Review Article

Health Informatics Journal

Health Informatics Journal
Vol. 31(4): 1–20

DOI: 10.1177/14604582251388879
journals.sagepub.com/home/jhi

Quality and readability of chatbot responses to patient questions: A systematic cross-sectional meta-synthesis

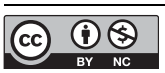
Peter Whittaker[1] **and Mengyan Sun**[2,3]

Abstract

Introduction: Patients increasingly use chatbots to obtain medical information, a trend that has provoked both optimism and pessimism. Numerous studies have evaluated the quality and readability of these outputs. This study synthesizes these findings through a cross-sectional meta-synthesis. **Methods:** We identified studies that evaluated responses using the DISCERN instrument, designed to assess the quality of written material. Additionally, we only included studies that also evaluated readability. We recorded the chatbot used, DISCERN scores, the number of words in each question, the number of questions asked, the number of DISCERN evaluators, the readability of responses, and the year the study was conducted. We also assessed the influence of each publication's journal ranking using the Journal Citation Indicator. **Results:** We identified 42 studies that conducted 86 tests. Chatbot response readability decreased as response quality increased. Forty-nine tests produced responses ranked "good" or better, and only 10 scored below college-level readability. We significantly increased readability by adding the phrase "write responses at sixth-grade reading level" to prompts that previously produced post-graduate reading level responses in published studies. **Discussion:** Variable quality and poor readability of chatbot responses reinforce pessimism about their utility. Nevertheless, appropriate "prompt engineering" provides scope to enhance response quality and readability.

[1]Green Templeton College, Oxford, UK
[2]Harris Manchester College, University of Oxford, Oxford, UK
[3]SMY Consulting, Shanghai, China

Corresponding author:
Peter Whittaker, Green Templeton College, University of Oxford, 43 Woodstock Road, Oxford OX2 6HG, UK.
Email: peter.whittaker@gtc.ox.ac.uk



Keywords
chatbots, ChatGPT, DISCERN, large language model, patient engagement, prompt engineering, readability

Introduction

Healthcare providers have anticipated patients' use of chatbots to seek medical information with optimism and pessimism.[1,2] Such usage is a continuation of internet search engine use.[3,4] Nevertheless, chatbots may fail to provide accurate information, and the information that is provided may not be written at an easily comprehensible level.[5,6] Numerous studies across various medical fields have examined chatbot responses to patient questions. These data have yet to be synthesized. There are at least three reasons why synthesis at this early stage of chatbot use by patients could be helpful. First, healthcare providers need to be able to assess the accuracy and readability of chatbot responses. This ability is vital because providers may never have the opportunity to corroborate the chatbots' answers with the patients. Patients may also benefit from learning the chatbots' strengths and weaknesses. Secondly, synthesis may identify factors that influence response quality and readability, thereby enabling improvement in both. Thirdly, synthesis may reveal failings in how chatbots are currently used and studied in this context.

We sought to address these issues through a systematic analysis of the current literature evaluating chatbot responses to both real and hypothetical patient questions. Our cross-sectional meta-synthesis includes findings from research published between 2022 and July 2025.

Methods

Assessment tools used

The DISCERN instrument was originally designed to *"assess the quality of written information on treatment choices for a health problem"*.[7,8] The instrument uses 16 questions, each graded on a one-to-five scale. A grade of one means the answer did not address a question component, three means the element was partially answered, and five means the component was fully answered. Therefore, the possible range is 16–80. Response quality is classified as: very poor (16–27); poor (28–38); fair (39–50); good (51–62); and excellent (≥63).

We assessed readability using the Flesch Reading Ease Score (FRES)[9]; the higher the score, the easier the text is to read. FRES ranges from 0 to 100, with values ascribed to different education levels.[10] In the United States (US), the recommendation is that patient material is written at a sixth-grade reading level[11]; corresponding to a FRES ≥80.[10] FRES values in the 30–50 range are considered college-level, 10–30 post-graduate, and <10 professional reading levels.

We included a second readability index: the Flesch-Kincaid Grade Level (FKGL).[12] This measure assigns text to a US school grade based on the ability to read the material. Like FRES, FKGL is calculated based on a formula containing the average number of words in each sentence and the average number of syllables in each word. The two methods use different weightings of these components.

Study identification and data extraction

The primary aim of our study was to evaluate the clinical quality of chatbot-generated responses using the DISCERN tool. To preserve a connection between our evaluation framework (DISCERN) and the assessed responses, we limited our literature search to PubMed-indexed studies. This approach ensured that the chatbot responses assessed were grounded in clinical and patient-oriented contexts.

We identified potential studies using the PubMed search engine with three queries: (1) *"chat* AND DISCERN"*, (2) *"chat* AND Flesch", and (3) "chat* DISCERN Kincaid"*. The reference lists of all the identified studies were searched. As an additional check to confirm that we found relevant studies, every study included was located in the Web of Science database (Clarivate, London, UK), and papers citing those studies were examined. We constructed a PRISMA flowchart to illustrate the search and selection process, which was conducted in mid-July 2025.

We read the abstracts of all identified publications. If the abstracts indicated that the study did not use the DISCERN score along with either FRES or FKGL, or that the DISCERN score was used to assess brochures or other printed patient information, those studies were excluded. All other studies were downloaded and examined in detail to determine whether or not they met the inclusion-exclusion criteria. Studies that used a modified or brief DISCERN score were excluded, except for studies that used the first 15 of the 16 DISCERN questions. The final question assesses the overall quality of the information provided. In those cases, we averaged the scores for the first 15 questions and then added that value to the total. In studies that provided scores for individual DISCERN questions, we found that this extrapolation approach yielded a score that differed from the reported score by less than one point. Therefore, our averaging approach provides a reasonable approximation.

We included studies that examined either real or hypothetical patient questions. We excluded studies that did not list the complete text of the questions. We also excluded studies that sought answers to physicians' questions.

We examined each study's chatbot query protocol, calculated the average number of words used in the question prompts, and recorded the number of questions asked. We extracted the DISCERN score and, when available, FRES and FKGL from data presented in each manuscript. Of the 10 studies (comprising 20 tests) that provided no FRES data, eight provided the entire responses for all questions posed to the chatbots. Therefore, we determined the average FRES for these responses using an online FRES calculator (Flesch Kincaid Calculator - Flesch Reading Ease Calculator). Both authors independently extracted the data, and any discrepancies were resolved by discussion.

Analysis

To determine whether there was an association between response quality and readability, we plotted a graph of FRES and DISCERN scores and performed regression analysis. The included studies used different chatbots known to exhibit different performance, which introduces heterogeneity. Therefore, to minimize this heterogeneity, we also plotted the same relationship using only results from versions of ChatGPT.

We examined five factors that could be associated with the DISCERN score and, therefore, response quality.

(1) Number of prompt words: The hypothesis was that more words would provide more detail and context, which could enhance response quality. We calculated the average and median

number of words in all questions in each study. The DISCERN score was then plotted as a function of the number of words.

(2) Number of questions: A greater number of questions could enhance response quality by providing context and thereby a foundation to answer subsequent questions. However, such enhancement would depend on whether the investigators cleared the chat history between questions. The DISCERN score was plotted as a function of the number of questions in each study. In a sensitivity analysis, we examined the relationship in tests that reported a query protocol consistent with deleting the question history and in tests that did not report query protocols.

(3) Number of evaluators: The hypothesis was that scores averaged from multiple evaluators could be expected to be more accurate. Extensive psychology literature supports the superiority and effectiveness of group decision-making (even when there is no communication within the group), attributed to "crowd wisdom" and aggregation effects.[13] We recorded the number of evaluators who determined the DISCERN score in each study. When studies posted scores for different evaluator groups, we recorded the average score posted by the most knowledgeable group; for example, one study assessed scores from patients, senior dentistry students, and orthodontists.[14] In this example, we only recorded the scores from the orthodontists. The DISCERN score was plotted as a function of the number of evaluators.

(4) Journal ranking: We hypothesized that studies reporting positive results (high DISCERN scores) could be favored by positive outcome bias and be published in higher-ranking journals (for example, see Easterbrook et al. and Emerson et al.).[15,16] There are several journal ranking metrics. We elected to use the Journal Citation Indicator (JCI: Clarivate, London, UK). JCI measures the citation impact, averaged over 3 years, for various research fields. Thus, a journal with a JCI of 1.25 has a 25% greater citation impact than the average in that category. Because the parameter is normalized, it enables comparison across fields, which, given the range of disciplines covered in the included studies, offers an advantage over parameters such as Journal Impact Factor and SCImago Journal Rank. We plotted the DISCERN score as a function of each journal's JCI.

(5) Chatbot version/year studied: Examination of these potential influences on response quality is challenging because of differences in chatbot capabilities and because of chatbot updates. Therefore, we created four groups based on the categorization of the chatbots' ability to provide quality responses and a temporal component reflecting when the study was conducted. For chatbot categorization, we divided the chatbots used into those with higher- versus lower-quality response capabilities. We based these decisions on the performance of chatbots answering medical questions, such as those in the United States Medical Licensing Examination (USMLE),[17–19] and an assessment of the grey literature. The higher-quality group included all ChatGPT-4 versions, as well as Perplexity, Gemini Advanced and Gemini 1.5 versions, and all Claude versions. The lower-quality group included all ChatGPT-3 versions, as well as Copilot, Grok, Chatsonic, and Gemini 1.0. The temporal component was divided into two periods: 2023 and 2024. One study was conducted in 2025 and was grouped with the 2024 studies. Many studies provided the date when the questions were asked or the year was inferred from the submission date. When the date was unspecified and could not be determined from submission dates, we excluded the study from this analysis.

Multiple linear regression model. The parameters identified as potentially influencing the DISCERN score were incorporated into a multiple linear regression model. We aimed to construct an explanatory rather than a predictive model for the DISCERN scores. There was missing data in the final model. Therefore, we determined if the data were missing completely at random (MCAR) using Little's test. We calculated variance inflation factors (VIF) in the final model to determine if collinearity was present. In addition, we assessed heteroskedasticity, skewness, and kurtosis using Cameron and Trivedi's decomposition test.

Prompt engineering. Appropriate use of prompt wording when asking questions of chatbots is an emerging field.[20,21] The apparent high level of education required to read the responses in most of the included studies led us to examine how prompts could be modified to enhance readability. In a pilot study, we selected six questions from four of the included studies.[22–25] This convenience sample of questions was chosen because the original responses left ample room for improvement in readability. These six responses fell within either the college-level or the post-graduate/professional ranges. We conducted three tests for each selected question. (1) Using the same prompt as in the original study but using ChatGPT-4o and adding the phrase, *"write the response all in text with no tables or bullet points"*. Tables and bullet points cause problems for FRES calculators because the calculations are based on sentence length.[26] (2) Using the original prompt with the following addition, *"Write your response all in text with no tables or bullet points. Write your response at a sixth-grade reading level."* (3) Using the following prompt, *"You are a physician responding to a patient's question. [original prompt question inserted here]. Write your response all in text with no tables or bullet points. You are unsure of the patient's knowledge of this topic, so write your response at a sixth-grade reading level."* The responses were assessed using an online FRES/FKGL calculator (as mentioned above). We cleared the chatbot history between each prompt and varied the order in which the prompts were used.

Certainty of evidence. We used the GRADE tool to assess the certainty of evidence in the included studies.[27]

Statistical analysis. Values are presented as means with 95% confidence intervals (CI). All analyses were conducted using Stata (version 18.0; StataCorp, College Station, TX).

Results

We identified 42 studies that conducted a total of 86 tests (Figure 1).[22–25,28–65] The most common reasons for exclusion were: (1) that the texts evaluated were patient educational material (for example, brochures), (2) the lack of DISCERN scores because studies used other quality assessment methods, and (3) the lack of both a DISCERN score and FRES or FKGL.

The studies primarily originated from the US (40%) and Türkiye (36%). The others were from Europe (17%), with two studies from Australia, and one from South Korea.

The study characteristics are provided in the supplementary material. Most studies (25 of 42) tested a single chatbot: ChatGPP-3/3.5 was used in 76% of these. Four studies tested two chatbots, six studies tested three, four studies tested four, and three studies tested six. The chatbots tested were mainly ChatGPT variants (version 3 – 35%; version 4 – 27%). Gemini and Copilot each contributed 16% of the tests. Claude was used in three tests, while Chatsonic, Perplexity, and Grok were used once.

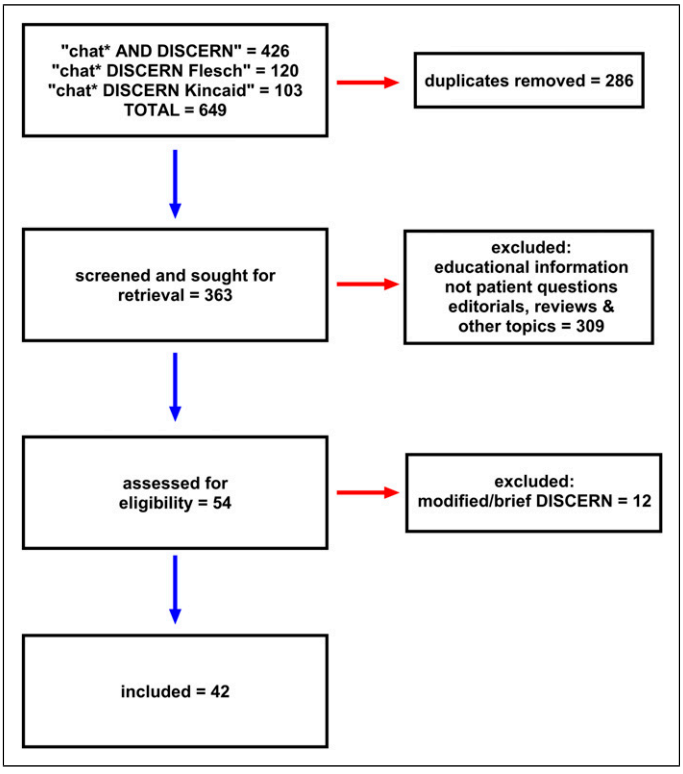


Figure 1. Flowchart showing the study selection process.

Query protocols

Thirty-one studies (74%) provided no details of their query protocols. Eleven studies (26%) provided protocol details, suggesting that they cleared the question history before asking the next one. However, the deletion of query history was not always explicitly stated.

Quality and readability

Forty-nine (57%) of the 86 DISCERN tests were ranked in the "good" quality range or higher. In contrast, only three sets of test responses were scored below college-level readability using the FRES calculation (FRES > 50); two were from Gemini and one from ChatGPT-3.5. An additional seven tests, that only reported FKGL, were below college reading level (FKGL $\leq$ 12); two were from Gemini, two from Gemini 1.5, and one each from Copilot, Perplexity, and ChatGPT-4o. Thus, a total of 10 test responses (12%) were below college reading levels.

We found an association between the DISCERN score and FRES: as response quality increased, readability decreased (adjusted $R^2 = 0.06$; $p = 0.02$; $n = 74$; Figure 2). There was no evidence of heteroskedasticity ($p = 0.48$). When we restricted analysis to only ChatGPT-3 and ChatGPT-3.5, the adjusted R-squared value increased (0.14; $p = 0.028$; $n = 29$; Figure 3). In contrast, we found no association for ChatGPT-4 (adjusted $R^2 = 0.03$; $p = 0.23$; $n = 20$; Figure 3). For ChatGPT-4 tests, both the DISCERN score and FRES increased, relative to the ChatGPT3/3.5 values. The latter

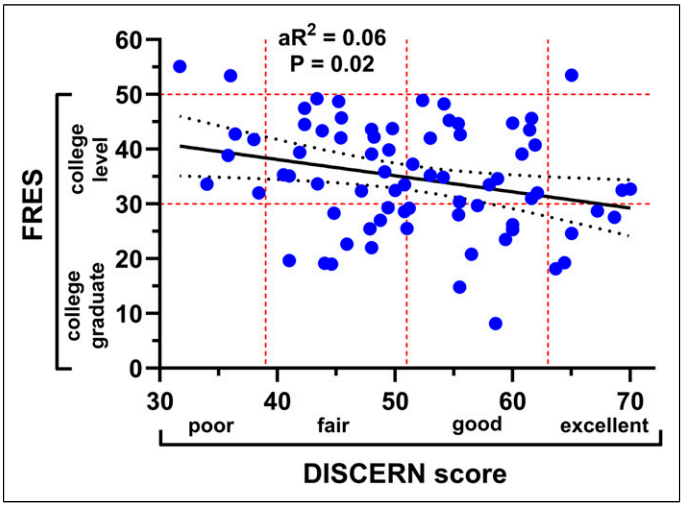


Figure 2. Readability (Flesch Reading Ease Score - FRES) plotted as a function of response quality (DISCERN score) for the 74 tests that measured both parameters. Response readability decreased as quality increased. The red vertical dashed lines indicate the categorical divisions of the DISCERN score, while the red horizontal dashed lines indicate the divisions of the FRES. The black dotted lines represent the 95% confidence intervals of the regression. aR^2 is the adjusted R-squared value.

increase was smaller than the former. These changes shifted the data up and to the right, thereby flattening the regression (Figure 3).

There was no relationship between the DISCERN score and FKGL ($p = 0.80$; $n = 75$). This lack of association appeared to be due to the FKGL data being range-restricted: the range between grades 10 and 14 contained 83% of the data.

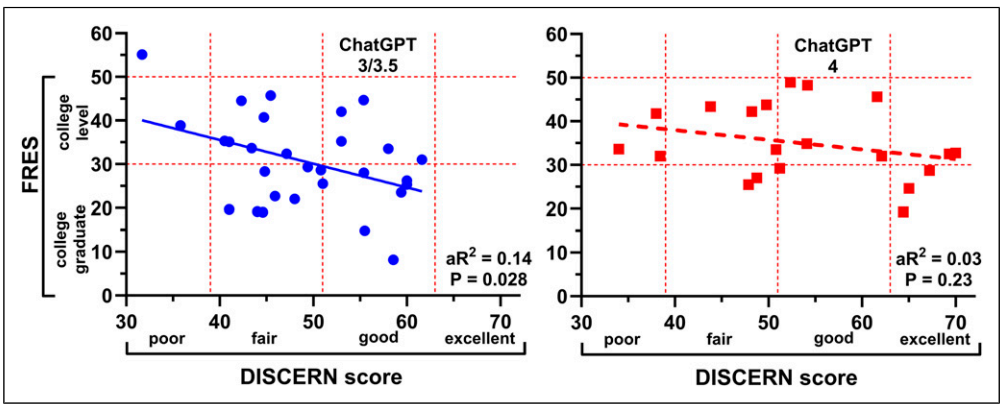


Figure 3. Readability (FRES) is plotted as a function of the DISCERN score for different versions of ChatGPT. For ChatGPT-3 versions (left panel), we found an association between FRES and DISCERN. This association was absent for ChatGPT-4 versions (right panel).

Number of prompt words

We found an association between the number of prompt words and the DISCERN score (adjusted $R^2 = 0.04$; $p = 0.03$; $n = 86$; Figure 4); the DISCERN score increased as the number of prompt words increased. Because this result was potentially influenced by the clustering of studies at approximately 10 and 40 words, we applied a reciprocal transform (1/X) to mitigate this effect. Again, there was an association (adjusted $R^2 = 0.05$; $p = 0.017$; Figure 4 inset), with no evidence of heteroskedasticity ($p = 0.22$).

Number of questions

Again, we found data clustering, so a reciprocal transform was applied. There was an association between the DISCERN score and the reciprocal of the number of questions: that is, the DISCERN score increased as the number of questions increased (adjusted $R^2 = 0.07$; $p = 0.009$; $n = 86$), with no evidence of heteroskedasticity ($p = 0.61$).

In the sensitivity analysis, we found that tests with no query protocol exhibited a stronger association between the DISCERN score and the reciprocal of the number of questions (adjusted $R^2 = 0.39$; $p < 0.001$; $n = 44$). Although there was also an association for tests that likely deleted query history ($p = 0.01$; $n = 42$), the sign of the coefficient was opposite; i.e., the DISCERN score decreased as the number of questions increased.

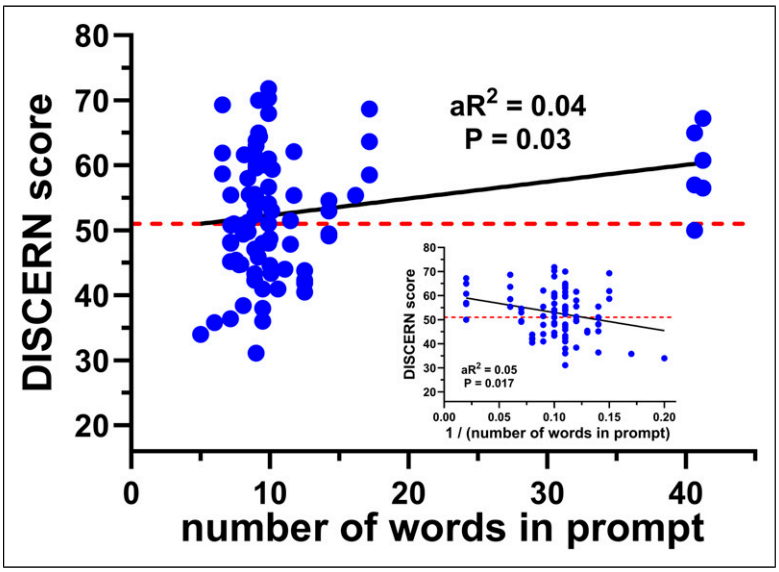


Figure 4. The DISCERN score is plotted as a function of the average number of words in the prompt for each test. DISCERN score increased as the number of prompt words increased. The horizontal red dashed line represents the threshold for a "good" quality response. The inset graph shows the transformed data (1/X) and the de-clustering of the points.

Number of evaluators

There was considerable variation in the number of evaluators: one study used a single evaluator, 18 used two, 12 used three, one used four, eight used more than four, and two studies did not specify the number (one of these used a "panel" – we assumed ≥3). We converted the number of evaluators into a binary parameter, dividing studies with fewer than three evaluators (38 tests) and those with three or more evaluators (47 tests). We excluded the one test in which the number of evaluators was unspecified.

The data were again clustered. After applying a reciprocal transform, there was no association between DISCERN score and the number of evaluators (adjusted $R^2 = 0.02$; $p = 0.09$; $n = 82$). As a binary variable, tests with more than two evaluators had higher DISCERN scores (55.1: 95% CI [52.6 to 57.5]) than tests with one or two evaluators (50.8: 95% CI [47.4 to 54.2]; $p = 0.04$). The increase represents a shift from "fair" to "good" for the quality category.

Journal ranking

We found a negative association between the DISCERN score and JCI. The DISCERN score increased as the JCI decreased (adjusted $R^2 = 0.04$; $p = 0.045$; $n = 72$; Figure 5), with no heteroskedasticity ($p = 0.33$).

Fourteen tests were published in journals that did not have a JCI. The DISCERN score in these did not differ from that for tests published in journals with a JCI (52.3: 95% CI [46.2–58.3] vs 52.9: 95% CI [50.7–55.1]; $p = 0.84$).

Chatbot version/year

We observed small temporal increases in DISCERN score and slightly greater increases with chatbot capabilities (Table 1). These changes are consistent with chatbot updates and the use of specific chatbots, producing higher DISCERN scores. The data in Table 1 also indicated a temporal

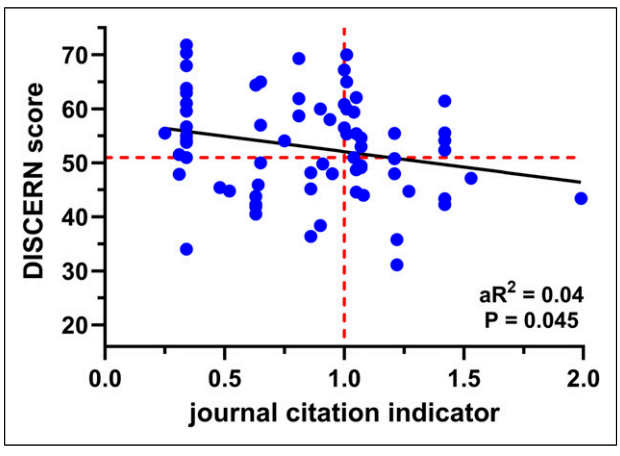


Figure 5. DISCERN score plotted as a function of the Journal Citation Indicator (JCI). The DISCERN score increased as JCI decreased. The vertical red dashed line indicates the average citation impact in a journal's subject category. The horizontal red dashed line represents the threshold for a "good" quality response.

Table 1. Chatbot temporal and version changes in DISCERN score.

	2023	2024
Lower capability	48.9 [44.8–53.0] (n = 18)	51.6 [48.3–54.9] (n = 30)
Higher capability	54.8 [48.7–60.9] (n = 6)	56.3 [51.6–61.0] (n = 22)

shift toward increased testing of chatbots with greater capability for providing good-quality medical responses.

We were unable to determine the year when the tests were performed in 10 of the studies. There was no difference in the DISCERN score for such tests (54.7: 95% CI [47.9–61.4) than in the test for which the year was identified (52.6: 95% CI [50.4–54.7]; p = 0.51).

Multiple linear regression model

None of the individual parameters alone provided a good explanation of the variance in the DISCERN score. Therefore, we examined different parameter combinations to maximize the model's parsimony and explanatory ability. Table 2 shows the optimal combination. The associations with the reciprocal of the number of questions and the chatbot version/year parameter remained in the adjusted model. There was weak evidence of an association with JCI. The model's adjusted R^2 value was 0.23. There was no evidence of heteroskedasticity (p = 0.41), skewness (p = 0.40) or kurtosis (p = 0.12). The average VIF score was 1.08 (range 1.02–1.12), indicating no collinearity.

Prompt engineering

ChatGPT-4o responses to the six chosen questions produced marginally higher FRES than those in the original study (28 [95% CI [18–37] vs 23 [95% CI [7–38]]), but the average remained in the post-graduate range (Figure 6). In contrast, readability increased dramatically when the prompt included an instruction to write responses at a sixth-grade reading level (FRES 75 [95% CI [67–83]]). Two of the 12 responses achieved a sixth-grade level. The majority achieved a seventh-grade reading level. The prompt containing additional context produced a similar increase in the average FRES (72 [95% CI [69–75]]).

Certainty of evidence

All of the studies were observational and so started at a GRADE level of "low".[66] The GRADE categories of "inconsistency" and "indirectness" did not apply.

We classified the risk of bias category for the outcomes of DISCERN score and readability as serious for all studies. None of the studies indicated that the evaluators received training and practice in using the DISCERN tool. The knowledge and experience of the evaluators were seldom reported. The knowledge and experience range spanned patients (we excluded data from patient evaluations), medical students, residents, fellows, and board-certified physicians. The failure to report the query protocol suggests that some studies did not clear the chat history after each question. Consequently, those studies may have provided context for the chatbots, which could lead to better-quality responses to subsequent questions. Very few studies reported the exact method for evaluating

Table 2. Multiple linear regression model (n = 65).

	Coefficient	95% confidence intervals	p-value
Number of questions (reciprocal)	-53	-90 to -16	0.006
Chatbot version/year	2.1	0.2 to 4.0	0.029
JCI	-4.8	-10.3 to 0.8	0.09
Constant	58	51 to 65	

The mean VIF = 1.08.

readability; i.e., which calculator was used. Therefore, we concluded that both outcomes were potentially biased.

The imprecision GRADE category was also a source of concern. When studies reported individual evaluator DISCERN scores for the same question, there was sometimes a considerable range (as much as 60 points). Also, different readability calculators handle the presence of bullet points, numbered points or sections, and references in various ways. Consequently, different calculators assign FRES scores to the same piece of text that can differ by as much as 10 points. Thus, both outcomes were associated with imprecision.

Therefore, we downgraded all studies to the "very low" category of certainty.

Discussion

We found that the readability of chatbot responses to patient questions decreased as the quality of the responses increased. Furthermore, response quality was associated with the number of questions asked and a parameter that reflected the chatbot version and the study year. We demonstrated that response readability can be readily affected by the choice of prompt wording; improving readability is a straightforward process. Our synthesis of the early investigations into patients' potential use of chatbots to seek medical information provides insight into the advantages and disadvantages of this approach.

Readability

The high level of reading ability required to comprehend the responses is expected given the inherent complexity of medicine, especially the frequent occurrence of multisyllabic words. Nonetheless, several studies report increased readability by adjusting the prompt, a technique known as prompt engineering. One study added the phrase "in simple terms" to the prompt, which increased FRES from college to high school level (39–55).[28] However, there was also a slight decrease in the DISCERN score: from 36 to 32 (both in the "poor" quality range). Similarly, investigators in another study (excluded from our analysis because they did not use DISCERN) included the instruction "*to produce easily readable material at a sixth-grade reading level*" in the prompt. The responses had an average FRES of 58, corresponding to ninth grade.[67] This was a higher FRES than achieved by any study included in our analysis (Figure 2). We also used prompt engineering to improve readability using a small sample of questions from the included studies. Readability was enhanced to eighth-grade or lower reading levels by including the phrase, *"write the response at a sixth-grade reading level"* (Figure 6).

The concept that how questions are asked is a critical determinant of chatbot response has only recently started to spread from computer science to medicine and the public. To date, this progress is

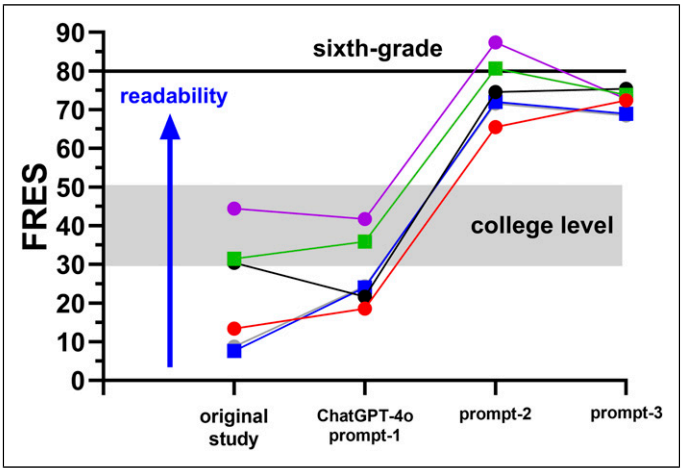


Figure 6. Readability (FRES) can be influenced by the prompt (see text for details). Six questions from four of the included studies were run using the same prompt in ChatGPT-4o and then with additional wording requesting that the response be written at a sixth-grade reading level. This approach produced a substantial increase in response readability.

limited. Searches (in Google Trends) for the term *"prompt engineering"* have not increased in the past 16 months, despite increases in searches for *"chatbot"* and *"ChatGPT"* (data not shown).

Similarly, a PubMed search for the term *"prompt engineering"* indicated that the phrase first appeared there in 2022 (two articles), increased to 27 articles in 2023, 213 in 2024, and 254 so far (August 7th) in 2025. If the search term *"Chat*"* was added to the query, this change reduced the number of articles to 18 in 2023, 102 in 2024, and 92 so far in 2025. Therefore, it is unsurprising that few prompts were explicitly designed to enhance accuracy or readability for studies conducted between 2022 and 2024. Prompt engineering is a crucial skill for physicians and patients.[68] More effective prompts should increase both the accuracy and readability of chatbot responses. Primers on writing effective prompts are now published.[69]

There will be a trade-off between response quality and readability; however, prompt engineering will enable both to be tailored to desired levels.

Response quality

Quality is a subjective measure that depends on how the assessment is done and the assessors' knowledge. Despite its systematic approach, validation, and demonstrated reliability, the DISCERN instrument remains subjective. Many of the included studies reported DISCERN scores with considerable variation among evaluators. For example, evaluator scores had a 60-point range, from "very poor" to "excellent", for the same responses in one study.[14] Such wide ranges contribute to the scatter and low R-squared values, as well as the residual heterogeneity of our model.

Training using the DISCERN instrument is recommended.[8,70] However, the number of studies that included training in their protocols is unknown; none reported training. A lack of training may lead physicians to unconsciously rely on their professional expectations and, therefore, assign greater value to technical detail over clarity. Such bias could skew DISCERN scores downward for highly readable resources. This issue likely contributes to score heterogeneity.

We should also consider whether the DISCERN instrument is appropriate for evaluating chatbot responses. This tool has been validated,[8] and its structured format provides consistency. On the other hand, DISCERN was designed to evaluate written material rather than the dynamic conversational interaction patients would have with chatbots. Modifying DISCERN to incorporate chatbot-specific elements or combining it with tools that assess different aspects of chatbot-patient interaction might enhance evaluation.

Nevertheless, DISCERN is currently the most widely used tool for evaluating response quality. Other evaluation tools, such as the modified and brief versions of DISCERN, DISCERN-AI, EQIP, global quality score (GQS) and Likert scales, also have deficiencies. Some are yet to be validated, some are more subjective and lack consistent application between studies, and some lack the granularity required to provide thorough evaluations.

Factors associated with DISCERN score

Prompt words. Our original premise was that response quality would be associated with the number of prompt words. However, the analysis did not support this idea. Although we found an association between the DISCERN score and prompt words in the initial regression, a small number of questions, each containing around 40 words, appeared to be responsible. We based our hypothesis on the concept that more words equal more specific questions and increased context. Hence, such questions yield more accurate and definitive answers.[68] That concept may, in general, be correct and is supported by five of the six ~40-word prompts achieving "good-to-excellent" DISCERN scores. Nevertheless, even short questions can be specific enough to yield accurate and definitive answers. For example, the four-word question, *"What are kidney stones?"* requires no context and is explicit. Therefore, even such short questions could be expected to yield a response with "good" to "excellent" quality.

Number of questions. The number of questions remained associated with the DISCERN score in the multiple linear regression model (Table 2). This parameter would not be expected to affect the DISCERN score directly if the query history is deleted before each question, thereby eliminating potential contextual gains. Conversely, if the history is not deleted, the additional context could boost DISCERN scores. The query protocol was unspecified in the majority of studies, which suggests no history deletion. The sensitivity analysis results were consistent with the idea that the lack of a query protocol (and hence likely no history deletion) and asking a larger number of questions provided context, which increased the average DISCERN score.

The number of questions could also indirectly affect the DISCERN score because a larger number increases the likelihood of including questions that play to the chatbot's strengths, thereby increasing the average score. Furthermore, more questions would reflect the chatbot's actual performance level because outlier scores would have less influence on the overall average.

Chatbot version/year. Studies have indicated differences in accuracy between chatbots in other types of tests. For example, 200 questions in the style of those used in the USMLE were posed to five chatbots. Claude and ChatGPT-4 scored the highest percentage of correct answers (83% and 82%, respectively). Three other chatbots achieved lower scores: Copilot 60%, ChatGPT-3.5 58%, and Gemini 54%.[19]

Such comparisons are challenging because chatbot capabilities are evolving. We combined chatbot type and temporal changes to create a categorical parameter. This grouping was positively associated with the DISCERN score. However, our chatbot categorizations may be questioned

because there are conflicting assessments of their quality. Therefore, as a sensitivity analysis, we restricted the categorical parameter to the chatbots for which quantitative assessment is available: ChatGPT versions, Claude, Gemini 1.0, and Copilot. In the multiple linear regression model, the adjusted R-squared value increased to 0.24 (from 0.23), and the *p*-values for reciprocal questions and chatbot version/year remained unchanged. The P-value for JCI increased to 0.16 from 0.086. These results indicate the robustness of the chatbot version/year parameter.

Journal citation indicator. JCI can serve as a proxy for journal impact. Because JCI values are normalized, comparisons can be made across fields. However, the values represent three-year averages, which may not accurately reflect journal impact in rapidly evolving fields such as chatbot research.

We anticipated that studies with higher DISCERN scores would be published in journals with a higher JCI. Therefore, we were surprised to find the opposite in the crude regression analysis. The observed relationship could be consistent with higher-ranking journals being biased against chatbot use based on current perceptions of poor response accuracy, errors, and their propensity to hallucinate.[2,71] Still, there are many factors involved in journals' publication decisions. Moreover, the coefficient was relatively small; contributing only 9.6 DISCERN units over the entire range, which is less than the width of a single DISCERN category. In the adjusted model, there was only weak evidence of an association. Therefore, JCI represents, at most, a minor factor.

Number of evaluators. We found that the average DISCERN score in studies with more than two evaluators was higher and in the "good" range, compared to studies with one or two evaluators (in the "fair" range). More evaluators will reduce the influence of outlier scores. The evaluator's knowledge and training in using DISCERN will also contribute to the score. Training was never mentioned in any of the studies, while the evaluators' knowledge was sometimes mixed. For example, evaluators in one study included residents and faculty.[56] Therefore, the lack of association between DISCERN score and number of evaluators in the crude regression was not surprising. These issues, combined with the lack of a clear expectation regarding whether, or how, more evaluators would influence the DISCERN score, prompted us to exclude this parameter from the regression model.

Limitations. There are several limitations to our analysis. First, as evident in all the graphs, there is considerable heterogeneity. In addition to the frequent lack of experimental detail provided in the studies, other variables, such as subject matter and the specific questions asked, could not be adjusted for. Nonetheless, even under these circumstances, the observed association between response readability and quality, the ability to enhance readability through prompt engineering, and the interpretation of the regression model provide insight.

Second, chatbot research is rapidly evolving; additional studies will have been published since the conclusion of our literature search. Furthermore, ongoing updates may result in improved response quality over time. We only searched PubMed, and, therefore, it is possible that studies published in engineering fields that included physician authors were missed.

Third, almost all the studies asked each question only once. Therefore, the repeatability and reproducibility of the results are unknown. Additionally, this piecemeal approach differs from how patients typically interact with chatbots. Patients could seek clarification for some responses; none of the included studies did this. Thus, the questions posed do not reflect how patients' conversations with chatbots might progress. Clarification and context will likely enhance the quality of responses.

The included studies were predominantly from the US and Türkiye and were all conducted in English, which may result in bias. Although the influences of location and language have yet to be systematically evaluated, one study reported that responses to the same healthcare question appeared to depend on the user's location.[72] The authors assessed responses from ChatGPT-3.5, Google Bard and Bing in the US, Indonesia, Nigeria, and Taiwan in November 2023. Further work is required to determine the extent and precise nature of potential location bias.

The final model was missing 21 chatbot tests; almost a quarter of the total. Therefore, missing data might introduce bias. Little's test provided no evidence against MCAR; however, it does not prove MCAR. The sensitivity analyses we performed supported MCAR. Consequently, we do not believe that missing data had an adverse effect on the results or our interpretations.

Pessimism or optimism. There has been optimism that chatbots will benefit patients, physicians, and their interactions. Although generally optimistic, many articles exploring the possibilities have outlined the challenges alongside the potential advantages of chatbots and large language models.[73–75] On the other hand, comments in chatbot studies are more pessimistic. For example, *"ChatGPT-generated responses… were outdated and failed to provide an adequate foundation for patients' understanding…"*[24] and *"… the information presented was difficult to read, with varying quality, understandability, accuracy, and comprehensiveness…"*.[76] Even studies that reported high DISCERN scores were circumspect in their appraisal and focused on readability and reference sources, *"There was generally high quality in the answers given… but there was a high reading level required… However, it is unclear where the answers originated, with no source material cited"*[47] Therefore, the current overall opinion appears to be one of skepticism.

Some chatbots that provide higher-quality responses currently require paid subscriptions. The free-to-use chatbots generally yielded lower-quality responses. This difference could result in access disparity.

Future work

The assessment of chatbot responses to patients' questions is a new area of research; the first studies included in our analysis were published in 2023. Consequently, there are no established standards for conducting such studies. The GRADE assessment of "very low" and the considerable heterogeneity in the data indicate that some standards should be established. We recommend that evaluators be trained in using the DISCERN tool and practice its use before conducting a study. Having more than two evaluators appears warranted to mitigate the influence of overly positive or negative scores. Even with the subjectivity of DISCERN, it is a validated instrument that provides a granular assessment of responses. This is an advantage over Likert scales, which, although easy to apply, fail to allow inter-study comparisons. The query protocol should be provided, especially a statement of whether the chat history was deleted between questions. The current studies primarily pose questions that do not accurately reflect how patients are likely to interact with chatbots. More natural question sequences will likely improve response quality because of the added context. Greater experimental detail would help synthesis; for example, the date the study was conducted, the specific chatbot used, and the calculator used to assess readability.

Conclusion

The current quality and readability of chatbot responses are inadequate for patient use. Many responses did not reach "good" quality, few were written below a college reading level, and none achieved the recommended sixth-grade level. These shortcomings could taint the perceptions of healthcare professionals and patients, particularly if widely disseminated. However, the potential utility of chatbots should not be dismissed, especially at this early stage in their development. Improvements in question structure, achieved through prompt engineering, will enable readability and quality to be tailored to the knowledge levels of specific patient populations.

ORCID iD

Peter Whittaker https://orcid.org/0000-0002-5627-2166

Author contributions

Both authors contributed to all aspects of the study.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The data that support the findings of this study are available from the authors upon reasonable request.

Registration

Neither the review nor the protocol was registered.

Supplemental Material

Supplemental material for this article is available online.

References

1. Hopkins AM, Logan JM, Kichenadasse G, et al. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr* 2023; 7: pkad010.
2. Webster P. Medical AI chatbots: are they safe to talk to patients? *Nat Med* 2023; 29: 2677–2679.
3. Lee K, Hoti K, Hughes JD, et al. Dr Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. *J Med Internet Res* 2014; 16: e262.
4. Stukus DR. How Dr Google is impacting parental medical decision making. *Immunol Allergy Clin* 2019; 39: 583–591.
5. Omiye JA, Gui H, Rezaei SJ, et al. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med* 2024; 177: 210–220.

6. Sun M, Reiter E, Kiltie AE, et al. Effectiveness of ChatGPT in explaining complex medical reports to patients. Epub ahead of print 23 June 2024. DOI: 10.48550/arXiv.2406.15963.
7. Charnock D. *The DISCERN handbook quality criteria for consumer health information on treatment choices*. Radcliffe Medical Press, 1998.
8. Charnock D, Shepperd S, Needham G, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 1999; 53: 105–111.
9. Flesch R. A new readability yardstick. *J Appl Psychol* 1948; 32: 221–233.
10. Arnold CL, Davis TC, Frempong JO, et al. Assessment of newborn screening parent education materials. *Pediatrics* 2006; 117: S320–S325.
11. Hersh L, Salzman B and Snyderman D. Health literacy in primary care practice. *Am Fam Physician* 2015; 92: 118–124.
12. Kincaid JP, Fishburne J, Robert PR, et al. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Defense Technical Information Center. DOI: 10.21236/ADA006655. Epub ahead of print 1 February 1975.
13. Tindale RS and Winget JR. *Group decision-making*. Oxford University Press, 2019.
14. Kurt Demirsoy K, Buyuk SK and Bicer T. How reliable is the artificial intelligence product large language model ChatGPT in orthodontics? *Angle Orthod* 2024; 94: 602–607.
15. Easterbrook PJ, Berlin JA, Gopalan R, et al. Publication bias in clinical research. *Lancet* 1991; 337: 867–872.
16. Emerson GB, Warme WJ, Wolf FM, et al. Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. *Arch Intern Med* 2010; 170: 1934–1939.
17. Bahir D, Zur O, Attal L, et al. Gemini AI vs. ChatGPT: a comprehensive examination alongside ophthalmology residents in medical knowledge. *Graefes Arch Clin Exp Ophthalmol* 2025; 263: 527–536.
18. Bolgova O, Ganguly P and Mavrych V. Comparative analysis of LLMs performance in medical embryology: a cross-platform study of ChatGPT, claude, gemini, and copilot. *Anat Sci Educ* 2025; 18: 718–726.
19. Mavrych V, Yousef EM, Yaqinuddin A, et al. Large language models in medical education: a comparative cross-platform evaluation in answering histological questions. *Med Educ Online* 2025; 30: 2534065.
20. Chen B, Zhang Z, Langrené N, et al. Unleashing the potential of prompt engineering for large language models. *Patterns (N Y)* 2025; 6: 101260.
21. Russe MF, Reisert M, Bamberg F, et al. Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. *Rofo* 2024; 196: 1166–1170.
22. Anastasio AT, Mills FB, Karavan MP, et al. Evaluating the quality and usability of artificial intelligence–generated responses to common patient questions in foot and ankle surgery. *Foot Ankle Orthop* 2023; 8: 24730114231209919.
23. Eng E, Mowers C, Sachdev D, et al. Chat generative pre-trained transformer (ChatGPT) – 3.5 responses require advanced readability for the general population and may not effectively supplement patient-related information provided by the treating surgeon regarding common questions about rotator cuff repair. *Arthroscopy* 2025; 41: 42–52.
24. Johns WL, Martinazzi BJ, Miltenberg B, et al. ChatGPT provides unsatisfactory responses to frequently asked questions regarding anterior cruciate ligament reconstruction. *Arthroscopy* 2024; 40: 2067–2079.e1.
25. Lack BT, Mouhawasse E, Childers JT, et al. Can ChatGPT answer patient questions regarding reverse shoulder arthroplasty? *Journal of ISAKOS* 2024; 9: 100323.
26. Jindal P and MacDermid J. Assessing reading levels of health information: uses and limitations of flesch formula. *Educ Health* 2017; 30: 84–88.
27. Granholm A, Alhazzani W and Møller MH. Use of the GRADE approach in systematic reviews and guidelines. *Br J Anaesth* 2019; 123: 554–559.

28. Abou-Abdallah M, Dar T, Mahmudzade Y, et al. The quality and readability of patient information provided by ChatGPT: can AI reliably explain common ENT operations? *Eur Arch Otorhinolaryngol* 2024; 281: 6147–6153.
29. Alyanak B, Dede BT, Bağcıer F, et al. Parental education in pediatric dysphagia: a comparative analysis of three large language models. *J Pediatr Gastroenterol Nutr* 2025; 81: 18–26.
30. Artioli E, Veronesi F, Mazzotti A, et al. Assessing ChatGPT responses to common patient questions regarding total ankle arthroplasty. *J Exp Orthop* 2025; 12: e70138.
31. Aydilek A and Karadamar ÖL. How reliable are ChatGPT and Google's answers to frequently asked questions about unicondylar knee arthroplasty from a scientific perspective? *J Orthop Surg* 2025; 33: 10225536251350411.
32. Collins CE, Giammanco PA, Guirgus M, et al. Evaluating the quality and readability of generative artificial intelligence (AI) chatbot responses in the management of achilles tendon rupture. *Cureus* 2025; 17: e78313.
33. Crook BS, Park CN, Hurley ET, et al. Evaluation of online artificial intelligence-generated information on common hand procedures. *J Hand Surg Am* 2023; 48: 1122–1127.
34. Demir S. Evaluation of responses to questions about keratoconus using ChatGPT-4.0, google gemini and microsoft copilot: a comparative study of large language models on keratoconus. *Eye Contact Lens* 2025; 51: e107–e111.
35. Demir S. Investigating the role of large language models on questions about refractive surgery. *Int J Med Inf* 2025; 195: 105787.
36. Durmaz Engin C, Karatas E and Ozturk T. Exploring the role of ChatGPT-4, BingAI, and gemini as virtual consultants to educate families about retinopathy of prematurity. *Children* 2024; 11: 750.
37. Fahy S, Oehme S, Milinkovic D, et al. Assessment of quality and readability of information provided by ChatGPT in relation to anterior cruciate ligament injury. *J Phys Math* 2024; 14: 104.
38. Fahy S, Oehme S, Milinkovic DD, et al. Enhancing patient education on the role of tibial osteotomy in the management of knee osteoarthritis using a customized ChatGPT: a readability and quality assessment. *Front Digit Health* 2024; 6: 1480381.
39. Fahy S, Niemann M, Böhm P, et al. Assessment of the quality and readability of information provided by ChatGPT in relation to the use of platelet-rich plasma therapy for osteoarthritis. *J Phys Math* 2024; 14: 495.
40. Gezer MC and Armangil M. Assessing the quality of ChatGPT's responses to commonly asked questions about trigger finger treatment. *Ulus Travma Acil Cerrahi Derg* 2025; 31: 389–393.
41. Giammanco PA, Collins CE, Zimmerman J, et al. Evaluating the quality and readability of information provided by generative artificial intelligence chatbots on clavicle fracture treatment options. *Cureus* 2025; 17: e77200.
42. Gilmore N, Kushner JN, Redden A, et al. Assessing ChatGPT responses to common patient questions on knee osteoarthritis. *J Orthop Exp & Innov*. DOI: 10.60118/001c.121815. Epub ahead of print 1 November 2024.
43. Günay AE, Özer A, Yazıcı A, et al. Comparison of ChatGPT versions in informing patients with rotator cuff injuries. *JSES Int* 2024; 8: 1016–1018.
44. Gupta S, Haislup BD, Hoffman RA, et al. Assessing information provided via artificial intelligence regarding distal biceps tendon repair surgery. *J Exp Orthop* 2025; 12: e70281.
45. Gupta S, Tarapore R, Haislup B, et al. Microsoft copilot provides more accurate and reliable information about anterior cruciate ligament injury and repair than ChatGPT and google gemini; however, no resource was overall the best. *Arthrosc Sports Med Rehabil* 2025; 7: 101043.
46. Hones K, Krisanda E and Chim H. Caution regarding ChatGPT's appropriateness and reliability regarding surgery for wrist arthritis. *Hand* 2025; 20: 910–916.

47. Hurley ET, Crook BS, Lorentz SG, et al. Evaluation high-quality of information from ChatGPT (artificial intelligence—Large language model) artificial intelligence on shoulder stabilization surgery. *Arthroscopy* 2024; 40: 726–731.e6.
48. Incerti Parenti S, Bartolucci ML, Biondi E, et al. Online patient education in obstructive sleep apnea: ChatGPT versus google search. *Healthcare* 2024; 12: 1781.
49. Kalem M, Balaban K, Kocaoğlu H, et al. Evaluation of ChatGPT responses to common patient questions on ankle fusion. *Foot Ankle Surg* 2025; S1268-7731(25): 00117.
50. Kayabaşı M, Köksaldı S and Durmaz EC. Evaluating the reliability of the responses of large language models to keratoconus-related questions. *Clin Exp Optom* 2024: 1–8.
51. Keating M, Bollard SM and Potter S. Assessing the quality, readability, and acceptability of AI-Generated information in plastic and aesthetic surgery. *Cureus* 2024; 16: e73874.
52. Kılınç DD and Mansız D. Examination of the reliability and readability of chatbot generative pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. *Am J Orthod Dentofacial Orthop* 2024; 165: 546–555.
53. Kolac UC, Karademir OM, Ayik G, et al. Can popular AI large language models provide reliable answers to frequently asked questions about rotator cuff tears? *JSES Int* 2025; 9: 390–397.
54. Lieu B, Crawford E, Laubach L, et al. Patient education strategies in pediatric orthopaedics: using ChatGPT to answer frequently asked questions on scoliosis. *Spine Deform* 2025; 13(5): 1377–1389. DOI: 10.1007/s43390-025-01087-y, Epub ahead of print 5.
55. Lim B, Lirios G, Sakalkale A, et al. Assessing the efficacy of artificial intelligence to provide peri-operative information for patients with a stoma. *ANZ J Surg* 2025; 95: 464–496.
56. Lim B, Seth I, Cuomo R, et al. Can AI answer my questions? Utilizing artificial intelligence in the perioperative assessment for abdominoplasty patients. *Aesthetic Plast Surg* 2024; 48: 4712–4724.
57. Lois A, Yates R, Ivy M, et al. Accuracy of natural language processors for patients seeking inguinal hernia information. *Surg Endosc* 2024; 38: 7409–7415.
58. Mishra A, Begley SL, Chen A, et al. Exploring the intersection of artificial intelligence and neurosurgery: let us be cautious with ChatGPT. *Neurosurgery* 2023; 93: 1366–1373.
59. Ozdemir ZM and Yapici E. Evaluating the accuracy, reliability, consistency, and readability of different large language models in restorative dentistry. *J Esthetic Restor Dent* 2025; 37: 1740–1752.
60. Sahin S, Erkmen B, Duymaz YK, et al. Evaluating ChatGPT-4's performance as a digital health advisor for otosclerosis surgery. *Front Surg* 2024; 11: 1373843.
61. Şahin Ş, Tekin MS, Yigit YE, et al. Evaluating the success of ChatGPT in addressing patient questions concerning thyroid surgery. *J Craniofac Surg* 2024; 35: e572–e575.
62. Şan H, Bayrakcı Ö, Çağdaş B, et al. Reliability and readability analysis of ChatGPT-4 and google bard as a patient information source for the most commonly applied radionuclide treatments in cancer patients. *Rev Española Med Nucl Imagen Mol* 2024; 43: 500021.
63. Warren CJ, Edmonds VS, Payne NG, et al. Prompt matters: evaluation of large language model chatbot responses related to Peyronie's disease. *Sex Med* 2024; 12: qfae055.
64. Xavier JL, Khoury J, Phen HM, et al. Evaluating the efficacy of natural language processing artificial intelligence models as a patient education tool for stature lengthening surgery and reconstruction. *Journal of Limb Lengthening & Reconstruction* 2024; 10: 22–27.
65. Yun JY, Kim DJ, Lee N, et al. A comprehensive evaluation of ChatGPT consultation quality for augmentation mammoplasty: a comparative analysis between plastic surgeons and laypersons. *Int J Med Inf* 2023; 179: 105219.
66. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011; 64: 401–406.

67. Kianian R, Sun D and Giaconi J. Can ChatGPT aid Clinicians in educating patients on the surgical management of glaucoma? *J Glaucoma* 2024; 33: 94–100.
68. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res* 2023; 25: e50638.
69. Lin Z. How to write effective prompts for large language models. *Nat Hum Behav* 2024; 8: 611–615.
70. Rees CE, Ford JE and Sheard CE. Evaluating the reliability of DISCERN: a tool for assessing the quality of written patient information on treatment choices. *Patient Educ Counsel* 2002; 47: 273–275.
71. Walters WH and Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep* 2023; 13: 14045.
72. Gumilar KE, Indraprasta BR, Hsu Y-C, et al. Disparities in medical recommendations from AI-based chatbots across different countries/regions. *Sci Rep* 2024; 14: 17052.
73. Kanjilal S. Flying into the future with large language models. *Clin Infect Dis* 2024; 78: 867–869.
74. Lee P, Bubeck S and Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; 388: 1233–1239.
75. Li R, Kumar A and Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or pandora's box? *JAMA Intern Med* 2023; 183: 596–597.
76. Bellinger JR, Kwak MW, Ramos GA, et al. Quantitative comparison of chatbots on common rhinology pathologies. *Laryngoscope* 2024; 134: 4225–4231.