [Medline: 27923397]
73. Bizer B, Garcia Valencia O, Arriola-Montenegro J, Thongprayoon C, Miao J, Craici IM, et al. Generative AI and large language model evaluation of kidney donation websites: benchmarking health literacy, sentiment, and digital engagement to optimize donor recruitment. Digit Health. 2026;12:20552076261437915. [FREE Full text] [doi: 10.1177/20552076261437915] [Medline: 41907358]
74. Li C, Zhao Y, Bai Y, Zhao B, Tola YO, Chan CW, et al. Unveiling the potential of large language models in transforming chronic disease management: mixed methods systematic review. J Med Internet Res. 2025;27:e70535. [FREE Full text] [doi: 10.2196/70535] [Medline: 40239198]
75. Zhang J, Sun K, Jagadeesh A, Falakaflaki P, Kayayan E, Tao G, et al. The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. J Am Med Inform Assoc. 2024;31(9):1884-1891. [doi: 10.1093/jamia/ocae184] [Medline: 39018498]
76. Bergling K, Wang LC, Shivakumar O, Nandorine Ban A, Moore L, Ginsberg N, et al. From bytes to bites: application of large language models to enhance nutritional recommendations. Clin Kidney J. 2025;18(4):sfaf082. [FREE Full text] [doi: 10.1093/ckj/sfaf082] [Medline: 40226366]
77. Ma H, Cui H, Yu H, Pei R, Li S, Wang A, et al. Development and evaluation of an large language model-integrated chatbot intervention for physical activity habit formation in adults with prehypertension. Digit Health. 2026;12:20552076261421367. [FREE Full text] [doi: 10.1177/20552076261421367] [Medline: 41696083]
78. Asfuroğlu ZM, Yağar H, Gümüşoğlu E. High accuracy but limited readability of large language model-generated responses to frequently asked questions about Kienböck's disease. BMC Musculoskelet Disord. 2024;25(1):879. [FREE Full text] [doi: 10.1186/s12891-024-07983-0] [Medline: 39497130]
79. Cacciamani GE, Layne E, Asmundo MG, Russo GI. Bridging the gap: the role of large language model refinement in readability in urology research. BJU Int. 2025;136(3):356-358. [doi: 10.1111/bju.16774] [Medline: 40387322]
80. Butler JJ, Harrington M, Tong Y, Rosenbaum A, Samsonov AP, Walls RJ, et al. From jargon to clarity. improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. Foot & Ankle Orthopaedics. 2024;9(4):331-337. [doi: 10.1177/2473011424s00301]
81. Super I, Bawa H, Asan O. Exploring the role of large language models in primary care: qualitative study of physicians in the United States and the Netherlands. JMIR Med Inform. 2026;14:e91652. [FREE Full text] [doi: 10.2196/91652] [Medline: 42430718]
82. Stephenson CR, Edakkanambeth Varayil J, Aakre CA, Croghan IT, Hurt RT. Prompting strategies for large language models in primary care: a primer for clinician-artificial intelligence interaction. J Prim Care Community Health. 2026;17:21501319261455393. [FREE Full text] [doi: 10.1177/21501319261455393] [Medline: 42403380]

XSL•FO
RenderX

83. Low YS, Jackson ML, Hyde RJ, Brown RE, Sanghavi NM, Baldwin JD, et al. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. Digit Health. 2025;11:20552076251348850. [FREE Full text] [doi: 10.1177/20552076251348850] [Medline: 40510193]
84. Ma J, Zhang Y, Tang H, Yi X, Zhong T, Li X, et al. Evaluating the quality of large language model-generated preoperative patient education material: a comparative study across models and surgery types. Front Med (Lausanne). 2025;12:1701344. [FREE Full text] [doi: 10.3389/fmed.2025.1701344] [Medline: 41458519]
85. Alaofè H, Hounkpatin WA, Djrolo F, Ehiri J, Rosales C. Knowledge, attitude, practice and associated factors among patients with type 2 diabetes in Cotonou, Southern Benin. BMC Public Health. 2021;21(1):339. [FREE Full text] [doi: 10.1186/s12889-021-10289-8] [Medline: 33579243]
86. Boyle A, Huo B, Sylla P, Calabrese E, Kumar S, Slater BJ, et al. Large language model-generated clinical practice guideline for appendicitis. Surg Endosc. 2025;39(6):3539-3551. [doi: 10.1007/s00464-025-11723-3] [Medline: 40251310]
87. Ho CM, Guan S, Mok PK, Lam CH, Ho WY, Mak CH, et al. Development and validation of a large language model-powered chatbot for neurosurgery: mixed methods study on enhancing perioperative patient education. J Med Internet Res. 2025;27:e74299. [FREE Full text] [doi: 10.2196/74299] [Medline: 40663377]
88. Goudrar R, Zekraoui O, Moussa I, Nguyen D, Bouhadana D, Li T, et al. Large language model chatbots for patient education in kidney stones: a scoping review. World J Urol. 2025;43(1):641. [doi: 10.1007/s00345-025-06019-z] [Medline: 41160267]
89. Başaranoğlu M, Akbay E, Erdem E. From digital assistants to clinical partners: revolutionizing pediatric urology through large language model-powered decision support and patient education. World J Urol. 2025;43(1):606. [doi: 10.1007/s00345-025-05990-x] [Medline: 41074935]
90. Yin J, Wang W, Wu L, Li Z, Wang W, He T, et al. Structured clinical input–guided large language model workflow for acute ischemic stroke discharge education: a multicenter feasibility study. Stroke. 2026;57(9):2709-2718. [doi: 10.1161/strokeaha.126.055189]
91. Jing L, Ye B, Liu S. Evaluation of an AI-assisted digital health follow-up system integrating humanistic care for patients undergoing chemotherapy: a prospective quasi-experimental study. Front. Public Health. 2026;14:1863027. [doi: 10.3389/fpubh.2026.1863027]
92. Hu S, Lin B, Lin F. Evaluation of large language model-generated information in diabetes health patient education: a scoping review. Front Public Health. 2026;14:1850486. [FREE Full text] [doi: 10.3389/fpubh.2026.1850486] [Medline: 42415766]
93. Kivity S, Barilan YM, Noham R, Saban M. Using a large language model to support thematic analysis of patient experiences in chronic illness management: comparative qualitative study. J Med Internet Res. 2026;28:e88677. [FREE Full text] [doi: 10.2196/88677] [Medline: 42372263]
94. Klang E, Apakama D, Abbott EE, Vaid A, Lampert J, Sakhuja A, et al. A strategy for cost-effective large language model use at health system-scale. NPJ Digit Med. 2024;7(1):320. [FREE Full text] [doi: 10.1038/s41746-024-01315-1] [Medline: 39558090]
95. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023;6(1):210. [FREE Full text] [doi: 10.1038/s41746-023-00958-w] [Medline: 37973919]
96. Feldman MJ, Hoffer EP, Conley JJ, Chang J, Chung JA, Jernigan MC, et al. Dedicated AI expert system vs generative AI with large language model for clinical diagnoses. JAMA Netw Open. 2025;8(5):e2512994. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.12994] [Medline: 40440012]
97. Barile J, Margolis A, Cason G, Kim R, Kalash S, Tchaconas A, et al. Diagnostic accuracy of a large language model in pediatric case studies. JAMA Pediatr. 2024;178(3):313-315. [doi: 10.1001/jamapediatrics.2023.5750] [Medline: 38165685]
98. Wu H, Su Z, Pan X, Shao A, Xu Y, Wang Y, et al. Enhancing diabetic retinopathy query responses: assessing large language model in ophthalmology. Br J Ophthalmol. 2025;109(11):1272-1278. [FREE Full text] [doi: 10.1136/bjo-2024-325861] [Medline: 40588331]
99. Srinivasan S, Chen Q, Tham YC. Context for large language model evaluation in ophthalmology-reply. JAMA Ophthalmol. 2026;144(3):288. [doi: 10.1001/jamaophthalmol.2025.5847] [Medline: 41569551]
100. Nasra M, Jaffri R, Pavlin-Premrl D, Kok HK, Khabaza A, Barras C, et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. Intern Med J. 2025;55(1):20-34. [doi: 10.1111/imj.16607] [Medline: 39720869]

Abbreviations

GenAI: generative AI
KAP: knowledge-attitude-practice
LLM: large language model
NIH: National Institutes of Health
PCC: population, concept, and context
PEM: patient education material

XSL•FO
RenderX

PEMAT: Patient Education Materials Assessment Tool
PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RAG: retrieval-augmented generation

Edited by I Steenstra; submitted 14.Apr.2026; peer-reviewed by K Satasiya, G Nzeigwe, F Legare; comments to author 25.Aug.2026; revised version received 21.Sep.2026; accepted 14.Oct.2026; published 06.Nov.2026

Please cite as:
Li C, Li J, Liu X, Gao Z, Xu Y, Xu X, Liu X, Zhao T, Cheng Y, Li M
Large Language Models for Generating Patient Education Materials: Scoping Review
J Med Internet Res 2026;28:e98084
URL: https://www.jmir.org/2026/1/e98084
doi: 10.2196/98084
PMID:

©Chengjin Li, Juan Li, Xuezhu Liu, Zefang Gao, Yuxiao Xu, Xiuze Xu, Xiaoyu Liu, Tong Zhao, Yuanyuan Cheng, Ming Li. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 06.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX