Journal of Clinical and Nursing Research, 2026, Volume 10, Issue 10
https://bbwpublisher.com/journal/JCNR
Online ISSN: 2208-3693
Print ISSN: 2208-3685

Advances in the Assessment of Generative AI–Generated Health Information: A Review of Evaluation Tools

Qinxue Chen[1], Mingyan Shen[2]*, Jie Lang[3], Yanping Shen[4]

[1]School of Public Health and Nursing, Hangzhou Normal University, Hangzhou 311121, Zhejiang, China
[2]Key Laboratory of Artificial Organs and Computational Medicine of Zhejiang Province, Shulan (Hangzhou) Hospital, Shulan International Medical College, Zhejiang Shuren University, Hangzhou 310015, Zhejiang, China
[3]Shulan International Medical College, Zhejiang Shuren University, Hangzhou 310015, Zhejiang, China
[4]Shulan (Hangzhou) Hospital, Hangzhou 310022, Zhejiang, China

***Corresponding author*:** Mingyan Shen, 289155169@qq.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Generative artificial intelligence (GenAI) is increasingly used to generate health information, raising concerns about its quality, reliability, and suitability for different users. This review examines the applications and limitations of existing assessment instruments for GenAI-generated health information, with particular attention to nursing practice. Established instruments provide complementary measures of information quality, understandability, actionability, and readability, while emerging AI-specific tools address additional concerns related to factual accuracy and evidence support. However, current instruments have limitations in assessing context-dependent recommendations, dynamic interactions, and cross-linguistic applicability. No single instrument adequately captures all dimensions of GenAI-generated health information. A context-sensitive, multidimensional approach combining complementary assessment methods with appropriate clinical verification may therefore be beneficial. Further research is needed to establish the applicability and validity of these approaches across different users and clinical contexts.

Keywords: Generative artificial intelligence; Large language models; Assessment tools; Quality of health information

Online publication: November 25, 2026

1. Introduction

Generative artificial intelligence (GenAI), particularly large language models (LLMs), has shown potential in producing patient-oriented health information, simplifying complex medical content, and supporting health communication [1,2]. In nursing practice, these technologies may facilitate patient education and discharge guidance while also providing nursing-related information and care recommendations for healthcare professionals [3,4]. However, the fluency and coherence of AI-generated responses do not necessarily reflect

their clinical accuracy, completeness, or suitability for the intended audience [5]. Evaluating the quality of such information therefore requires consideration of factors beyond linguistic presentation alone.

These limitations have been illustrated in studies evaluating GenAI-generated health information. Pan et al. [6] evaluated 100 cancer-related responses from four AI chatbots and found that favorable overall information quality scores were accompanied by moderate understandability and poor actionability. Similarly, Zaretsky et al. [7] reported substantial improvements in readability and understandability after GPT-4 was used to transform 50 discharge summaries into patient-friendly formats. Nevertheless, independent reviewers identified omissions and inaccuracies with potential safety implications [8]. These findings illustrate that improvements in information presentation do not necessarily ensure the completeness or clinical accuracy of AI-generated health information.

Various assessment instruments have been applied to evaluate GenAI-generated health information, addressing dimensions such as reliability, content quality, understandability, actionability, and readability [9]. More recently, AI-specific tools have emerged to assess additional concerns, including fabricated information and insufficient evidence support [10]. However, these instruments differ in their measurement constructs, and their applicability to GenAI-generated content cannot be assumed solely from their established use in conventional health information assessment [11]. This review examines the characteristics, applications, and limitations of existing tools for evaluating text-based health information generated by GenAI. It further considers their complementary roles and implications for assessing information intended for different users, with particular attention to nursing practice.

2. Traditional health information assessment tools applied to AI outputs

Established instruments originated in assessments of written materials or online information. Applying them to AI outputs requires a defined assessment object, task, and instrument version. The following tools address distinct constructs and should not be treated as interchangeable on the basis of their total scores.

2.1. Information reliability evaluation tool: DISCERN

The DISCERN instrument, developed by Charnock et al. in 1999, was designed to assess the quality of written consumer health information on treatment choices [12]. It consists of 16 items organized into three sections: the reliability of information, the quality of information on treatment options, and an overall quality rating. By examining information sources, transparency, balance, and the presentation of treatment benefits and risks, DISCERN provides a structured approach to evaluating patient-oriented health information. Its established framework has also supported its increasing use in assessing health information generated by generative artificial intelligence [6,13].

Several studies have applied DISCERN to compare the quality of GenAI-generated health information across different clinical topics. Chen et al. [14] evaluated responses to frequently asked questions concerning intertrochanteric fractures in older adults using five conversational AI systems, including ChatGPT, Copilot, Gemini, ERNIE Bot, and ChatGLM. Differences in DISCERN scores were observed across models, with ChatGPT and ERNIE Bot demonstrating relatively favorable performance. Similarly, Yıldız et al. [13] incorporated DISCERN alongside the Ensuring Quality Information for Patients (EQIP) instrument and the Global Quality Score (GQS) to evaluate AI-generated information on sexually transmitted diseases. Together, these studies illustrate the utility of DISCERN in identifying differences in the presentation and reliability

of AI-generated health information. However, such findings should not be interpreted as direct evidence of clinical accuracy or patient safety.

The growing diversity of GenAI applications has also prompted efforts to adapt DISCERN beyond its original scope. Modified versions, commonly referred to as mDISCERN, have been used in health information evaluations, although their item selection and scoring approaches may vary across studies[15,16]. Hua et al.[17] further explored the use of AI-DISCERN in evaluating AI-generated content, reflecting broader efforts to address the specific characteristics of AI-mediated health communication. In addition, Shan et al.[18] translated, culturally adapted, and validated a simplified Chinese version of DISCERN (C-DISCERN), providing a language-specific instrument for assessing Chinese patient-oriented information on treatment choices. Nevertheless, evidence supporting the application of these adaptations to dynamic GenAI interactions remains limited. Although DISCERN provides a structured assessment of information quality, its applicability to GenAI-generated nursing recommendations remains limited. In particular, it does not directly assess the clinical accuracy of procedural instructions, individualized recommendations, or potential patient safety risks. Therefore, DISCERN may serve as a foundational assessment tool but should be supplemented by clinical evidence verification and context-specific safety evaluation.

2.2. Content quality assessment tools

2.2.1. Ensuring quality information for patients

The Ensuring Quality Information for Patients (EQIP) instrument was developed by Moult et al. in 2004 to assess the quality of written patient information [9]. The original 20-item instrument focuses on content completeness, presentation, understandability, and usefulness, providing a structured approach to evaluating how health information is presented to patients.

EQIP has increasingly been applied to evaluate GenAI-generated health information. Ozduran et al. [19] compared responses from ChatGPT, Gemini, and Perplexity to pain-related questions and found that Perplexity achieved significantly higher EQIP scores. Similarly, Şahin et al. [20] evaluated five AI chatbots responding to questions about erectile dysfunction, with Copilot demonstrating superior EQIP performance. These findings highlight differences in information quality across AI systems, although the results remain specific to the models and clinical topics examined. To provide a more comprehensive assessment, Charvet-Berard et al. [21] expanded the original instrument to 36 items, covering content, identification, and structure. The modified EQIP-36 has also been used to evaluate medical information generated by ChatGPT [22].

Nevertheless, well-organized and seemingly comprehensive responses do not necessarily provide sufficient information for safe clinical decisions. GenAI may omit relevant risks, contraindications, or alternative treatment options despite producing coherent explanations [7]. EQIP therefore offers a useful framework for assessing the content and presentation of AI-generated patient information, including nursing education materials. However, its findings should be interpreted alongside independent assessments of clinical accuracy, guideline consistency, and patient safety.

2.2.2. Global quality score

The Global Quality Score (GQS) is a five-point rating scale used to assess the overall quality, information flow, and usefulness of health information [23]. Its simplicity facilitates rapid evaluation and has led to its use alongside more structured instruments, such as DISCERN and EQIP. Gibson et al. [24] applied GQS to assess GPT-4-generated information on prostate cancer, reporting a high mean score of 4.46 out of 5. Similarly,

Ozduran et al. [19] compared ChatGPT, Gemini, and Perplexity in answering pain-related questions and found significant differences in GQS ratings, with Gemini achieving higher scores.

Despite its practical advantages, GQS provides limited insight into the specific factors underlying an overall quality rating. A coherent and professionally presented AI response may receive a favorable score while containing factual inaccuracies or omitting clinically relevant information. Notably, Gibson et al. [24] identified readability difficulties and fabricated references despite high GQS ratings, highlighting the distinction between perceived overall quality and specific information deficiencies. GQS may therefore serve as a supplementary measure of overall information quality, including in the evaluation of AI-generated patient education materials, but should not be used alone to establish clinical accuracy or safety.

2.2.3. Patient education materials assessment tool

The Patient Education Materials Assessment Tool (PEMAT), developed by the Agency for Healthcare Research and Quality (AHRQ), assesses the understandability and actionability of patient education materials [25]. It includes separate versions for printable (PEMAT-P) and audiovisual (PEMAT-A/V) materials, with independent scores for each domain. Understandability concerns how clearly information is presented, whereas actionability reflects whether readers are provided with specific, manageable steps. PEMAT thus addresses aspects of patient-oriented communication that are not fully captured by conventional information quality assessments.

Its application to GenAI-generated health information has revealed notable differences between overall information quality and practical usability. Pan et al. [6] evaluated cancer-related responses from four AI chatbots and reported relatively high information quality but moderate understandability (66.7%) and poor actionability (20.0%). In contrast, Zaretsky et al. [7] demonstrated that GPT-4 substantially improved the understandability of patient-friendly discharge summaries compared with the original documents (81% vs. 13%), although actionability was not assessed. Efforts to adapt PEMAT have also emerged. Gibson et al. [24] developed PEMAT-AI by modifying items unsuitable for text-only AI responses, while Shan et al. [26] translated and validated the simplified Chinese version of PEMAT-P (C-PEMAT-P).

Despite these developments, PEMAT scores reflect characteristics of educational materials rather than patients' actual comprehension or ability to perform recommended actions [25]. Modified versions also require further validation across different GenAI applications. PEMAT may therefore help identify communication barriers in AI-generated patient education, particularly in nursing, where clear instructions are essential for supporting self-care.

2.4. Readability assessment tools

Readability formulas, including the Flesch Reading Ease (FRE), Flesch–Kincaid Grade Level (FKGL), and Simple Measure of Gobbledygook (SMOG) index, estimate the reading difficulty of written materials using linguistic features such as sentence length and syllable complexity [27-29]. FRE provides a reading ease score, whereas FKGL and SMOG estimate the educational grade level required to read a text. These measures offer a standardized and efficient approach to examining the linguistic complexity of GenAI-generated health information.

Evidence suggests that fluent AI-generated responses are not necessarily accessible to general audiences. Singh et al. [30] found that patient education materials on inferior vena cava filters generated by ChatGPT and Microsoft Copilot exceeded recommended reading grade levels. A meta-analysis comparing ChatGPT with

Bard/Gemini also identified differences in readability, although neither consistently achieved recommended levels for patient-facing information [31]. Similar concerns have been reported across other medical topics [32–34].

However, formula-based readability scores capture only selected linguistic features and should not be equated with actual patient comprehension. Necessary medical terminology may increase estimated reading difficulty even when accompanied by clear explanations, while shorter sentences do not necessarily convey complex concepts more effectively [25]. Furthermore, FRE, FKGL, and SMOG were developed for English-language texts and cannot be directly transferred to Chinese without linguistic adaptation and validation. Their principal value therefore lies in identifying potentially inaccessible language rather than establishing whether health information is understandable to a particular patient. Future assessments should incorporate language-specific measures and direct evaluation with intended readers, particularly when GenAI is used to develop patient education materials.

3. AI-specific health information assessment tools

The CLEAR tool was developed by Sallam et al. to assess the quality of AI-generated health information across five dimensions: Completeness, Lack of false information, Evidence support, Appropriateness, and Relevance [10]. Each dimension is assessed using a five-point rating scale. Following a literature review and expert assessment, the instrument underwent pilot testing involving 32 healthcare professionals, including nurses, physicians, pharmacists, and laboratory technicians. Preliminary findings supported its internal consistency, although inter-rater agreement varied across AI models. The authors subsequently applied CLEAR to compare health information generated by four AI systems, identifying differences in their overall quality. Beyond model comparisons, CLEAR has also been used to investigate linguistic disparities. In a study of infectious disease-related questions, Sallam et al. [35] reported significantly higher CLEAR scores for English than Arabic responses, suggesting that the quality of AI-generated health information may vary across languages.

Unlike traditional instruments developed for conventional patient information, CLEAR explicitly incorporates factual accuracy and evidence support into its assessment framework. This represents a meaningful step toward addressing concerns specific to GenAI-generated content, particularly fabricated information and unsupported claims [5]. Nevertheless, its five-item structure may not adequately capture the complexity of individualized clinical recommendations or risks arising during multi-turn interactions. Furthermore, evidence regarding its measurement properties and applicability across clinical settings remains preliminary. CLEAR should therefore be considered a promising AI-specific assessment tool rather than a fully validated framework for comprehensive clinical quality evaluation.

4. Challenges and future directions in evaluating GenAI-generated health information

4.1. Value and limitations of existing assessment instruments

Established health information assessment instruments provide a valuable foundation for evaluating GenAI-generated content. DISCERN, EQIP, and GQS address different aspects of information quality, while PEMAT and readability formulas offer complementary perspectives on communication. Their established scoring procedures facilitate structured assessment and comparisons across studies [36]. Nevertheless, these

instruments were developed for different purposes, and their measurement constructs are neither identical nor entirely independent. Combining their scores without considering conceptual overlap may therefore introduce redundancy rather than provide a more comprehensive evaluation.

A further concern is the distinction between information quality and clinical validity. A response may satisfy criteria related to organization, completeness, or overall usefulness while containing unsupported clinical claims. Conversely, factual accuracy alone does not ensure that information is appropriate for its intended audience. AI-specific instruments such as CLEAR have begun to address these gaps by incorporating factual accuracy and evidence support [10]. However, preliminary validation evidence does not establish their suitability across all clinical contexts. Rather than seeking a single universal instrument, future research should examine which combinations of assessment methods provide valid, reliable, and clinically meaningful information for specific evaluation purposes.

4.2. From information presentation to comprehension and action

The increasing use of GenAI to simplify medical information has brought renewed attention to the distinction between linguistic accessibility and practical usefulness. Evidence reviewed in this article indicates that favorable overall quality assessments may coexist with limited actionability, while improvements in readability and understandability do not necessarily eliminate clinically important omissions [6,7].

This distinction has implications for the interpretation of assessment results.

Readability formulas primarily estimate linguistic complexity [27,37], whereas PEMAT evaluates characteristics of educational materials that may facilitate understanding and action [25]. Neither directly measures whether an individual has correctly interpreted the information or can apply it in practice. Moreover, simplified language may improve accessibility without adequately communicating uncertainty, conditional recommendations, or situations requiring professional assistance [5]. Future evaluations should therefore consider supplementing material-based scores with user-centered assessments, such as comprehension testing, interpretation of key instructions, and task-based evaluations. These approaches may provide a more meaningful indication of whether AI-generated information meets the needs of its intended users.

4.3. Challenges in evaluating dynamic AI interactions

Most established health information assessment instruments were designed for relatively static materials. GenAI systems, however, generate responses that may change as users provide additional information, clarify questions, or challenge previous answers. Evaluating a single response may therefore overlook important features of the interaction, including consistency across turns, recognition of conflicting information, and the ability to correct earlier inaccuracies.

These challenges are particularly relevant when an initially reasonable response becomes inappropriate after new clinical details are introduced. Conversely, a system may improve an incomplete answer through clarification. A response-level score cannot fully characterize either process [5,10]. Future assessment methods should consider the interaction, rather than an isolated response, as an additional unit of analysis. Standardized multi-turn scenarios, including changes in clinical context and opportunities for error correction, may help assess consistency, responsiveness, and the management of uncertainty. Such methods require further development and validation before they can be incorporated into routine evaluation.

4.4. Cross-linguistic and cross-cultural applicability

The applicability of existing instruments across languages and cultural contexts presents another methodological challenge. Readability formulas such as FRE, FKGL, and SMOG rely on linguistic features of English and cannot be directly transferred to languages with different writing systems and grammatical structures [27-29]. Although Chinese adaptations of DISCERN and PEMAT have undergone translation and psychometric evaluation [18,26], these findings do not automatically establish their validity for GenAI-generated content or interactive applications.

Language-related differences have also been observed in evaluations using CLEAR, with English responses receiving higher scores than Arabic responses in a study of infectious disease information [35]. Such findings raise questions about whether GenAI provides information of comparable quality across linguistic settings. However, differences in observed scores should not be interpreted as direct evidence of measurement equivalence or inequitable health outcomes. Future studies should examine cross-cultural validity, rating consistency, and the comparability of assessment results across languages. Particular attention should be given to culturally appropriate terminology, differences in health literacy, and the interpretation of clinical recommendations within local healthcare contexts.

4.5. Implications for nursing practice and future assessment

The selection of assessment tools should reflect not only the characteristics of GenAI-generated content but also its intended users and clinical purposes. This distinction is particularly important in nursing, where GenAI may be used to prepare patient education materials or provide professional information and care recommendations.

For patient-facing information, evaluation should address clinical accuracy, understandability, actionability, and accessibility. Instruments such as PEMAT and readability measures can help identify communication barriers, but they should be complemented by clinical content review and, where feasible, testing with intended users. Information intended for nurses presents different demands. AI-generated nursing recommendations may require assessment of evidence consistency, task relevance, patient-specific appropriateness, and the feasibility of proposed interventions within clinical workflows[3]. Tools originally developed for patient education cannot be assumed to measure these professional requirements adequately.

These differences suggest the need for a context-sensitive, multidimensional assessment approach. Rather than applying every available instrument to each AI response, evaluators should select measures according to the intended audience, information type, and potential clinical risk. Independent verification against current clinical evidence and relevant guidelines remains particularly important when generated content may influence care decisions [2]. Future research should also examine the measurement properties of existing instruments in professional nursing contexts and develop methods for evaluating interactive, individualized recommendations.

Importantly, the available evidence remains concentrated on patient-oriented health information, limiting conclusions about the performance of these tools when used to assess professional nursing recommendations [13]. Addressing this gap will require studies involving practicing nurses, representative clinical scenarios, and clearly defined assessment outcomes. Such work could help establish more appropriate evaluation strategies without extending existing instruments beyond their validated purposes.

This review is based on a selective synthesis of the available literature, and its conclusions should be considered within this scope. As existing evidence predominantly concerns patient-facing health information,

further studies are warranted to examine the applicability of assessment instruments to professional nursing recommendations and other clinical contexts.

5. Conclusion

GenAI is reshaping the production and delivery of health information by enabling increasingly interactive and context-dependent responses. Established assessment instruments provide a foundation for evaluating information quality, understandability, actionability, and readability, while emerging AI-specific tools place greater emphasis on factual accuracy and evidence support. Nevertheless, no single instrument adequately captures all dimensions relevant to GenAI-generated health information [36]. In particular, favorable scores for presentation or overall quality should not be interpreted as sufficient evidence of clinical reliability or practical usefulness [6,7,24].

Future evaluation should therefore adopt a context-sensitive, multidimensional approach that combines complementary assessment methods according to the intended audience, information purpose, and potential risks. In nursing practice, this requires distinguishing patient-facing educational materials from professional information and care recommendations. Further validation across languages, clinical contexts, and interactive scenarios is needed to establish the applicability of existing instruments and support the responsible use of GenAI-generated health information.

Funding

Key Laboratory of Artificial Organs and Computational Medicine of Zhejiang Province (Project No.: SZD2026B015); "Four New" Research and Practice Project of Zhejiang Shuren University (Project No.: xyk202601)

Disclosure statement

The authors declare no conflict of interest

Reference

[1] Sallam M, 2023, ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare, 11(6): 887.

[2] Park YJ, Pillai A, Deng J, et al., 2024, Assessing the Research Landscape and Clinical Utility of Large Language Models: A Scoping Review. BMC Medical Informatics and Decision Making, 24(1): 72.

[3] Li X, Wu J, Li Y, et al., 2025, The Application of Large Language Models in the Field of Nursing. International Journal of Nursing, 44(20): 3829–3835.

[4] Amankwaa I, Ekpor E, Cudjoe D, et al., 2025, Patterns, Advances, and Gaps in Using ChatGPT and Similar Technologies in Nursing Education: A PAGER Scoping Review. Nurse Education Today, 153: 106822.

[5] Ji Z, Lee N, Frieske R, et al., 2023, Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12): 1–38, Article 248.

[6] Pan A, Musheyev D, Bockelman D, et al., 2023, Assessment of Artificial Intelligence Chatbot Responses to Top

Searched Queries About Cancer. JAMA Oncology, 9(10): 1437–1440.

[7] Zaretsky J, Kim JM, Baskharoun S, et al., 2024, Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. JAMA Network Open, 7(3): e240357.

[8] Menz BD, Kuderer NM, Bacchi S, et al., 2024, Current Safeguards, Risk Mitigation, and Transparency Measures of Large Language Models Against the Generation of Health Disinformation: Repeated Cross-Sectional Analysis. BMJ, 384: e078538.

[9] Moult B, Franck LS, Brady H, 2004, Ensuring Quality Information for Patients: Development and Preliminary Validation of a New Instrument to Improve the Quality of Written Health Care Information. Health Expectations, 7(2): 165–175.

[10] Sallam M, Barakat M, Sallam M, 2023, Pilot Testing of a Tool to Standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-Based Models. Cureus, 15(11): e49373.

[11] Charvet-Berard AI, Chopard P, Perneger TV, 2008, Measuring Quality of Patient Information Documents with an Expanded EQIP Scale. Patient Education and Counseling, 70(3): 407–411.

[12] Bernard A, Langille M, Hughes S, et al., 2007, A Systematic Review of Patient Inflammatory Bowel Disease Information Resources on the World Wide Web. The American Journal of Gastroenterology, 102(9): 2070–2077.

[13] Shoemaker SJ, Wolf MS, Brach C, 2014, Development of the Patient Education Materials Assessment Tool (PEMAT): A New Measure of Understandability and Actionability for Print and Audiovisual Patient Information. Patient Education and Counseling, 96(3): 395–403.

[14] Liao W, Han Y, Ma C, 2025, Evaluation Paradigms for Conversational AI in Healthcare: A Systematic Review. Chinese Journal of Health Policy, 18(7): 78–86.

[15] Charnock D, Shepperd S, Needham G, et al., 1999, DISCERN: An Instrument for Judging the Quality of Written Consumer Health Information on Treatment Choices. Journal of Epidemiology & Community Health, 53(2): 105–111.

[16] Yıldız HA, Söğütdelen E, 2025, AI Chatbots as Sources of STD Information: A Study on Reliability and Readability. Journal of Medical Systems, 49(1): 43.

[17] Chen W, Bai X, Wang G, et al., 2025, Reliability and Feasibility of Interactive Artificial Intelligence Language Model Evaluation of Common Problems of Femoral Intertrochanteric Fractures in the Elderly. Chinese Journal of Bone and Joint Injury, 40(6): 575–578.

[18] Onder CE, Koc G, Gokbulut P, et al., 2024, Evaluation of the Reliability and Readability of ChatGPT-4 Responses Regarding Hypothyroidism During Pregnancy. Scientific Reports, 14(1): 243.

[19] Kacer EO, 2025, Evaluating AI-Based Breastfeeding Chatbots: Quality, Readability, and Reliability Analysis. PLOS ONE, 20(3): e0319782.

[20] Hua HU, Kaakour AH, Rachitskaya A, et al., 2023, Evaluation and Comparison of Ophthalmic Scientific Abstracts and References by Current Artificial Intelligence Chatbots. JAMA Ophthalmology, 141(9): 819–824.

[21] Shan Y, Xing Z, Dong Z, et al., 2023, Translating and Adapting the DISCERN Instrument into a Simplified Chinese Version and Validating Its Reliability: Development and Usability Study. Journal of Medical Internet Research, 25: e40733.

[22] Ozduran E, Akkoc I, Büyükçoban S, et al., 2025, Readability, Reliability and Quality of Responses Generated by ChatGPT, Gemini, and Perplexity for the Most Frequently Asked Questions About Pain. Medicine, 104(11): e41780.

[23] Şahin MF, Ateş H, Keleş A, et al., 2024, Responses of Five Different Artificial Intelligence Chatbots to the Top

Searched Queries About Erectile Dysfunction: A Comparative Analysis. Journal of Medical Systems, 48(1): 38.

[24] Walker HL, Ghani S, Kuemmerli C, et al., 2023, Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. Journal of Medical Internet Research, 25: e47479.

[25] Gibson D, Jackson S, Shanmugasundaram R, et al., 2024, Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment. Journal of Medical Internet Research, 26: e55939.

[26] Shan Y, Ji M, Dong Z, et al., 2023, The Chinese Version of the Patient Education Materials Assessment Tool for Printable Materials: Translation, Adaptation, and Validation Study. Journal of Medical Internet Research, 25: e39808.

[27] Flesch R, 1948, A New Readability Yardstick. Journal of Applied Psychology, 32(3): 221–233.

[28] Kincaid JP, Fishburne RP Jr, Rogers RL, et al., 1975, Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel (Research Branch Report No. 8-75). Naval Technical Training Command.

[29] McLaughlin GH, 1969, SMOG Grading—A New Readability Formula. Journal of Reading, 12(8): 639–646.

[30] Singh SP, Jamal A, Qureshi F, et al., 2024, Leveraging Generative Artificial Intelligence Models in Patient Education on Inferior Vena Cava Filters. Clinics and Practice, 14(4): 1507–1514.

[31] DeTemple DE, Meine TC, 2025, Comparison of the Readability of ChatGPT and Bard in Medical Communication: A Meta-Analysis. BMC Medical Informatics and Decision Making, 25(1): 325.

[32] Gül Ş, Erdemir İ, Hanci V, et al., 2024, How Artificial Intelligence Can Provide Information About Subdural Hematoma: Assessment of Readability, Reliability, and Quality of ChatGPT, BARD, and Perplexity Responses. Medicine, 103(18): e38009.

[33] Hanci V, Ergün B, Gül Ş, et al., 2024, Assessment of Readability, Reliability, and Quality of ChatGPT, BARD, Gemini, Copilot, and Perplexity Responses on Palliative Care. Medicine, 103(33): e39305.

[34] Kara M, Ozduran E, Kara MM, et al., 2025, Evaluating the Readability, Quality, and Reliability of Responses Generated by ChatGPT, Gemini, and Perplexity on the Most Commonly Asked Questions About Ankylosing Spondylitis. PLOS ONE, 20(6): e0326351.

[35] Wang LW, Miller MJ, Schmitt MR, et al., 2013, Assessing Readability Formula Differences with Written Health Information Materials: Application, Results, and Recommendations. Research in Social and Administrative Pharmacy, 9(5): 503–516.

[36] Sallam M, Al-Mahzoum K, Alshuaib O, et al., 2024, Language Discrepancies in the Performance of Generative Artificial Intelligence Models: An Examination of Infectious Disease Queries in English and Arabic. BMC Infectious Diseases, 24(1): 799.

[37] Leonforte F, Cavallaro A, Nicosia V, et al., 2026, Artificial Intelligence in Medical and Surgical Information: A Systematic Review of Accuracy, Readability, Reliability, and Emotional Resonance. AI, 7(10): 396.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.