International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	ARTIFICIAL INTELLIGENCE CHATBOTS IN PATIENT EDUCATION: OPPORTUNITIES, RISKS, AND ETHICAL CHALLENGES
DOI	https://doi.org/10.31435/ijitss.2(50).2026.5829
RECEIVED	12 March 2026
ACCEPTED	15 May 2026
PUBLISHED	28 May 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

ARTIFICIAL INTELLIGENCE CHATBOTS IN PATIENT EDUCATION: OPPORTUNITIES, RISKS, AND ETHICAL CHALLENGES

Wiktoria Sęk *(Corresponding Author, Email: sekwitoria04@gmail.com)*
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0001-3452-7025

Agata Kowalczyk
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0000-9027-7270

Magdalena Fąfara
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0001-8524-9870

Margarita Jeczkało
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0003-1103-090X

Monika Kosiorowska
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0002-5447-6365

Joanna Lasoń
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0004-7463-7626

Marcelina Makulska
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0008-2075-106X

Weronika Płachta
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0003-2460-9435

Amelia Sieradzka
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0005-9588-3065

Cezary Wegenko
Collegium Medicum Jan Kochanowski University in Kielce, Kielce, Poland
ORCID ID: 0009-0005-4920-7963

ABSTRACT

Artificial intelligence chatbots and large language models are increasingly used to obtain, simplify, and personalize medical information for patients. This narrative medical review synthesizes 40 recent studies on chatbot-supported patient education across emergency medicine, cardiology, oncology, ophthalmology, gastroenterology, dermatology, orthopedics, otolaryngology, urology, pain medicine, nursing, and health-literacy research. The review evaluates opportunities, risks, and ethical challenges associated with patient-facing AI communication. The evidence suggests that chatbots can generate fluent explanations, improve access to plain-language materials, support appointment preparation, and reduce the workload of drafting educational content. However, important limitations recur across the literature, including excessive reading level, incomplete or unsafe advice, weak source transparency, fabricated references, model variability, language-related inequity, and patient overtrust in fluent but unverified answers. Ethically acceptable implementation should therefore treat chatbots as supervised communication-support tools rather than autonomous medical educators. Clinical use should require clinician-approved source material, explicit AI disclosure, readability testing, escalation instructions, model-version documentation, multilingual validation, and post-deployment monitoring. Under such safeguards, AI chatbots may strengthen patient education, but unsupervised deployment may intensify existing problems in online medical information.

KEYWORDS

Artificial Intelligence, Chatbots, Patient Education, Health Literacy, Large Language Models; Medical Ethics

CITATION

Wiktoria Sęk, Agata Kowalczyk, Magdalena Fąfara, Margarita Jeczkało, Monika Kosiorowska, Joanna Lasoń, Marcelina Makulska, Weronika Płachta, Amelia Sieradzka, Cezary Wegenko. (2026) Artificial Intelligence Chatbots in Patient Education: Opportunities, Risks, and Ethical Challenges. *International Journal of Innovative Technologies in Social Science*. 2(50). doi: 10.31435/ijitss.2(50).2026.5829

1. Introduction

Patient education is a core clinical function because it affects symptom recognition, informed consent, treatment adherence, self-management, and shared decision-making (Kim et al., 2025). The emergence of conversational artificial intelligence has changed the information environment in which this education occurs because patients can now receive immediate, personalized, and dialogue-like medical explanations outside clinical encounters (Armbruster et al., 2024). The relevance of this change extends beyond technology adoption because chatbot answers can influence how patients interpret symptoms, prepare for appointments, and decide whether to seek urgent care (Yau et al., 2024).

The patient-facing use of AI chatbots is especially important because conventional online patient education materials frequently exceed recommended reading levels and may not be actionable for people with limited health literacy (Will et al., 2025). Several studies suggest that large language models can rewrite complex clinical information into shorter, more coherent, and more patient-oriented explanations when appropriately prompted (Sudharshan et al., 2024). This potential aligns with a broader public-health need to make medical information understandable, timely, and accessible across languages, diseases, and care settings (Michalska et al., 2025).

The clinical value of chatbot-based patient education depends on more than linguistic fluency because a useful educational tool must also be accurate, complete, cautious, readable, actionable, and transparent about uncertainty (Tiller et al., 2026). Evidence from oncology, cardiology, emergency care, ophthalmology, dermatology, gastroenterology, and otolaryngology shows that chatbot outputs can be acceptable for many routine questions but inconsistent when clinical complexity or safety-sensitive decisions increase (Yau et al., 2024). Because patient education often occurs at moments of anxiety or vulnerability, even small omissions or overly reassuring statements can have disproportionate consequences (Armbruster et al., 2024).

An additional concern is that the conversational form of chatbots can simulate empathy, certainty, and authority even when the underlying answer lacks sufficient evidence or omits appropriate escalation advice

(Ayers et al., 2023). Patients may evaluate an answer as trustworthy because it is fluent, personalized, and responsive, whereas clinicians may identify clinical risks that are not obvious to lay users (Armbruster et al., 2024). This divergence between perceived helpfulness and clinical reliability creates an ethical challenge for autonomy, nonmaleficence, accountability, and trust in health systems (Tiller et al., 2026).

This review examines AI chatbots in patient education as a socio-technical medical problem rather than only a software-performance problem (Ayers et al., 2023). The aim is to synthesize evidence on opportunities, risks, and ethical challenges in patient-facing chatbot communication, with particular attention to clinical accuracy, readability, health literacy, misinformation, trust, equity, and implementation safeguards (Tiller et al., 2026). The topic fits the scope of innovative technologies in social science because chatbot-mediated health communication reshapes relationships among patients, clinicians, institutions, technology providers, and public-health information systems (Armbruster et al., 2024).

2. Methodology

A narrative medical review design was used because the included literature is methodologically heterogeneous and includes cross-sectional evaluations, comparative experiments, quasi-experimental studies, methodological studies, readability analyses, content-quality assessments, and patient or clinician perception studies (Golan et al., 2023). The review did not aim to calculate a pooled effect size because the studies differed in clinical domain, model version, prompt design, comparator, rating scale, outcome definition, and whether patient or expert raters were used (Will et al., 2025).

Eligible publications addressed artificial intelligence chatbots or large language models in relation to patient-facing health information, patient education materials, medical question answering, readability, health literacy, clinical communication, misinformation, or ethical risk (Ayre et al., 2024). The evidence base was intentionally broad because patient education is not confined to one specialty and because risks differ across urgent-care questions, chronic-disease self-management, procedural counseling, cancer education, and public-health communication (Yau et al., 2024).

The included studies were extracted into a structured evidence table containing author, year, clinical context, model or chatbot focus, educational task, comparator where applicable, principal outcomes, and contribution to the review question (Ayers et al., 2023). Findings were synthesized thematically into domains of clinical accuracy and completeness, readability and health literacy, patient experience and trust, misinformation and source transparency, model variability and equity, ethical governance, and clinical implementation (Tiller et al., 2026).

The main limitation of this method is that narrative synthesis cannot eliminate publication bias, model-version bias, prompt-specific bias, or heterogeneity in outcome assessment (Golan et al., 2023). This limitation is particularly relevant for AI studies because model behavior can change after deployment, repeated generations may differ, and prompt wording can alter readability, tone, and completeness (Zhang et al., 2025). The synthesis therefore emphasizes convergent patterns and implementation implications rather than treating any single study as definitive evidence of safety or effectiveness (Yau et al., 2024).

3. Results

The included studies show that AI chatbots are being evaluated across a wide range of patient education settings, including emergency care, hypertension, atrial fibrillation, obstructive sleep apnea, prostate cancer, lung cancer, breast cancer surgery, cervical cancer screening, inflammatory bowel disease, glaucoma, otolaryngology, urological malignancies, dermatology, orthopedic trauma, hand conditions, chronic pain medication, ostomy care, rheumatology, and plastic surgery (Yau et al., 2024). This diversity indicates that chatbot-based education is not a specialty-specific phenomenon but a cross-cutting innovation in medical communication (Armbruster et al., 2024).

The evidence base also shows that studies often evaluate different dimensions of patient education that should not be conflated, including factual accuracy, completeness, readability, understandability, actionability, empathy, perceived trust, safety, referencing, and clinical appropriateness (Golan et al., 2023). A chatbot can perform well on one dimension while performing poorly on another, which means that implementation decisions require multidimensional evaluation rather than a single overall score (Golan et al., 2023). Table 1 summarizes the clinical context and main contribution of the 40 included publications (Tiller et al., 2026).

Table 1. Included studies and main contribution to the medical review

Citation	Clinical education context	Main contribution
Ayers et al., 2023	General medical questions	AI answers were rated highly for quality and empathy, supporting supervised communication use.
Armbruster et al., 2024	Patient perspective	Patients valued usefulness and empathy, but specialist review identified harmful advice.
Yau et al., 2024	Emergency care	Chatbot answers varied in accuracy, completeness, source reporting, and safety.
Tiller et al., 2026	Misinformation-prone topics	Multiple chatbots produced problematic answers, weak references, and fabricated citations.
Laily et al., 2026	HPV vaccine information	Responses were generally evidence-based but limited by jargon, reading level, and links.
Gibson et al., 2024	Prostate cancer	ChatGPT-4 performed well on quality and appropriateness, with readability and citation concerns.
Gencer, 2024	Lung cancer	Lung cancer responses remained at college-level difficulty across repeated generations.
Richlitzki et al., 2025	Lung cancer radiotherapy	Clinician and patient perspectives were favorable, but readability and usability limits remained.
Almagazzachi et al., 2024	Hypertension	Most answers to common hypertension questions were appropriate and reproducible.
Lee et al., 2024	Atrial fibrillation	Most responses were correct, but references were rarely supplied and readability was high.
Kim et al., 2025	Heart attack education	AI and human conversations improved knowledge, while human conversation was stronger for some outcomes.
Sudharshan et al., 2024	Text simplification	ChatGPT reduced text complexity, with language- and material-specific limitations.
Ayre et al., 2024	Health literacy	Simplification preserved many key messages but often remained above ideal reading grade.
Kirchner et al., 2023	Orthopedics	AI rewriting reduced readability level without reviewer-detected inaccuracies.
Zalzal et al., 2024	Otolaryngology	Most answers were medically correct, although lay trust was lower than physician ratings.
Thia & Saluja, 2024	Urological malignancies	Age-targeted prompting improved accessibility relative to default outputs.
Abdelmalek et al., 2025	Hand conditions	GPT-4 performed better than GPT-3.5, but many outputs exceeded sixth-grade targets.
Gondode et al., 2024	Chronic pain medications	Traditional leaflets were more readable in some medication contexts than chatbot materials.
Will et al., 2025	Online PEM rewriting	Several LLMs improved readability while occasional inaccuracies persisted.
Kacmaz et al., 2025	Ostomy education	AI-assisted materials achieved high understandability and actionability despite remaining readability limits.
Roster et al., 2024	Dermatology	Sequential prompting improved readability, especially for sunscreen and melanoma education.
Stephenson-Moe et al., 2025	Chemotherapy cardiotoxicity	Materials were high quality but frequently required college-to-graduate reading ability.
Zhang et al., 2025	Inflammatory bowel disease	Multiple models handled basic education but showed variability for complex advice.
Campbell et al., 2023	Obstructive sleep apnea	Most answers were correct or nearly correct, but readability remained too high.
Michalska et al., 2025	Polish screening education	Polish cervical-screening outputs were coherent but required high formal education.
Rossettini & Palese, 2026	Accessibility critique	Accuracy was framed as insufficient when outputs are not understandable or usable.
Mendoza-Pinto et al., 2025	Rheumatology literature	LLMs improved accessibility of specialist literature but did not reach ideal patient levels.
Rodler et al., 2025	Oncological guidelines	GPT-4 generated readable guideline-aligned oncology materials in randomized assessment.

Citation	Clinical education context	Main contribution
Dejenie et al., 2026	Plastic surgery research	Several LLMs improved accessibility of research summaries while performance differed by model.
Sriram et al., 2026	Breast cancer surgery	Simplified prompting reduced reading level while maintaining content quality.
Kufta & Djalilian, 2025	Ophthalmology	Default LLM materials were harder than AAO materials, but optimized prompts helped.
Naufal et al., 2025	Breast cancer brochures	ChatGPT and Gemini generated comparable brochures, with different complexity profiles.
Lv et al., 2024	Patient access/self-management	Expert and AI-generated content differed in appropriateness, readability, and stability.
Golan et al., 2023	Quality/readability evaluation	ChatGPT did not reliably replace established readability tools or human DISCERN ratings.
Covarrubias et al., 2025	Trauma orthopedics	Readability improved substantially, but expert editing remained necessary.
He et al., 2025	IBD communication	ChatGPT was comparable for accuracy and more complete, but patients still valued doctors.
Wu et al., 2026	Ophthalmology handouts	ChatGPT-4o handouts compared favorably with AAO materials, including Spanish versions.
Todorov et al., 2025	Dermatology leaflets	Outputs were not misleading but missed ideal readability and completeness targets.
Yan et al., 2025	IBD doctor comparison	Overall quality was similar to specialists, but complex integrated questions remained challenging.
Cohen et al., 2024	Glaucoma	ChatGPT was more accurate than Google but harder to read before simplification.

Source: own elaboration based on the included studies.

Clinical accuracy and completeness

Several studies reported that chatbots can provide clinically acceptable responses to common, well-bounded patient questions when the requested information is general rather than individualized (Almagazzachi et al., 2024). Hypertension, obstructive sleep apnea, atrial fibrillation, otolaryngology, glaucoma, and inflammatory bowel disease studies suggest that many routine explanations are correct or nearly correct, although conclusions differ by model, prompt, and evaluation method (Campbell et al., 2023).

Accuracy becomes less stable when questions require triage, individualized judgment, source evaluation, or integration of symptoms with personal context (Yau et al., 2024). Emergency-care evaluations showed variability in accuracy, completeness, source reporting, and safety across four free-access chatbots, which is clinically important because urgent-care answers can influence whether patients seek immediate help (Yau et al., 2024). In inflammatory bowel disease, chatbot performance was promising for general education but more fragile when questions required nuanced doctor-patient integration or treatment-specific judgment (Yau et al., 2024).

Oncology studies reinforce this mixed picture by showing both guideline-aligned educational potential and persistent risks related to completeness, readability, and citation quality (Richlitzki et al., 2025). GPT-4 generated oncology materials that were readable and aligned with current guidelines in a randomized assessment, whereas prostate cancer and lung cancer evaluations still identified concerns regarding readability, hallucinated references, or patient usability (Rodler et al., 2025).

These findings suggest that AI chatbots are strongest when the educational message is already established and the task is to translate, summarize, or organize that message for patients (Kirchner et al., 2023). They are weakest when asked to operate as autonomous clinical reasoners, safety triage agents, or sources of individualized advice without clinician oversight (Yau et al., 2024).

Readability, health literacy, and accessibility

Readability was the most recurrent limitation across the reviewed literature, even when content was judged clinically acceptable (Gencer, 2024). ChatGPT responses on lung cancer remained at a college-level reading difficulty, atrial fibrillation responses remained difficult despite prompt variation, and chemotherapy cardiotoxicity materials often required college-to-graduate reading ability (Gencer, 2024).

A consistent opportunity is that prompt-based rewriting can reduce reading level and make patient education materials more accessible (Ayre et al., 2024). Orthopedic, online health information, trauma, breast cancer surgery, and ophthalmology studies suggest that large language models can simplify existing materials

rapidly, although the quality of simplification depends on model, prompt, language, and clinical domain (Will et al., 2025).

Simplification does not automatically guarantee health literacy because patients need actionable, complete, and contextually meaningful information rather than only shorter sentences (Ayre et al., 2024). Ostomy education research showed high understandability and actionability after AI-assisted development, but other studies showed that outputs can remain above recommended levels or require expert editing even after simplification (Kacmaz et al., 2025).

The accessibility problem is also visible in specialty-specific contexts where technically accurate material remains too difficult for the general population (Thia & Saluja, 2024). Urological malignancy, hand-condition, and ophthalmology studies indicate that age-targeted or readability-targeted prompting can improve access but rarely removes the need for clinician or plain-language review (Abdelmalek et al., 2025).

Readability should therefore be treated as a patient-safety metric rather than a cosmetic editing outcome (Rossettini & Palese, 2026). A response that is accurate but incomprehensible may fail to support informed decision-making and may widen disparities for patients with lower literacy, anxiety, limited schooling, or unfamiliarity with medical terminology (Rossettini & Palese, 2026).

Patient experience, empathy, and trust

Conversational AI can improve perceived access to health information because it provides immediate explanations, adaptable tone, and patient-specific phrasing (Armbruster et al., 2024). In a comparison of chatbot and physician responses to public patient questions, chatbot responses were rated higher for quality and empathy, suggesting that language generation can meet important communication needs that traditional online materials often leave unmet (Ayers et al., 2023).

Patient experience studies also show that perceived usefulness is not the same as clinical safety (Armbruster et al., 2024). Patients may rate chatbot answers positively even when specialists identify potentially harmful advice, and patients may still prefer doctors for some treatment-related questions even when AI answers are complete and readable (Armbruster et al., 2024).

Educational benefit is more likely when chatbots are used as complements to human communication rather than replacements for it (Kim et al., 2025). Quasi-experimental evidence in heart attack education showed that both AI and human-delivered conversations improved knowledge, but human conversation produced stronger gains in some outcomes and remains important for social support and judgment (Kim et al., 2025).

The trust issue is ethically important because chatbot fluency can mask uncertainty and because patients may not know whether an answer is based on current evidence, institutional policy, or generated probability (Tiller et al., 2026). A safe patient education system must therefore explain the role of AI, disclose limitations, identify urgent symptoms, and direct users to clinicians when individual decisions or emergencies are involved (Yau et al., 2024).

Misinformation, referencing, and epistemic transparency

The most serious epistemic risk is that chatbots can produce confident medical language without reliable source grounding (Tiller et al., 2026). An audit of generative AI chatbots in misinformation-prone health topics found problematic answers, weak referencing, and fabricated citations, which directly undermines patient ability to verify medical information (Tiller et al., 2026).

Referencing limitations recur even in clinical areas where response accuracy appears otherwise acceptable (Lee et al., 2024). Atrial fibrillation and prostate cancer studies found that correct or useful answers could still lack adequate references or include hallucinated citations, indicating that source quality should be evaluated separately from answer fluency (Gibson et al., 2024).

A related problem is that chatbots may not reliably assess the quality or readability of medical information when asked to act as evaluators (Golan et al., 2023). Cross-sectional evidence showed that ChatGPT did not consistently match established readability tools or human DISCERN ratings, which means that AI self-evaluation should not replace validated assessment or expert review (Golan et al., 2023).

Misinformation risk is particularly relevant for public-health topics such as vaccination, cancer screening, and online controversy because patients may combine chatbot answers with social media narratives or prior misinformation (Laily et al., 2026). HPV-vaccine and cervical-screening studies show that AI systems can generate coherent and broadly evidence-based messages, but accessibility, links, jargon, and reading level still influence whether information supports informed decisions (Laily et al., 2026).

Model variability, multilinguality, and equity

Model selection and model version influence patient education quality, which makes results difficult to generalize across platforms and time (Zhang et al., 2025). Comparisons among ChatGPT, Google Gemini, Claude, Bard, Grok, and other models show differences in readability, appropriateness, stability, and content complexity, even when the educational task is similar (Lv et al., 2024).

Multilingual performance is a major equity issue because patient education must be understandable in the language patients actually use (Michalska et al., 2025). Polish cervical cancer screening information was generally coherent but required high formal education, while ophthalmology handout studies showed that Spanish outputs can perform well when specifically evaluated rather than assumed to be equivalent to English (Michalska et al., 2025).

Equity also concerns who benefits from immediate AI explanations and who is most exposed to harm from unclear, incomplete, or overconfident advice (Rossettini & Palese, 2026). Patients with lower health literacy, chronic illness, limited digital skills, poor access to clinicians, or anxiety after diagnosis may gain from accessible explanations but may also be less able to identify unsafe advice or invented references (Tiller et al., 2026).

A socially responsible implementation strategy should therefore test chatbot education in multiple languages, literacy levels, ages, diseases, and care contexts before deployment (Wu et al., 2026). Without such evaluation, AI patient education may reproduce the same inequities already present in online health information while adding new risks related to automation and perceived authority (Tiller et al., 2026).

4. Discussion

The reviewed evidence supports a balanced conclusion that AI chatbots are promising patient education tools but should not be treated as autonomous medical educators (Tiller et al., 2026). Their strongest value lies in communication support, including drafting explanations, simplifying validated materials, preparing patient questions, translating content, generating handouts, and organizing complex disease information (Will et al., 2025).

This interpretation is consistent across multiple specialties because the best-performing use cases involve transformation of already approved information rather than independent clinical decision-making (Rodler et al., 2025). When source material is clinically validated and outputs are reviewed by clinicians, chatbots can make health information more readable, personalized, and scalable (Kacmaz et al., 2025).

The main clinical risk is unsafe confidence because patients may receive fluent language that omits warning signs, overgeneralizes, or fails to recommend urgent care when needed (Yau et al., 2024). The risk is intensified by weak source transparency and hallucinated citations because patients and clinicians may be unable to distinguish reliable education from plausible but unsupported statements (Tiller et al., 2026).

The second major risk is a mismatch between generated text and patient health literacy (Rossettini & Palese, 2026). Even when a chatbot is accurate, patient education may fail if the information is written at a college level, lacks actionability, or does not clearly state when professional care is required (Lee et al., 2024).

From an ethical perspective, patient autonomy requires more than simply increasing the quantity of available information (Michalska et al., 2025). Autonomy is strengthened only when patients can understand the message, evaluate uncertainty, recognize the difference between general education and individualized advice, and know when to contact a clinician (Tiller et al., 2026).

Nonmaleficence requires explicit limits on chatbot scope because education can become advice when a patient asks about symptoms, medication effects, or treatment choices (Yau et al., 2024). Beneficence supports chatbot use when the tool reduces communication barriers, increases access to understandable materials, and helps patients prepare for clinical visits (Kim et al., 2025).

Justice requires that chatbot education be tested for populations that are usually underserved by traditional medical communication, including non-English speakers, patients with low health literacy, older adults, patients with chronic illness, and people with limited access to clinicians (Michalska et al., 2025). Accountability requires that healthcare organizations know which model, prompt, source material, reviewer, language, and approval date produced the patient-facing content (Will et al., 2025).

Socio-technical implications for patient education

From a socio-technical perspective, the main innovation of patient-facing chatbots is not only automated text production but the redistribution of educational work across patients, clinicians, and healthcare organizations (Armbruster et al., 2024). When patients use chatbots before clinical encounters, the technology may shape the questions they ask, the confidence with which they interpret symptoms, and the expectations they bring to a consultation (Ayers et al., 2023). Evidence from heart-attack education shows that AI-based

conversation can increase knowledge, while the stronger performance of human-delivered conversation in some outcomes indicates that social interaction remains clinically meaningful (Kim et al., 2025). This means that chatbot-based patient education should be assessed as part of a communication infrastructure rather than as an isolated information product (Will et al., 2025).

Equity implications are especially important because digital health tools can make information easier to access while simultaneously excluding patients who have limited literacy, limited language concordance, or limited confidence in online information (Michalska et al., 2025). The Polish cervical-cancer screening study shows that coherent AI-generated text may still require a high level of formal education, which limits practical accessibility for some target users (Michalska et al., 2025). The ophthalmology handout study indicates that non-English outputs can perform well when they are specifically evaluated, but it also shows why equivalence across languages should not be assumed without testing (Wu et al., 2026). Readability-oriented health-literacy research supports the view that simplification must be accompanied by assessment of meaning, actionability, and suitability for the intended audience (Ayre et al., 2024). Patients with lower health literacy may benefit greatly from conversational explanations, but they may also be less able to detect unsupported claims, missing warnings, or invented references (Rossettini & Palese, 2026).

Ethical governance should therefore focus on source grounding, disclosure, scope limitation, and escalation rather than on a general declaration that chatbots are helpful or harmful (Tiller et al., 2026). Emergency-care evidence shows why open-ended deployment is risky, because questions that appear educational can quickly become symptom-triage questions requiring individualized clinical judgment (Yau et al., 2024). Prostate-cancer assessments demonstrate that useful answers may still contain limitations in completeness, referencing, or source quality, which makes expert review necessary even when the overall response appears clinically acceptable (Gibson et al., 2024). The finding that ChatGPT may not reliably evaluate the quality or readability of online medical information reinforces the need for independent assessment rather than AI self-certification (Golan et al., 2023). Documentation of the model, prompt, source material, language, and approval date is also required because outputs can vary across systems and over time (Naufal et al., 2025).

The most defensible early use cases involve transforming validated content into patient-facing language under human supervision (Rodler et al., 2025). Orthopedic evidence shows that AI revision can improve readability when the starting point is an existing educational material rather than an unconstrained request for clinical advice (Kirchner et al., 2023). Ostomy education research adds that understandability and actionability should be validated explicitly, because a text can be short and fluent without helping patients decide what to do (Kacmaz et al., 2025). Trauma education research supports the practical value of using AI to improve readability in procedural and post-injury contexts, but it also implies the need to check whether simplified wording preserves essential safety information (Covarrubias et al., 2025). Breast-cancer surgery and reconstruction materials show that content quality and readability must be evaluated together when AI-generated explanations address emotionally sensitive and preference-sensitive decisions (Sriram et al., 2026).

Evaluation should include both clinicians and patients because these groups judge different dimensions of quality (Richlitzki et al., 2025). Clinicians are better positioned to identify factual omissions, guideline inconsistency, and unsafe reassurance, whereas patients are better positioned to report whether the explanation is understandable, emotionally acceptable, and useful for preparing questions (Richlitzki et al., 2025). Inflammatory-bowel-disease evidence comparing gastroenterologist and ChatGPT responses illustrates that apparent answer quality must be interpreted in relation to the clinical role that the answer is expected to play (He et al., 2025). Quasi-experimental evidence in heart-attack education suggests that chatbot conversations may contribute to learning outcomes, but it does not remove the need for human counseling where anxiety, uncertainty, or treatment decisions are involved (Kim et al., 2025). Studies that frame the problem as whether ChatGPT can replace doctors are valuable because they clarify the boundary between communication assistance and professional responsibility (Yan et al., 2025).

Research methods in this field should become more transparent and reproducible because small technical choices can change the educational output that patients receive (Will et al., 2025). Readability studies on lung-cancer information show that repeated generations may remain too difficult for patients even when the answer appears organized and clinically plausible (Gencer, 2024). Chemotherapy-cardiotoxicity analyses demonstrate that readability and quality should be examined together because a clinically important topic can be compromised by both complexity and incomplete patient guidance (Stephenson-Moe et al., 2025). HPV-vaccine information studies demonstrate the need to test misinformation-sensitive public-health topics, because chatbot communication may influence decisions that affect both individual and community health

(Laily et al., 2026). Research-summary simplification in plastic surgery illustrates another promising use case, but it also shows why simplification must be checked against the original scientific meaning before patients or clinicians rely on it (Dejenie et al., 2026).

Clinical and ethical implementation framework

A practical governance model should classify chatbot education by clinical risk, patient vulnerability, and degree of personalization (Yau et al., 2024). Low-risk uses include rewriting approved leaflets, generating appointment-preparation questions, summarizing general disease information, and translating clinician-approved materials for review (Kirchner et al., 2023). Medium-risk uses include medication explanations, post-procedure instructions, and chronic-disease self-management content that requires explicit escalation advice and clinician approval (Gondode et al., 2024). High-risk uses include symptom triage, emergency advice, individualized treatment recommendations, and content for patients at high risk of misunderstanding or harm (Yau et al., 2024).

Table 2 translates the reviewed evidence into minimum safeguards for clinical integration (Tiller et al., 2026). The central principle is that the degree of human review should increase as the clinical stakes, personalization, vulnerability, and uncertainty increase (Armbruster et al., 2024).

Table 2. Clinical opportunities, risks, and safeguards in AI chatbot patient education

Use case	Opportunity	Risk	Minimum safeguard
Drafting approved educational material	Fast creation of summaries, FAQs, and handouts	Factual omissions or unsupported additions	Clinician-approved source base and expert review
Plain-language rewriting	Improved readability and accessibility	Loss of nuance or incomplete simplification	Readability testing plus fidelity review
Multilingual communication	Improved access for non-English-speaking patients	Translation errors and cultural mismatch	Native-language clinical and patient validation
Appointment preparation	Better patient questions and shared decision-making	Overtrust in generic advice	AI disclosure and instruction to discuss decisions with clinicians
Medication or post-procedure education	Structured self-management support	Unsafe omissions about adverse effects or warning signs	Escalation instructions and clinician approval
Symptom or emergency questions	Rapid orientation when patients are uncertain	False reassurance and delayed care	Avoid autonomous triage and require urgent-care prompts
Source-supported public-health education	Scalable response to misinformation	Fabricated references or weak links	Verified citations and institutional review
Research-summary simplification	Accessible explanation of specialist literature	Oversimplification or distorted implications	Expert check against original source and intended audience

Source: own elaboration based on the evidence synthesis.

Practical recommendations for healthcare organizations

First, early deployment should be limited to supervised and source-grounded tasks rather than open-ended patient counseling (Gibson et al., 2024). Health organizations should begin with clinician-approved source materials and use chatbots to produce plain-language summaries, FAQs, discharge-support explanations, or appointment-preparation prompts (Kirchner et al., 2023).

Second, every patient-facing chatbot output should include safety boundaries and escalation pathways (Yau et al., 2024). The educational message should state that it does not replace professional medical advice, identify symptoms or situations requiring urgent care, and encourage patients to contact clinicians for individualized decisions (Yau et al., 2024).

Third, readability should be measured before release and not assumed from model fluency (Gencer, 2024). Prompts should specify target audience, reading level, tone, length, and actionability, but outputs should still be tested with objective readability tools and expert review because prompt optimization does not guarantee comprehension (Will et al., 2025).

Fourth, source integrity should be treated as a clinical safety requirement rather than a formatting preference (Tiller et al., 2026). Patient-facing AI systems should cite only verified sources, avoid invented references, and make it clear when information is general, uncertain, or not applicable to individual clinical decisions (Tiller et al., 2026).

Fifth, implementation should include documentation and audit trails because model behavior may differ across tools and over time (Naufal et al., 2025). A safe record should include the model name, version or access date, prompt, source material, language, reviewer, approval date, and scope of intended use (Zhang et al., 2025).

Sixth, post-deployment monitoring should evaluate patient understanding, complaints, near misses, clinician feedback, and whether patients know when to seek help (Kim et al., 2025). The clinical success of chatbot education should be measured by patient comprehension, self-efficacy, appropriate care-seeking, and shared decision-making rather than only by text quality metrics (Kim et al., 2025).

Minimum quality assessment before deployment

Before a healthcare organization releases chatbot-generated patient education, evaluation should include more than a single demonstration that the model can answer a common question (Yau et al., 2024). A minimum test set should include routine informational questions, questions about warning signs, questions about adverse effects, questions that require referral to a clinician, and prompts that contain misinformation or misleading assumptions (Tiller et al., 2026).

Quality assessment should separate accuracy, completeness, readability, actionability, empathy, and source reliability because these dimensions may move in different directions (Stephenson-Moe et al., 2025). A response may be accurate but too difficult to read, readable but incomplete, empathic but unsafe, or well-structured but weakly referenced (Tiller et al., 2026).

Testing should include both clinician review and patient-facing assessment because experts and lay users evaluate different aspects of quality (Richlitzki et al., 2025). Clinicians can identify factual omissions, unsafe advice, and guideline inconsistency, whereas patients can evaluate perceived clarity, usefulness, emotional tone, and whether the answer supports action (Richlitzki et al., 2025).

Readability assessment should combine automated metrics with human evaluation because formulas alone do not determine whether patients can apply information in real situations (Golan et al., 2023). Human review is needed to determine whether simplified text preserves essential meaning, avoids misleading reassurance, and states what patients should do next (Kacmaz et al., 2025).

Model stability should be assessed through repeated generations and documentation of the exact model environment (Gencer, 2024). This is necessary because patient education materials generated on different days, by different systems, or under different prompt instructions may vary in complexity, completeness, and clinical caution (Zhang et al., 2025).

A deployment evaluation should also test edge cases rather than only typical questions because patient harm often occurs at the boundary between education and advice (Yau et al., 2024). Edge-case prompts should include urgent symptoms, medication interactions, ambiguous complaints, vulnerable-patient scenarios, culturally sensitive topics, and requests for certainty when uncertainty should be disclosed (Yau et al., 2024).

Patient education chatbots should be evaluated with a source-grounding requirement when they are deployed by healthcare institutions (Gibson et al., 2024). If a system cannot reliably provide verified references or link its answer to approved materials, its use should be restricted to drafting for clinician review rather than direct patient-facing delivery (Tiller et al., 2026).

The evaluation process should include equity testing because performance for English-language, well-educated, digitally skilled users may not generalize to patients with different languages or literacy levels (Michalska et al., 2025). Materials should be tested with target users or proxies for the target population whenever possible, especially for cancer screening, chronic disease, postoperative care, and medication education (Kacmaz et al., 2025).

A final approval decision should be tied to the intended use case rather than to a generic claim that a chatbot is safe or unsafe (Rodler et al., 2025). The same model may be acceptable for rewriting a validated leaflet, questionable for answering medication side-effect questions, and inappropriate for unsupervised emergency triage (Yau et al., 2024).

Quality assessment should be repeated after major model updates or changes in institutional source material because apparent safety at one time point may not remain stable (Naufal et al., 2025). Continuous review is especially important when patient-facing systems are used in high-volume settings such as oncology education, chronic disease management, or public-health communication (Richlitzki et al., 2025).

Research priorities and limitations

Future research should move beyond isolated output assessment toward patient-centered outcomes (Kim et al., 2025). Studies should evaluate whether patients understand information, remember warning signs, ask better questions, experience less decisional conflict, follow safer self-management steps, and seek appropriate care when needed (Kim et al., 2025).

Research should also test chatbot education in real workflows rather than only as stand-alone prompts (Kacmaz et al., 2025). A chatbot used before a visit, after discharge, in a patient portal, during chronic disease management, or in public-health outreach creates different responsibilities and risks (Almagazzachi et al., 2024).

Reporting standards for AI patient education studies should become more transparent because reproducibility depends on prompt text, model version, access date, temperature or generation settings, language, source material, number of generations, and evaluator background (Will et al., 2025). Without such reporting, results cannot be compared across models or safely translated into institutional practice (Dejenie et al., 2026).

This review is limited by heterogeneity in study design, specialty, model type, outcome measures, and publication date (Will et al., 2025). It is also limited by rapid technological change because the performance of a commercial chatbot evaluated in one month may not represent its later behavior after model updates or changes in guardrails (Naufal et al., 2025).

Despite these limitations, the convergence of findings across specialties strengthens the main conclusion that supervised use is more defensible than autonomous deployment (Tiller et al., 2026). The evidence consistently supports a model in which chatbots assist clinicians and patients with communication, while clinicians, institutions, and validated sources remain responsible for clinical quality and safety (Yau et al., 2024).

5. Conclusions

AI chatbots have substantial potential to improve patient education by making medical information more immediate, conversational, personalized, and adaptable to different literacy levels (Ayre et al., 2024). The most credible current uses are supervised communication tasks such as drafting, simplifying, translating, summarizing, and organizing clinician-approved information (Kirchner et al., 2023).

The reviewed evidence also shows that autonomous deployment would be clinically and ethically premature because outputs may be too difficult to read, incomplete, weakly sourced, fabricated in their references, or unsafe in urgent and individualized situations (Tiller et al., 2026). Patient satisfaction and perceived empathy should not be treated as proof of safety because patients may not reliably detect harmful or unsupported advice (Armbruster et al., 2024).

Ethically acceptable integration requires human oversight, source grounding, explicit AI disclosure, readability targets, multilingual validation, escalation instructions, model documentation, and ongoing monitoring (Tiller et al., 2026). Within such a governance framework, AI chatbots can become valuable instruments of patient-centered education, but without it they may intensify existing problems in online health information (Rossettini & Palese, 2026).

Funding: No external funding was declared for the preparation of this review manuscript.

Conflicts of Interest: No conflicts of interest to declare.

Ethics Approval: Not applicable. This narrative review used previously published studies and did not involve human participants, animals, or identifiable personal data.

Data Availability: All data synthesized in this review are derived from published full-text articles listed in the References section.

REFERENCES

1. Abdelmalek, G., Uppal, H., Garcia, D., Farshchian, J., Emami, A., & McGinniss, A. (2025). Leveraging ChatGPT to produce patient education materials for common hand conditions. *Journal of Hand Surgery Global Online, 7*, 37–40. https://doi.org/10.1016/j.jhsg.2024.10.002
2. Almagazzachi, A., Mustafa, A., Eighaei Sedeh, A., Vazquez Gonzalez, A. E., Polianovskaia, A., Abood, M., Abdelrahman, A., Muyolema Arce, V., Acob, T., & Saleem, B. (2024). Generative artificial intelligence in patient education: ChatGPT takes on hypertension questions. *Cureus, 16*(2), e53441. https://doi.org/10.7759/cureus.53441
3. Armbruster, J., Bussmann, F., Rothhaas, C., Titze, N., Grutzner, P. A., & Freischmidt, H. (2024). Doctor ChatGPT, can you help me? The patient's perspective: Cross-sectional study. *Journal of Medical Internet Research, 26*, e58831. https://doi.org/10.2196/58831
4. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine, 183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838
5. Ayre, J., Mac, O., McCaffery, K., McKay, B. R., Liu, M., Shi, Y., Rezwan, A., & Dunn, A. G. (2024). New frontiers in health literacy: Using ChatGPT to simplify health information for people in the community. *Journal of General Internal Medicine, 39*(4), 573–577. https://doi.org/10.1007/s11606-023-08469-w
6. Campbell, D. J., Estephan, L. E., Mastrolonardo, E. V., Amin, D. R., Huntley, C. T., & Boon, M. S. (2023). Evaluating ChatGPT responses on obstructive sleep apnea for patient education. *Journal of Clinical Sleep Medicine, 19*(12), 1989–1995. https://doi.org/10.5664/jcsm.10728
7. Cohen, S. A., Fisher, A. C., Xu, B. Y., & Song, B. J. (2024). Comparing the accuracy and readability of glaucoma-related question responses and educational materials by Google and ChatGPT. *Journal of Current Glaucoma Practice, 18*(3), 111–115. https://doi.org/10.5005/jp-journals-10078-1448
8. Covarrubias, O., Ghanem, D., Murdock, C., Domes, C., & Shafiq, B. (2025). Improving the readability of trauma patient education materials: A ChatGPT solution demonstrated using materials by the Orthopaedic Trauma Association. *OTA International, 8*(2), e417. https://doi.org/10.1097/OI9.0000000000000417
9. Dejenie, R. A., Gantt, B. K., Alexis, M. M., Shockley, S. M., Foote, K., Woolfolk, M. S., Banyard, D., & Butler, P. D. (2026). Bridging the gap: A pilot study using artificial intelligence to make plastic surgery research accessible. *Plastic and Reconstructive Surgery—Global Open, 14*, e7539. https://doi.org/10.1097/GOX.0000000000007539
10. Gencer, A. (2024). Readability analysis of ChatGPT's responses on lung cancer. *Scientific Reports, 14*, 17234. https://doi.org/10.1038/s41598-024-67293-2
11. Gibson, D., Jackson, S., Shanmugasundaram, R., Seth, I., Siu, A., Ahmadi, N., Kam, J., Mehan, N., Thanigasalam, R., Jeffery, N., Patel, M. I., & Leslie, S. (2024). Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: Multimetric assessment. *Journal of Medical Internet Research, 26*, e55939. https://doi.org/10.2196/55939
12. Golan, R., Ripps, S. J., Reddy, R., Loloi, J., Bernstein, A. P., Connelly, Z. M., Golan, N. S., & Ramasamy, R. (2023). ChatGPT's ability to assess quality and readability of online medical information: Evidence from a cross-sectional study. *Cureus, 15*(7), e42214. https://doi.org/10.7759/cureus.42214
13. Gondode, P., Duggal, S., Garg, N., Sethupathy, S., Asai, O., & Lohakare, P. (2024). Comparing patient education tools for chronic pain medications: Artificial intelligence chatbot versus traditional patient information leaflets. *Indian Journal of Anaesthesia, 68*(7), 631–636. https://doi.org/10.4103/ija.ija_204_24
14. He, Z., Liu, Y., Wang, Z., Zhou, X., Fan, X., He, M., Wang, C., Fu, H., Wang, Z., & Bai, Y. (2025). Revolutionizing inflammatory bowel disease healthcare communication: A head-to-head comparison of gastroenterologist and ChatGPT responses. *Gastroenterology Report, 13*, goaf087. https://doi.org/10.1093/gastro/goaf087
15. Kacmaz, H. Y., Kahraman, H., Akutay, S., & Dagdelen, D. (2025). Development and validation of an artificial intelligence-assisted patient education material for ostomy patients: A methodological study. *Journal of Advanced Nursing, 81*, 3859–3867. https://doi.org/10.1111/jan.16542
16. Kim, D. D., Zhang, J., Sagae, K., Devon, H. A., Hoffmann, T. J., Rountree, L., & Fukuoka, Y. (2025). Human-delivered conversation versus AI chatbot conversation in increasing heart attack knowledge in women in the United States: Quasi-experimental studies. *Journal of Medical Internet Research, 27*, e73184. https://doi.org/10.2196/73184
17. Kirchner, G. J., Kim, R. Y., Weddle, J. B., & Bible, J. E. (2023). Can artificial intelligence improve the readability of patient education materials? *Clinical Orthopaedics and Related Research, 481*, 2260–2267. https://doi.org/10.1097/CORR.0000000000002668
18. Kufta, A. Y., & Djalilian, A. R. (2025). Enhancing patient education with AI: A readability analysis of AI-generated versus American Academy of Ophthalmology online patient education materials. *Journal of Clinical Medicine, 14*, 6968. https://doi.org/10.3390/jcm14196968
19. Laily, A., Schwab-Reese, L. M., Davish, M., Cahue, E., LaRoche, K. J., Rodriguez, N. M., Duncan, R. J., Hubach, R. D., & Kasting, M. L. (2026). Examining artificial intelligence chatbots' responses in providing human papillomavirus vaccine information for young adults: Qualitative content analysis. *JMIR Public Health and Surveillance, 12*, e79720. https://doi.org/10.2196/79720
20. Lee, T. J., Campbell, D. J., Rao, A. K., Hossain, A., Elkattawy, O., Radfar, N., Lee, P., & Gardin, J. M. (2024). Evaluating ChatGPT responses on atrial fibrillation for patient education. *Cureus, 16*(6), e61680. https://doi.org/10.7759/cureus.61680

21. Lv, X., Zhang, X., Li, Y., Ding, X., Lai, H., & Shi, J. (2024). Leveraging large language models for improved patient access and self-management: Assessor-blinded comparison between expert- and AI-generated content. *Journal of Medical Internet Research, 26*, e55847. https://doi.org/10.2196/55847
22. Mendoza-Pinto, C., Munguia-Realpozo, P., Etchegaray-Morales, I., Ramirez-Lara, E., & Solis-Poblano, J. C. (2025). Artificial intelligence in patient education: Evaluating large language models for understanding rheumatology literature. *Frontiers in Digital Health, 7*, 1623399. https://doi.org/10.3389/fdgth.2025.1623399
23. Michalska, A. A., Stefaniak, M. M., & Gotlib-Malkowska, J. (2025). Can ChatGPT provide patient-friendly and reliable information on cervical cancer screening? A study of ChatGPT-generated information in Polish. *Medical Science Monitor, 31*, e947992. https://doi.org/10.12659/MSM.947992
24. Naufal, M., Kannan, S., Senthilkumar, S., Palaniappan, T., Kumar, S. S., & Kulkarni, P. P. (2025). Patient education using AI: A cross-sectional study comparing ChatGPT and Google Gemini-generated patient education brochures on various surgical management of breast cancer. *Cureus, 17*(11), e97525. https://doi.org/10.7759/cureus.97525
25. Richlitzki, C., Mansoorian, S., Kasmann, L., Stoleriu, M. G., Kovacs, J., Sienel, W., Kauffmann-Guerrero, D., Duell, T., Schmidt-Hegemann, N. S., Belka, C., Corradini, S., & Eze, C. (2025). Assessing ChatGPT's educational potential in lung cancer radiotherapy from clinician and patient perspectives: Content quality and readability analysis. *JMIR Cancer, 11*, e69783. https://doi.org/10.2196/69783
26. Rodler, S., Cei, F., Ganjavi, C., Checcucci, E., De Backer, P., Rivero Belenchon, I., Taratkin, M., Puliatti, S., Veccia, A., Piazza, P., Baekelandt, L., Kowalewski, K. F., Gomez Rivas, J., Fankhauser, C. D., Moschini, M., Gandaglia, G., Campi, R., De Castro Abreu, A., Russo, G. I., ... Cacciamani, G. E. (2025). GPT-4 generates accurate and readable patient education materials aligned with current oncological guidelines: A randomized assessment. *PLOS ONE, 20*, e0324175. https://doi.org/10.1371/journal.pone.0324175
27. Rossettini, G., & Palese, A. (2026). Accurate but not accessible: Artificial intelligence chatbots and the limits of patient education. *Proceedings (Baylor University Medical Center), 39*(2), 294–295. https://doi.org/10.1080/08998280.2026.2613616
28. Roster, K., Kann, R. B., Farabi, B., Gronbeck, C., Brownstone, N., & Lipner, S. R. (2024). Readability and health literacy scores for ChatGPT-generated dermatology public education materials: Cross-sectional analysis of sunscreen and melanoma questions. *JMIR Dermatology, 7*, e50163. https://doi.org/10.2196/50163
29. Sriram, N., Bajaj, A., Li, M., Adam, T., Nguyen, A., Chon, J., & Galiano, R. D. (2026). Comparing the readability and content quality of online patient education materials and ChatGPT-generated patient education materials for breast cancer surgery and reconstruction. *Archives of Plastic Surgery, 53*, 191–198. https://doi.org/10.1055/a-2794-9984
30. Stephenson-Moe, C. A., Behers, B. J., Gibons, R. M., Behers, B. M., De Jesus Herrera, L., Anneaud, D., Rosario, M. A., Wojtas, C. N., Bhambrah, S., & Hamad, K. M. (2025). Assessing the quality and readability of patient education materials on chemotherapy cardiotoxicity from artificial intelligence chatbots: An observational cross-sectional study. *Medicine, 104*(15), e42135. https://doi.org/10.1097/MD.0000000000042135
31. Sudharshan, R., Shen, A., Gupta, S., & Zhang-Nunes, S. (2024). Assessing the utility of ChatGPT in simplifying text complexity of patient educational materials. *Cureus, 16*(3), e55304. https://doi.org/10.7759/cureus.55304
32. Thia, I., & Saluja, M. (2024). ChatGPT: Is this patient education tool for urological malignancies readable for the general population? *Research and Reports in Urology, 16*, 31–37. https://doi.org/10.2147/RRU.S440633
33. Tiller, N. B., Marcon, A. R., Zenone, M., Kidd, K. E., Jeukendrup, A. E., Master, Z., & Caulfield, T. (2026). Generative artificial intelligence-driven chatbots and medical misinformation: An accuracy, referencing and readability audit. *BMJ Open, 16*, e112695. https://doi.org/10.1136/bmjopen-2025-112695
34. Todorov, D., Park, J. Y., Ng Hing Cheung, J. A., Avramidou, E., & Gnanappiragasam, D. (2025). Assessing the readability of dermatological patient information leaflets generated by ChatGPT-4 and its associated plugins. *Skin Health and Disease, 5*, 14–21. https://doi.org/10.1093/skinhd/vzae015
35. Will, J., Gupta, M., Zaretsky, J., Dowlath, A., Testa, P., & Feldman, J. (2025). Enhancing the readability of online patient education materials using large language models: Cross-sectional study. *Journal of Medical Internet Research, 27*, e69955. https://doi.org/10.2196/69955
36. Wu, L. L., Hong, A. T., Davuluru, S. S., Moreno-Diaz, E. A., Gutierrez-Hernandez, J. C., Martinez-Camarillo, J. C., Xu, B. Y., & Zhang-Nunes, S. (2026). Utility of ChatGPT-4o in creating patient handouts in ophthalmology: A comparison with American Academy of Ophthalmology educational materials. *Translational Vision Science & Technology, 15*(2), 14. https://doi.org/10.1167/tvst.15.2.14
37. Yan, Z., Liu, J., Fan, Y., Lu, S., Xu, D., Yang, Y., Wang, H., Mao, J., Tseng, H.-C., Chang, T.-H., & Chen, Y. (2025). Ability of ChatGPT to replace doctors in patient education: Cross-sectional comparative analysis of inflammatory bowel disease. *Journal of Medical Internet Research, 27*, e62857. https://doi.org/10.2196/62857
38. Yau, J. Y.-S., Saadat, S., Hsu, E., Murphy, L. S.-L., Roh, J. S., Suchard, J., Tapia, A., Wiechmann, W., & Langdorf, M. I. (2024). Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: Experimental comparative study. *Journal of Medical Internet Research, 26*, e60291. https://doi.org/10.2196/60291
39. Zalzal, H. G., Abraham, A., Cheng, J., & Shah, R. K. (2024). Can ChatGPT help patients answer their otolaryngology questions? *Laryngoscope Investigative Otolaryngology, 9*, e1193. https://doi.org/10.1002/lio2.1193
40. Zhang, Y., Wan, X.-H., Kong, Q.-Z., Liu, H., Liu, J., Guo, J., Yang, X.-Y., Zuo, X.-L., & Li, Y.-Q. (2025). Evaluating large language models as patient education tools for inflammatory bowel disease: A comparative study. *World Journal of Gastroenterology, 31*(6), 102090. https://doi.org/10.3748/wjg.v31.i6.102090