Review paper

Quality assessment of ChatGPT-generated medical information for healthcare professionals and patients: a systematic review

Filip M. Tkaczyk[1], Joanna Gotlib-Małkowska[2], Zbigniew Siudak[3]

[1]Doctoral School, *Collegium Medicum*, Jan Kochanowski University, Kielce, Poland
[2]Department of Education and Research in Health Sciences, Faculty of Health Sciences, Medical University of Warsaw, Poland
[3]*Collegium Medicum*, Jan Kochanowski University, Kielce, Poland

Medical Studies
DOI: https://doi.org/10.5114/ms/208978

Key words: artificial intelligence, natural language processing, medical informatics.

Abstract

Introduction: ChatGPT, an artificial intelligence (AI)-based language model, has gained popularity in medicine, although concerns remain regarding the accuracy and clinical applicability of its generated content.
Methods: In accordance with PRISMA guidelines, a systematic review was conducted using the PubMed, EBSCO, and Scopus databases, including 29 studies published between 2022 and 2024. The review included publications evaluating the accuracy, readability, comprehensiveness, and clinical utility of medical information generated by ChatGPT.
Results: ChatGPT's content was generally consistent with clinical knowledge, although the quality varied depending on language, query formulation, and medical domain. Responses in English were usually more accurate. Limitations included lack of references, insufficient contextual and emotional understanding, and inconsistencies with medical guidelines in complex cases.
Conclusions: ChatGPT shows potential as an educational and clinical decision-support tool but requires further refinement and standardisation of evaluation methods.

Introduction

Artificial intelligence (AI) refers to advanced computational models that mimic human cognitive abilities. Examples include applications using artificial neural networks, which have the ability to self-learn and become more efficient at solving the tasks they are designed to perform. Examples of this are artificial neural networks – algorithms inspired by the functioning of the human brain – which are capable of learning from large datasets by recognising patterns and relationships, becoming increasingly effective over time at performing their designated tasks. Machine learning is the process through which a computer model "learns" from examples to independently generate responses or make decisions. AI algorithms can, among other things, recognise images, process speech, and analyse input text, enabling their widespread use, especially in medicine. These mathematical models are based on vast numbers of medical data, which are rapidly growing beyond the capabilities of manual human analysis. AI-based programs have produced remarkable results in many fields, and their development and implementation are being extensively researched and applied in almost every medical field. A review of medical information generated by ChatGPT is particularly important because it allows for an assessment of whether such tools can be used safely and effectively in patient education and in supporting the work of healthcare professionals – especially in light of the growing influence of artificial intelligence on the healthcare system [1, 2].

Analysis from the Digital 2024 Global Overview Report on the use of digital technologies and social media in 2023 shows that at the beginning of 2024, the number of mobile phone users reached 5.61 billion, and the number of internet users reached 5.35 billion, marking a 1.8% increase compared to the previous year. Social media recorded over 5 billion active users, spending an average of 2 hours and 23 minutes per day on these platforms, with TikTok and Instagram leading the way in terms of popularity [3, 4].

ChatGPT, developed by OpenAI, has become the fastest-growing application in history, reaching 100 million users in January 2024. Since its debut in November 2022, the application has attracted around 13 million users daily, outpacing the growth rate of platforms such as TikTok or Instagram. In its first month, ChatGPT attracted 57 million active users monthly, and in January it recorded 590 million visits. Analysts at the Union Bank of Switzerland (UBS) highlight the application's exceptional growth rate and point to its enormous potential and future opportunities. The evaluation of medical information generated by ChatGPT is important because, despite its popularity and ease of access, the model may provide incom-

plete or inaccurate responses, posing a risk of misinformation in the context of health and medical care [5].

The training process of ChatGPT uses advanced machine learning, including Reinforcement Learning from Human Feedback (RLHF) and supervised fine-tuning with the participation of human AI trainers. This model is trained on extensive datasets, enabling it to identify patterns and relationships in the data and then generate responses based on these patterns. A distinctive feature of ChatGPT is its iterative learning capability, which allows the quality of generated responses to be systematically improved through multiple iterations of the optimisation process. Due to the large variety and amount of data used during training, ChatGPT ranks among the most advanced natural language processing models, capable of processing and generating text at a near-human level [6, 7].

Earlier AI-based medical tools such as IBM Watson for Oncology, Babylon Health, Ada Health, and Buoy Health demonstrated significant limitations – from outdated data and simplified symptom analysis to low effectiveness in interpreting complex clinical cases. Their limited flexibility and lack of contextual understanding highlight the need to evaluate newer models like ChatGPT for their suitability in dynamic healthcare environments. Using ChatGPT for content retrieval is convenient and quick. However, there are risks involved. This model can provide answers to many questions and assist with research, but the information is not always accurate or up to date. This is due to limitations in access to the most recent data and a lack of full understanding of the context. There may also be unconscious bias in the content generated by ChatGPT. Therefore, it is important to verify the obtained information against reliable sources and maintain a critical approach to minimise the risk of misinformation [8, 9].

Objectives

This study aims to critically assess the quality of medical information generated by ChatGPT, with a specific focus on its accuracy, consistency, and clinical usefulness across different medical domains, languages, and question contexts. The review addresses a key gap in current knowledge by evaluating how ChatGPT performs in delivering reliable medical content when exposed to variations in prompt formulation, language (e.g. English vs. non-English), and clinical complexity. Two general research hypotheses were explored: 1) Medical information generated by ChatGPT is in line with the current state of clinical knowledge. 2) The quality of medical information generated by ChatGPT will vary depending on how the prompt is formulated.

Methods

Design

A systematic review was designed to search, assess, and synthesise the available scientific evidence following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [10].

Eligibility criteria

Inclusion criteria for the scientific publications under review were as follows:

1. Publications that directly analyse the quality of medical information generated by ChatGPT.
2. Publications that evaluate the readability, comprehensiveness, accuracy, and usefulness of information provided by ChatGPT.
3. Publications published in English language.
4. Publications published within the period 2022–2024.
5. Publications available via open access.

The inclusion criteria were established to ensure the relevance and quality of the review: limiting the timeframe to 2022–2024 captures only the most up-to-date data related to recent versions of ChatGPT; selecting publications in English and those available via open access ensures linguistic consistency and transparency of analysis; and focusing on studies that evaluate readability, comprehensiveness, accuracy, and usefulness allows for a comprehensive assessment of the actual medical value of the generated content.

Exclusion criteria for the study were as follows:

1. Publications that do not focus directly on the analysis of the quality of medical information generated by ChatGPT.
2. Publications discussing the use of ChatGPT in a non-medical context.
3. Publications in languages other than English.
4. Publications published before 2022.
5. Publications for which no full text is available.

The exclusion criteria were defined to narrow the scope of the review to studies most relevant for evaluating the quality of medical information generated by ChatGPT: publications not directly focused on medical content analysis or those addressing non-medical applications were excluded to maintain thematic coherence; studies published before 2022 and those in languages other than English were excluded to ensure the currency and comparability of data; and the exclusion of publications without full-text access enabled a thorough and reliable assessment of study methodology and findings.

Information sources and search strategies

The systematic review search strategy was based on the following research question: *What is the quality of medical information generated by ChatGPT in terms of accuracy, readability, comprehensiveness, and usefulness for patients and healthcare professionals? What methods can be used to assess the accuracy and reliability of ChatGPT-generated healthcare-related information?*

A systematic review of the literature published between 2022 and 2024 was conducted in three da-

tabases (PubMed, EBSCO, and Scopus), following the PRISMA guidelines. Two reviewers (FT and JGM) searched the databases independently, using the same pre-established protocol. The entire search was performed in June 2024.

The query used in database search engines was as follows:

((“ChatGPT” OR “Generative Pre-trained Transformer” OR “GPT” OR “AI” OR “Artificial Intelligence” OR “Machine Learning” OR “Deep Learning” OR “Neural Networks” OR “Natural Language Processing” OR “Data Mining”) AND (“Health Information” OR “Health Education” OR “Medical Information” OR “Healthcare Information” OR “Health Literacy” OR “Patient Education”) AND (“Implementation” OR “Utilization” OR “Integration” OR “Usage” OR “Adoption” OR “Effectiveness” OR “Benefit”) AND (“Potential” OR “Future” OR “Prospective” OR “Emerging” OR “Innovative” OR “Novel” OR “Advancements” OR “Trends”) AND (“Clinical” OR “Diagnostic” OR “Treatment” OR “Management” OR “Monitoring” OR “Support” OR “Decision-Making” OR “Outcome” OR “Risk” OR “Care” OR “Patient” OR “Patient Education as Topic”))

The reference lists of the previously searched and qualified publications reviewed were also investigated.

Study selection

A three-stage study selection system was used to review the publications. In the first stage, duplicate studies were excluded, followed by screening based on: (1) title, (2) abstract, and (3) full text. Any discrepancies in article selection were resolved by review team consensus (Figure 1).

Analysis was performed by two independent researchers (FT and JGM) using Rayyan software.

Synthesis of results

The synthesis of the results included an analysis of 29 eligible systematic review publications. The results of each study were pooled and compared to identify common trends, differences, and potential sources of heterogeneity. A detailed analysis of the results enabled the formulation of collective conclusions reflecting the current state of knowledge on the topic under review.

Results

Characteristics of the studies

The studies under review included a multidimensional analysis of the use of ChatGPT in various fields of medicine, health education, and support for healthcare professionals and patients. The studies used different ChatGPT versions (3.5, 4.0, and turbo).

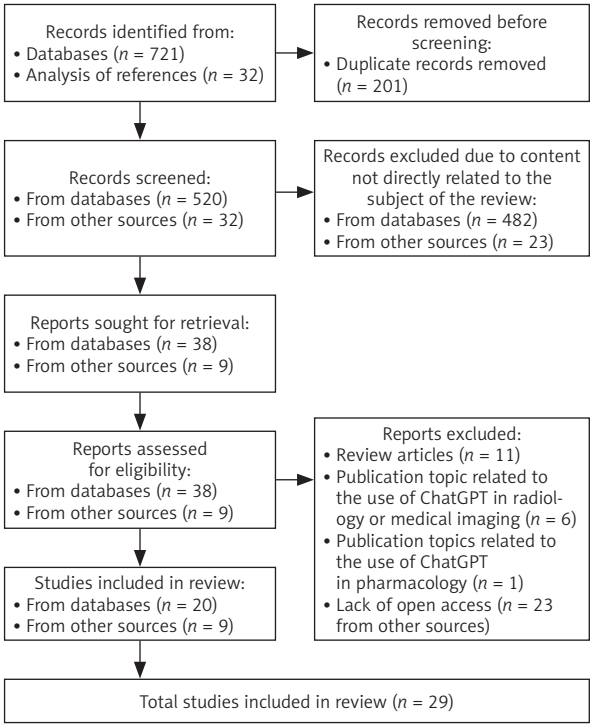

Figure 1. PRISMA flowchart of the systemic search and the selection process

The primary objective of all studies was to assess the quality, accuracy, relevance, and usefulness of ChatGPT responses. A variety of evaluation criteria were considered, such as compliance with medical guidelines, understandability, clarity, completeness, and potential risk of error. The results of the analysis suggest that the effectiveness of ChatGPT varies depending on the specificity of the question and the context. In many cases, the responses were rated as satisfactory; however, significant limitations were also identified, such as lack of references, emotional intelligence issues, and insufficient compliance with guidelines in more complex clinical situations.

The studies conducted were methodologically diverse, including both qualitative and quantitative approaches. Various assessment techniques were used, including guideline compliance analysis, expert review, thematic analysis, and tools such as DISCERN, Global Quality Score (GQS), EQIP, and Flesch-Kincaid. The studies employed a range of evaluation systems, such as scoring systems, Likert scales, and inter-expert agreement analyses. The focus was on different aspects of ChatGPT responses, including their accuracy, completeness, comprehensibility, relevance, and potential risk of error. The results of the studies showed that, in many cases, ChatGPT responses were satisfactory and useful, especially in English language (Table 1) [11–39].

Table 1. Evaluation of the Quality of Medical Information Generated by ChatGPT.

Author *et al.* (year)	Factual Accuracy	Completeness of Responses	Clarity and Understandability	Currency of Information	Ethical Compliance	Clinical Usefulness	Evaluation Method
Ibrahim *et al.* [11]	English: 90% satisfactory, Urdu: 10% satisfactory	English Responses: Satisfactory requiring minimal clarification: 9 out of 10 (90%) Satisfactory requiring moderate clarification: 1 out of 10 (10%) Urdu Responses (both Roman and Nastaliq scripts): Satisfactory requiring moderate clarification: 1 out of 10 (10%) Unsatisfactory requiring substantial clarification: 9 out of 10 (90%)	English: Clear, Urdu: Lacks clarity	English: Current, Urdu: Less rigorous	English: Ethical, Urdu: Mixed vocabulary issues	English: Useful, Urdu: Limited usefulness	Comparison with literature, Expert evaluation with 15+ years of experience, Panel evaluation, No statistical analysis, No surveys or interviews
Sallam *et al.* [12]	Accurate, but some bias and inaccuracies noted	The study does not provide specific numerical data regarding the completeness of responses from ChatGPT Generally complete, but missed some aspects	Clear, but some responses lacked clarity	Current, but limited to pre-2021 data	Ethical, but privacy concerns noted	Useful, but risks noted (critical thinking)	Literature comparison, Expert panel, No stats, No surveys
Hernandez *et al.* [13]	High, with 98.5% accuracy	Very complete, covering most common queries with a success rate of 98.5% appropriate responses out of 70 questions	Clear and understandable	Current as of the date of the study	Ethical, minor concerns noted	High usefulness for patient education	Literature comparison, Expert panel, No stats, No surveys
Alanzi *et al.* [14]	Generally accurate, some minor inaccuracies	The study does not provide specific numerical data regarding the completeness of responses from ChatGPT Mostly complete, but some gaps noted	Clear, but some responses lacked depth	Current, based on recent data	Ethical, minor privacy concerns noted	Useful for public awareness, some limitations	Literature comparison, Expert panel, Stats analysis, No surveys
Aliyeva *et al.* [15]	High accuracy, aligned with guidelines	Mostly complete, covers key aspects Accuracy of information: 100% (5/0) Clarity and understandability: 98% average score Relevance: 92% average score	Clear and understandable	Current, based on recent guidelines	Ethical, minor concerns noted	Useful for postoperative care	Literature comparison, Expert panel, No stats, No surveys

Table 1. Cont.

Author *et al.* (year)	Factual Accuracy	Completeness of Responses	Clarity and Understandability	Currency of Information	Ethical Compliance	Clinical Usefulness	Evaluation Method
Kasapovic *et al.* [16]	Moderate, 47% relevant keywords mentioned	ChatGPT mentioned an average of 47% of relevant keywords across the procedures. The range of relevant keyword mentions varied from 30.5% (lowest) to 68.6% (highest)	Clear but lacks depth	Current but mixed reliability	Ethical, minor privacy concerns	Moderate, 45% somewhat useful	Literature comparison, mDISCERN instrument, Expert panel, Statistical analysis
Anastasio *et al.* [17]	Generally accurate, with some discrepancies	Generally complete, but some gaps noted 4.5% of the responses were rated as bottom tier 27.3% of the responses were rated as middle tier 68.2% of the responses were rated as top tier	Clear, but some responses lacked depth	Current, but mixed reliability	Ethical, minor privacy concerns	Moderately useful for patient education	Literature comparison, DISCERN tool, AIRM tool, Statistical analysis, Expert panel
King *et al.* [18]	High, 95% accuracy for ChatGPT-4	ChatGPT-4: 84% of responses were comprehensive (82 out of 98 responses) ChatGPT-3.5: 79% of responses were comprehensive (66 out of 83 responses)	Clear, but readability at college level	Current as of September 2021	Ethical, with privacy considerations	Useful, but readability needs improvement	Literature comparison, Expert panel, Likert scale for clarity, Statistical analysis
Giorgino *et al.* [19]	High, 4.1/5 for sports medicine	For sports medicine, the average score for responses provided by Google Bard was 4 ±0.78, while the average score for responses generated by ChatGPT was 4.1 ±0.7 For paediatric orthopaedics, the average score for responses provided by Google Bard was 3.5 ±1, while the average score for responses generated by ChatGPT was 3.8 ±0.83	Clear, more readable than Google Bard	Up to date as of September 2021	Ethical, minor concerns noted	Useful, but requires further validation	Literature comparison, GQS tool, Expert panel, Statistical analysis

Table 1. Cont.

Author *et al.* (year)	Factual Accuracy	Completeness of Responses	Clarity and Understandability	Currency of Information	Ethical Compliance	Clinical Usefulness	Evaluation Method
Dergaa *et al.* [20]	Moderate, some inaccuracies noted	Incomplete, lacked depth in complex cases Comprehensive responses: 82% Responses needing minor clarifications: 14% Incomplete responses: 4%	Clear but lacked critical details	Up to date as of 2023	Ethical, but concerns on safety	Limited, potential risks noted	Literature comparison, Expert panel, No stats, Simulated patient interactions
Sahin *et al.* [21]	Generally accurate, 90% correct information	Mostly complete, some gaps noted One response classified as poor (6.7%) Six responses classified as fair (40%) Six responses classified as good (40%) Two responses classified as excellent (13.3%)	Clear, but high readability level	Up to date as of 2023	Ethical, minor privacy concerns	Useful, but high readability level	Literature comparison, DISCERN tool, Readability indices (FRE, FKGL, Gunning FOG, SMOG, CLI, ARI), Statistical analysis, Expert panel
Shahsavar & Choudhury [22]	Generally accurate, some inaccuracies noted	Mostly complete, with gaps in complex queries Appropriate Responses: 98.5% (69 out of 70 questions) Inappropriate Responses: 1.4% (1 out of 70 questions)	Clear, but readability can be improved	Up to date as of 2023	Ethical, with concerns on privacy and data security	Limited, potential for misuse noted	Literature comparison, Expert panel, PLS-SEM analysis, 4-point Likert scale
Xue *et al.* [23]	High, ChatGPT-4: 88%, ChatGPT-3.5: 74%, Bard: 25%	Generally complete, ChatGPT-4: > 80% ChatGPT-4: Responses were complete in over 80% of evaluable cases ChatGPT-3.5: Responses were complete in over 80% of evaluable cases Bard: Responses were complete in approximately 60% of evaluable cases	Clear, ChatGPT-4: 97%, ChatGPT-3.5: 98%, Bard: 83%	Up to date as of 2023	Ethical, with minor concerns	Useful, but needs verification of sources	Literature comparison, Expert panel, Statistical analysis, Adaptation prompts
Lv *et al.* [24]	High, ChatGPT-4: 88%, ChatGPT-3.5: 74%, Bard: 25%	Generally complete, ChatGPT-4: > 80% ChatGPT-4: Complete in over 80% of cases ChatGPT-3.5: Complete in over 80% of cases Bard: Complete in approximately 60% of cases	Clear, ChatGPT-4: 97%, ChatGPT-3.5: 98%, Bard: 83%	Up to date as of 2023	Ethical, with minor concerns	Useful, but needs verification of sources	Literature comparison, Expert panel, Statistical analysis, Adaptation prompts, Readability indices (FRE, FKGL, Gunning FOG, SMOG, CLI, ARI)

Table 1. Cont.

Author *et al.* (year)	Factual Accuracy	Completeness of Responses	Clarity and Understandability	Currency of Information	Ethical Compliance	Clinical Usefulness	Evaluation Method
Fahy *et al.* [25]	Moderate, DISCERN score: 48.74 for ChatGPT-4	Generally complete, ChatGPT-4 better than 3.5 ChatGPT-4: 92% of responses were complete ChatGPT-3.5: 75% of responses were complete Bard: 60% of responses were complete	Clear but high readability level	Up to date as of 2023	Ethical, with minor concerns	Useful but readability limits use	Literature comparison, DISCERN tool, Readability indices (FRE, FKGL, Gunning FOG, SMOG, CLI, ARI), Statistical analysis
Gray *et al.* [26]	High, 72% realism in responses	Generally complete, some gaps in complex queries ChatGPT-4: 82 out of 98 responses (84%) were comprehensive ChatGPT-3.5: 66 out of 83 responses (79%) were comprehensive	Clear, but readability targeted at fifth-grade level	Up to date as of 2023	Ethical, with careful review for realism	Useful for training, with expert review	Literature comparison, Expert panel, 5-point Likert scale, Statistical analysis
Sivarajkumar *et al.* [27]	High, precise information extraction	Generally complete, comprehensive ontology ChatGPT (few-shot): Average F1-score of 0.35 ChatGPT (zero-shot): Average F1-score of 0.37 Precision: 0.33 (zero-shot), 0.27 (few-shot) Recall: 0.8 (zero-shot), 0.82 (few-shot)	Clear, but readability varies	Up to date as of 2023	Ethical, with minor concerns	Highly useful for personalised treatment	Literature comparison, Expert panel, F1-score analysis, Various NLP models (rule-based, machine learning, LLM-based)
Liao *et al.* [28]	Moderate, 84.38% accuracy rate	Generally complete, lacks detail in complex queries Balanced Eating Scenario: Mean completeness score of 7.27 (SD 1.89), Median 8 Fit Weight Scenario: Mean completeness score of 6.70 (SD 1.93), Median 6.5. Dining Out Well Scenario: Mean completeness score of 6.73 (SD 2.29), Median 7 Fewer Processed Foods Scenario: Mean completeness score of 7.90 (SD 1.73), Median 8 Limit Sugary Drinks Scenario: Mean completeness score of 6.80 (SD 1.96), Median 7	Clear, readability excellent but understandability lacking	Up to date as of 2023	Ethical, but concerns on thoroughness	Limited, issues with practical application	Literature comparison, 30 dietitians' feedback, Likert scale, NL test, Descriptive statistical analysis

Table 1. Cont.

Author *et al.* (year)	Factual Accuracy	Completeness of Responses	Clarity and Understandability	Currency of Information	Ethical Compliance	Clinical Usefulness	Evaluation Method
Dennstädt *et al.* [29]	High, 94.3% valid answers, 60.61% correct	Generally complete, but some gaps in detail For multiple-choice questions, ChatGPT provided valid answers for 66 out of 70 questions (94.3%) The correctness of the valid answers: Clinical questions: 50% correct Physics questions: 78.57% correct Biology questions: 58.33% correct For open-ended questions, the evaluations by radiation oncologists were: Very good: 29.3% Good: 28% Acceptable: 19.3% Bad: 9.3% Very bad: 14%	Clear, readability targeted at professionals	Up to date as of September 2021	Ethical, minor concerns noted	Useful, but not reliable for clinical decisions	Literature comparison, 5-point Likert scale, Statistical analysis, Expert panel, Interrater agreement (rWG)
Thia & Saluja [30]	Moderate, validated with Flesch-Kincaid scores	Generally complete, but some gaps noted General information about prostate cancer: 43.0 ±2.9 General information about prostate cancer screening: 34.9 ±2.8 Low risk prostate cancer treatment: 18.6 ±1.8 Advanced prostate cancer treatment: 28.4 ±3.3 General information about renal cell cancer: 26.8 ±4.0 General information about bladder cancer: 36.2 ±2.9 Bladder cancer diagnosis and staging: 35.6 ±0.7 Non-muscle invasive bladder cancer treatment: 25.9 ±1.1 Muscle invasive bladder cancer treatment: 31.2 ±3.6	Clear, readability improved for younger audience	Up to date as of 2023	Ethical, minor concerns noted	Useful, but verification needed	Literature comparison, Flesch-Kincaid score, Expert panel, Statistical analysis

Table 1. Cont.

Author *et al.* (year)	Factual Accuracy	Completeness of Responses	Clarity and Understandability	Currency of Information	Ethical Compliance	Clinical Usefulness	Evaluation Method
Walker *et al.* [31]	Moderate, 60% agreement with guidelines	Generally complete, some gaps noted Gallstone Disease: Total Score: 16 (IQR 14.5-18) Content: 10 (IQR 9.5-12.5) Identification: 1 (IQR 1-1) Structure: 4 (IQR 4-5) Pancreatitis: Total Score: 19 (highest among conditions) Content: Varied between 9 and 14 Liver Cirrhosis: Total Score: 15 Pancreatic Cancer: Total Score: 17 Hepatocellular Carcinoma: Total Score: 14 (lowest among conditions)	Clear, but readability varies	Up to date as of 2023	Ethical, minor privacy concerns	Limited, useful for basic information	Literature comparison, EQIP tool, Expert panel, Statistical analysis
Mondal *et al.* 32]	Moderate, 1.83±0.37 accuracy score	Among 20 cases, the average score of accuracy was 1.83 ±0.37 and guidance was 1.9 ±0.21. Both scores were higher than the hypothetical median of 1.5, indicating a high level of completeness and accuracy in the responses	Clear, but college-level readability	Up to date as of 2023	Ethical, minor concerns noted	Useful for initial guidance	Literature comparison, 3-point Likert scale, FKES, Expert panel, Statistical analysis

Table 1. Cont.

Author *et al.* (year)	Factual Accuracy	Completeness of Responses	Clarity and Understandability	Currency of Information	Ethical Compliance	Clinical Usefulness	Evaluation Method
Yeo *et al.* [33]	Moderate, 79.1% correct for cirrhosis, 74% for HCC	Generally complete, but some gaps noted In gastroenterology questions (n = 15), the completeness of responses was as follows: Comprehensive: ChatGPT-4 (60%), ChatGPT-3.5 (53%) Correct but inadequate: ChatGPT-4 (20%), ChatGPT-3.5 (27%) Some correct and some incorrect: ChatGPT-4 (20%), ChatGPT-3.5 (13%) Completely incorrect: ChatGPT-4 (0%), ChatGPT-3.5 (7%) In neurology questions (n = 15), the completeness of responses was as follows: Comprehensive: ChatGPT-4 (67%) Correct but inadequate: ChatGPT-4 (20%) Some correct and some incorrect: ChatGPT-4 (13%) Completely incorrect: ChatGPT-4 (0%) Comprehensive responses for ChatGPT-4: (9/15 gastroenterology + 10/15 neurology) = 19/30 = 63.33% Comprehensive responses for ChatGPT-3.5: (8/15 gastroenterology + not available for neurology)	Clear, but varied by complexity	Up to date as of 2021	Ethical, minor concerns noted	Useful for basic information, not for decision-making	Literature comparison, Expert panel, Two independent reviewers, Statistical analysis

Table 1. Cont.

Author *et al.* (year)	Factual Accuracy	Completeness of Responses	Clarity and Understandability	Currency of Information	Ethical Compliance	Clinical Usefulness	Evaluation Method
Alan & Alan [34]	High, DISCERN scores indicate 'good' quality	Generally complete, ‚treatment choices' section scored lower total DISCERN Scores for ChatGPT-4 responses: Rater-1: Mean score of 55.60 ±2.27 Rater-2: Mean score of 54.00 ±4.14 Treatment Choices Section Scores (lower than other sections): Significantly lower scores in the treatment choices section for both raters	Clear, readability suitable for general population	Up to date as of 2023	Ethical, minor concerns noted	Useful for patient education	Literature comparison, DISCERN tool, Two independent raters, Statistical analysis (weighted kappa, Krippendorff alpha)
Haver *et al.* [35]	High, 88% appropriate responses	Generally complete, but some gaps noted ChatGPT responses were appropriate for 22 out of 25 questions (88%) in both patient-facing and chatbot contexts One response (4%) was deemed inappropriate. Two responses (8%) were deemed unreliable	Clear, suitable for general population	Up to date as of 2023	Ethical, minor concerns noted	Useful for patient education, but supervision needed	Literature comparison, Expert panel, Three independent reviewers, Statistical analysis (descriptive statistics)
Rahsepar *et al.* [36]	Moderate, 70.8% correct answers	Generally complete, some gaps noted Completeness (mean): 7.45 Completeness (median): 7.5	Clear, varied by question complexity	Up to date as of 2021 for ChatGPT, 2023 for Bard	Ethical, minor concerns noted	Useful for basic information, not for decision-making	Literature comparison, Expert panel, Statistical analysis, Consistency check

Table 1. Cont.

Author *et al.* (year)	Factual Accuracy	Completeness of Responses	Clarity and Understandability	Currency of Information	Ethical Compliance	Clinical Usefulness	Evaluation Method
Seth *et al.* [37]	Moderate, 60–70% accuracy	Generally complete, some gaps noted Correcting internal valve dysfunction: DISCERN score: 42 Correcting external valve collapse: DISCERN score: 42 Correcting caudal septal dislocation: DISCERN score: 44 Managing turbinate hypertrophy: DISCERN score: 41 Managing tip support after submucous resection: DISCERN score: 43 When to do nasal bone fractures in rhinoplasty: DISCERN score: 41	Clear, readability varied by model	Up to date as of 2021 for ChatGPT	Ethical, minor concerns noted	Useful for general information, not for detailed guidance	Literature comparison, Expert panel, Flesch Reading Ease Score, Flesch-Kincaid Grade Level, Coleman–Liau Index, Modified DISCERN score, Likert scale
Mondal *et al.* [38]	High, relational level of accuracy	Generally complete, some gaps noted The average text reading ease score was 46.94 ±8.23 The text similarity index was 27.07 ±11.46%	Clear, readability suitable for high school to college students	Up to date as of 2023	Ethical, minor concerns noted	Useful for patient education	Literature comparison, SOLO taxonomy, Flesch-Kincaid readability scores, Turnitin for text similarity, Statistical analysis
Ponzo *et al.* [39]	Moderate, 55.5%–73.3% appropriateness	Generally complete, but some gaps noted Sarcopenia: 55.5% NAFLD: 73.3% Obesity: 70% T2DM: 71.4% CKD: 70% Hypertension: 66.7% Hypercholesterolaemia: 71.4% Hypertriglyceridaemia: 71.4%	Clear, suitable for general population	Up to date as of 2023	Ethical, minor concerns noted	Useful for general dietary advice	Literature comparison, Panel of nutrition experts, Multiple prompt sessions, Statistical analysis

.A variety of methodological approaches were used to construct prompts in 29 articles analysing the quality of medical information generated by ChatGPT. Most studies used open-ended questions that allowed for more elaborate and detailed responses. These questions were typically simple and straightforward, consisting of 1–2 sentences with simple grammar. The prompts were formulated in formal and specialised medical language, ensuring precision and clarity of expected responses. Most studies did not consider context, such as patient history or medical details, limiting the potential for more personalised responses. The questions were generally clear and precise, aiming to minimise response ambiguity.

A significant proportion of the studies focused on specific medical fields, such as orthopaedics, diabetology, or oncology, enabling the performance of ChatGPT to be evaluated in the context of advanced terminology and detailed medical knowledge. Most prompts addressed easy or moderately difficult topics, facilitating the assessment of response accuracy and completeness. Requirements for providing references in responses were rarely specified, which could have affected the assessment of the reliability of information. Follow-up questions were infrequent, and the interactivity of the prompts was usually limited to single questions with no dialogue sequences. This approach allowed for an effective assessment of the quality and usefulness of the information generated by ChatGPT in various medical contexts, although it may have limited the depth and personalisation of the responses obtained. The results are summarised in Table 2.

Hypothesis 1: Medical information generated by ChatGPT is consistent with up-to-date clinical knowledge.

Research on patient use of ChatGPT

Research on ChatGPT responses shows varying levels of agreement with current clinical knowledge. A study by Ibrahim *et al.* [11], explored responses related to total hip arthroplasty. The results showed that 90% of the responses in English were satisfactory, requiring minimal or moderate clarification, whereas 90% of the responses in Urdu required significant clarification.

Hernandez *et al.* [13] examined ChatGPT responses regarding type 2 diabetes and found 98.5% of them to be consistent with current medical standards. A study by Alanzi *et al.* [14] analysed public awareness of obesity as a cancer risk factor and found that 65% of the study participants were not fully aware of this connection. However, ChatGPT demonstrated potential as an educational tool.

Aliyeva *et al.* [15] evaluated the effectiveness of ChatGPT-4 as an additional source of information for patients after cochlear implant surgery, showing a high level of compliance with medical guidelines and a high level of response accuracy.

Research on the use of ChatGPT by healthcare professionals

A study by Sallam *et al.* [12], evaluated the responses of ChatGPT regarding medical education, with an average score of 3.96 ±0.19 for accuracy, 2.96 ±0.19 for clarity, and 2.56 ±0.50 for conciseness on a 4-point scale.

Giorgino *et al.* [19] compared the responses of ChatGPT-3.5 and Google Bard in the fields of paediatric and sports orthopaedics, and found that ChatGPT achieved an average score of 4.1 ±0.7 on the quality scale, outperforming Google Bard.

Anastasio *et al.* [17] examined ChatGPT responses to foot and ankle surgery questions, finding 68.2% of them to be of the highest quality.

Fahy *et al.* [25] assessed the quality of information about platelet-rich plasma therapy for osteoarthritis and demonstrated that the average DISCERN score for ChatGPT-4 was 48.74, indicating moderate information quality.

Haver *et al.* [35] examined ChatGPT responses to questions about breast cancer prevention and found that 88% of the responses were appropriate, with only 4% containing inappropriate information.

To sum up, studies suggest that ChatGPT can provide information consistent with up-to-date clinical knowledge, although the quality of responses largely depends on the language and the context. Patient-oriented responses are often useful and in line with guidelines, but linguistic differences can affect their accuracy. Research has shown that responses in English tend to be more accurate than those in other languages, such as Urdu. Also, ChatGPT shows potential as an educational tool, helping patients to better understand their health condition.

In the context of healthcare professionals, ChatGPT achieves a high degree of alignment with current medical standards. Study results show that ChatGPT responses are often correct and in line with clinical guidelines, which can be particularly useful for medical education and clinical support. However, further refinement of the algorithms is needed to ensure more comprehensive and accurate responses to more complex queries. For instance, while responses on type 2 diabetes and platelet-rich plasma therapy were rated highly, other studies suggest a need to improve the accuracy and consistency of the information, especially for more complex health issues.

ChatGPT has the potential to support physicians and other healthcare professionals – such as nurses, pharmacists, and physiotherapists – in their daily clinical practice. For example, it can automate the generation of visit summaries, discharge reports, and care plans, streamlining medical documentation and saving time. It may also enhance communication with patients by translating complex medical

Table 2. Checklist of prompt characteristics in articles assessing the quality of medical information generated by ChatGPT

Prompt Aspect	Description	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
Question type	Open	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Closed																			▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Multiple choice																			▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
Complexity of the question	Simple (unambiguous)	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Complex (multi-part, requiring analysis)	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
Prompt length	Short (1-2 sentences)	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Medium (3-5 sentences)																													
	Long (more than 5 sentences)																													
Grammatical structure	Simple	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Complex																													
Context presence	With context (patient history, medical details)										▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Without context	▪	▪	▪	▪	▪	▪	▪	▪	▪																				
Linguistic form	Formal	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Informal																													
	Specialist	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
Prompt precision	Clear and precise	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Vague and general																													
Thematic area	Specific specialisation	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	General medicine																													
Terminological specificity	Using specialised medical language	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Without specialised language																													
Difficulty level of the issue	Easy	▪		▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Medium difficulty		▪																	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Difficult																													
Expected response time	Short (immediate)	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Long (after a few minutes)																													
Use of examples in prompts	With clinical examples		▪																											
	Without examples	▪		▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
Presence of additional questions	Yes	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪		▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	No	▪		▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
Prompt interactivity	Single prompt	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
	Sequence of questions (dialog)																			▪	▪		▪	▪	▪	▪	▪	▪	▪	▪
Expectation of sources	Request for sources													▪																
	No request for sources	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪		▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪

**Articles are marked with numbers 12, 13, 14 according to the reference list. ▪ This aspect was taken into account and fulfilled in the analysed article.*

terminology into plain language, thereby improving patients' understanding of their diagnosis and treatment recommendations. Future research should explore ChatGPT's potential in personalising treatment plans, analysing data from electronic health records (EHR), supporting clinical decision-making, and serving as an educational tool for various professional groups within the healthcare system.

Hypothesis 2: The quality of medical information generated by ChatGPT varies depending on how the prompt is formulated.

Research on the quality of medical information generated by ChatGPT suggests that the way prompts are formulated has a significant impact on the quality of responses. A study by Alanzi *et al.* [14], analysed public awareness of obesity as a risk factor for cancer and evaluated ChatGPT as an educational tool. The results showed that 65% of participants were not fully aware of the link between obesity and cancer, but precisely formulated questions to ChatGPT helped to improve the level of awareness. The study found that in 75% of cases, ChatGPT responses were rated as appropriate when the questions were precisely worded.

Another study, conducted by Aliyeva *et al.* [15], assessed the effectiveness of ChatGPT-4 in providing information to patients after cochlear implant surgery. The study showed that 90% of the responses were consistent with medical guidelines when the questions were precisely worded. This highlights the importance of detailed prompts in obtaining reliable information.

A study by Anastasiao *et al.* [17] evaluated the quality of ChatGPT responses to foot and ankle surgery questions. The results showed that 68.2% of the responses were rated as being of the highest quality, indicating the significance of clear and detailed prompts in obtaining high-quality medical information. The study also showed that only 4.5% of the responses were considered to be of poor quality when the questions were ambiguous.

Fahy *et al.* [25] investigated the quality of information provided by ChatGPT-4 on platelet-rich plasma therapy for osteoarthritis treatment and obtained an average DISCERN score of 48.74. The study found that 85% of the responses were accurate when the questions were precisely worded; suggesting that the way prompts are worded has a direct impact on the quality of the responses generated.

A study by Dergaa *et al.* [20] analysed ChatGPT responses to mental health questions. The results showed that the responses were more accurate and safer when the questions were precisely worded, whereas more general prompts led to less accurate responses. The responses were rated as accurate in 78% of the cases when the questions were detailed, compared to 45% accuracy with more general prompts.

In summary, the studies clearly indicate that the quality of medical information generated by ChatGPT largely depends on how the prompts are formulated. Precise, clear, and well-designed prompts result in responses that are more accurate, clear, and consistent with up-to-date clinical knowledge.

Discussion

The reviewed publications reveal diverse results regarding the quality of responses generated by language models such as ChatGPT. The evaluation of the relevance and accuracy of responses in various fields of medicine is an important aspect of these studies. The study by Ibrahim *et al.* [11] showed a significant difference in the quality of responses between English and Urdu, with responses in English generally being more accurate and requiring less additional detailed explanation. These results suggest that language and cultural barriers may affect the quality of information generated by artificial intelligence, despite the clear potential of language models.

A study by Sallam *et al.* [12] highlighted the benefits and limitations of ChatGPT in medical education. Although the responses provided by the model were accurate, they often lacked completeness and precision, which proved challenging in the context of medical education. The assessment system highlighted the need for further improvement of the language model to meet high medical standards by analysing the accuracy, clarity, and conciseness of responses. On the other hand, a study by Alanzi *et al.* [14] on public awareness of the link between obesity and cancer risk showed that ChatGPT can be a useful educational tool, although the responses required additional verification for accuracy and being up to date.

Another important aspect is the use of ChatGPT to support patient education. A study by Hernandez *et al.* [13] on responses regarding type 2 diabetes showed that ChatGPT can be an effective educational tool, with a high accuracy of information. However, the key challenge is to ensure the consistency and stay up to date with medical knowledge. Mechanisms for validating and regularly updating AI-generated content should be implemented to minimise the risk of misinformation among patients and healthcare professionals.

A study by Kasapovic *et al.* [16] revealed significant limitations in the quality of information generated by ChatGPT in the context of orthopaedics and trauma surgery. The average information quality rating using the modified DISCERN tool was 2.4 out of 5, suggesting moderate to low information quality. This suggests there is a need for further development of the language models to provide the users with more accurate and useful information.

Conversely, research by Shahsavar *et al.* [22] on perceived user decisions and intentions to use ChatGPT

for self-diagnosis found that most respondents were inclined to use ChatGPT for self-diagnosis, emphasising its potential as a support tool. High performance expectations and positive risk-benefit ratings were positively correlated with intentions to use ChatGPT for self-diagnosis. These findings highlight the need for further research on the integration of artificial intelligence into user decision-making processes and its potential to improve the quality of healthcare.

The highest quality of ChatGPT-generated responses was observed in areas related to basic health education. For example, in a study on type 2 diabetes, 98.5% of the responses were consistent with current medical guidelines [13], while patient education following cochlear implant surgery yielded 100% guideline compliance and an average understandability rating of 98% [15]. Similarly, in the context of breast cancer prevention, 88% of the responses were deemed appropriate [35].

In contrast, the quality of responses significantly declined in more complex clinical domains. In orthopaedics and trauma surgery, the average inclusion rate of relevant medical terminology was only 47%, and the overall completeness of information was rated as moderate [16]. In the field of foot and ankle surgery, only 68.2% of responses were classified as high quality [17], while in radiation oncology, the accuracy of responses to clinical questions reached just 50% [29].

Studies included in this review demonstrated that open-ended questions (e.g. "What are the recommendations for postoperative care after cochlear implant surgery?") resulted in responses with higher accuracy and greater alignment with clinical guidelines – for instance, 90–100% of ChatGPT's answers were consistent with clinical knowledge in the analysis by Aliyeva *et al.* [15]. In contrast, for multiple-choice questions, as shown in the study by Dennstädt *et al.* on radiation oncology [29], only 60.6% of answers were correct, and for clinical questions specifically, only 50% were deemed accurate, highlighting the limited effectiveness of this question format when interacting with the model.

The analysis demonstrated that the quality of ChatGPT-generated responses significantly improves when well-structured, precise prompts are used. For example, a poorly formulated question such as "How to care for a patient after surgery?" resulted in a general and clinically unhelpful response. In contrast, the prompt "What are the nursing care recommendations for adult patients following cochlear implant surgery, according to current medical guidelines?" yielded a detailed, guideline-consistent, and clinically useful answer. Similarly, the question "What to do for diabetes?" produced vague responses, whereas a precise prompt such as "What are the current recommendations for non-pharmacological management of type 2 diabetes in adult patients, according to the 2023 ADA guidelines?" led to an accurate and medically relevant reply. These examples clearly illustrate that well-crafted prompts – including specific clinical context, patient population, and references to guidelines – greatly enhance the accuracy, completeness, and utility of the information generated by ChatGPT.

AI-generated misinformation – such as incorrect drug dosages, contraindications, or misinterpretation of diagnostic results – can pose a direct threat to patient safety. In one of the reviewed studies, ChatGPT provided recommendations for platelet-rich plasma therapy in osteoarthritis that were inconsistent with current clinical guidelines, potentially leading to inappropriate treatment decisions [25]. To effectively mitigate the risk of misinformation, AI models must be integrated with reliable, real-time updated medical knowledge bases, such as PubMed. These models should automatically generate citations to specific sources and indicate the confidence level of their responses, for example, through accuracy scores or guideline compliance indicators. It is also essential to implement expert validation procedures for AI-generated content before clinical application. In parallel, targeted training should be provided to healthcare professionals and patients on how to interpret AI-generated information and use it responsibly and safely.

The phenomenon of "hallucination" – the generation of factually incorrect information by ChatGPT – has been noted in several studies included in this review, particularly in the context of medical responses. For example, in the study by Dennstädt *et al.* [29], only 50% of clinical responses were deemed accurate, while the remainder contained potentially misleading or incomplete information that could lead to inappropriate therapeutic decisions. Similarly, Sallam *et al.* [12] highlighted the presence of bias and factual errors in ChatGPT's outputs, emphasising the need for validation before clinical use. Despite these individual observations and cautionary findings, the current literature lacks systematic studies that thoroughly examine the hallucination phenomenon in ChatGPT-generated medical content, including assessments of clinical risk across different specialties.

In summary, the studies reviewed indicate the significant potential of GPT language models in medicine, but also the need for their further development and improvement. Future research should focus on the comparison of different language models and their application in specific clinical contexts. The introduction of systematic methods for evaluating the quality of information and the integration of feedback from medical professionals can contribute to increasing the effectiveness and safety of AI applications in medicine. It will also be crucial to monitor and update algorithms to ensure that the information provided is always up-to-date, accurate, and compliant with the latest medical guidelines.

Ethical, legal, and social implications

The use of ChatGPT in healthcare raises a range of specific ethical, legal, and social challenges that require urgent attention and regulation. Most notably, the model does not consistently provide sources or indicate the reliability of its responses, making it difficult for physicians and patients to assess the credibility of the information – an issue that could lead to risky clinical decisions. The lack of clear legal accountability for potentially harmful or incorrect content generated by the model raises critical questions about who bears responsibility in the event of patient harm – the physician, the healthcare institution, or the AI developer. There is also a risk of exacerbating health inequalities, as access to safe and effective AI tools may be limited by linguistic, digital, or economic barriers. Furthermore, in terms of privacy, the use of ChatGPT poses potential risks related to the handling of sensitive medical data without clearly defined regulatory frameworks, which could lead to violations of patient rights and professional ethical standards [40, 41].

Limitations

The developed systematic review has certain limitations. Firstly, the variety of ChatGPT model versions, such as 3.5 and 4.0, used in the studies leads to inconsistent results, making it difficult to draw general conclusions. Also, the reviewed studies have various objectives, including the evaluation of the accuracy, readability, and educational usefulness of ChatGPT-generated responses. This heterogeneity in aims and evaluation methods poses a significant challenge for synthesising and comparing results. Evaluation by experts from different medical specialties, such as orthopaedics, otolaryngology, or internal medicine, may introduce bias due to varying guidelines, affecting consistency and rigour. Moreover, differences in qualitative and quantitative approaches, as well as the lack of standardised rating systems like DISCERN or EQIP, make it difficult to synthesise data and draw definitive conclusions.

Another major limitation of the study is the variability in the characteristics of queries and command constructions, as shown in Table 2. The variety of question types, their complexity, and grammatical structure affect the quality and reliability of the responses generated. Furthermore, the studies include different demographic and geographical groups, introducing potential demographic bias and limiting the generalisability of the results. Inconsistent inclusion of sources and references in ChatGPT responses is another important issue that undermines the credibility of the medical information generated.

Most of the included studies were cross-sectional and did not evaluate the real-world impact of ChatGPT-generated responses on clinical decision-making or patient behaviour, which limits the assessment of its practical effectiveness and safety.

There is also a lack of long-term comparative studies examining ChatGPT against other medical decision support tools, such as clinical decision support systems (CDS), making it difficult to determine its actual added value in healthcare settings.

Additionally, the influence of query language (e.g. English vs. non-English) and local cultural context on response quality has not been systematically analysed, despite its potential significance for the model's applicability in diverse clinical environments.

To improve future research on medical information generated by ChatGPT, it is essential to use a consistent version of the model in studies and to introduce standard evaluation criteria, such as DISCERN. It is important for evaluation to be conducted by several specialists in relevant medical fields. It is also necessary to standardise query characteristics and commands and to ensure that responses include appropriate references. Furthermore, it is crucial to ensure the readability and comprehensibility of responses, tailoring them to the educational level of the recipients.

Conclusions

Research on the assessment of the quality of ChatGPT-generated medical information shows significant variation in the effectiveness and usefulness of responses depending on the language, query complexity, and clinical context. Furthermore, responses to simple and straightforward questions tend to be more accurate and complete than answers to complex questions that require deeper analysis. The clinical context also plays a crucial role, as responses that include additional clinical information are more valuable. Nevertheless, ChatGPT's inability to interact, its lack of emotional intelligence, and occasional inaccuracies in information highlight the need to further optimise the language models. Despite these limitations, the potential of ChatGPT as an educational and support tool in medicine is promising, especially in the context of providing rapid access to information and education. Based on this information, recommendations for ChatGPT users have been developed, as shown in Table 3.

The recommendations presented in the table are practical in nature and are directed at both end users (e.g. healthcare professionals) and developers of AI-based solutions. Their primary aim is to enhance the precision and usefulness of responses generated by ChatGPT by optimising prompt formulation – taking into account factors such as length, grammatical structure, inclusion of clinical context, level of terminological specificity, and the expectation of source citations. The key conclusion drawn from these recommendations is that the quality of generated content

Table 3. Recommendations for ChatGPT users

Question aspect	Description	Recommendation	Example
Prompt type	The type of question can affect the answer (open, closed, multiple choice).	Prefer open questions to get more detailed answers.	Open-ended prompt: *What are the key aspects of nursing care for patients with erectile dysfunction?* Closed-ended prompt: *Do patients with erectile dysfunction require specialised nursing care? (Yes/No)*
Prompt complexity	The simplicity or complexity of the question can affect the quality of the answer.	Use clear and unambiguous questions, avoid complex multi-part questions requiring analysis.	Simple prompt: *What are the basic care needs of patients with erectile dysfunction?* Complex prompt: *What are the differences in nursing care for patients with erectile dysfunction depending on the cause (psychogenic vs. organic)?*
Prompt length	Short questions are usually easier to understand and answer.	Prefer short questions (1–2 sentences) for better comprehension.	Short prompt: *What are the methods of supporting patients with erectile dysfunction?* Long prompt: *What are the most effective methods of supporting patients with erectile dysfunction that nurses can apply in clinical practice?*
Grammatical structure	Simple grammatical structures are more understandable for language models.	Use simple grammatical structures in questions.	Simple Structure: *How can a nurse assist a patient with erectile dysfunction?* Complex Structure: *What nursing interventions are most effective in improving the quality of life for patients with erectile dysfunction?*
Context presence	Providing context (e.g., patient history, medical details) can improve the quality of the answer.	Provide context in questions if it is relevant for obtaining a precise answer.	Without context: *How can nurses support patients with erectile dysfunction?* With context: *How can nurses support patients with erectile dysfunction after prostate surgery?*
Linguistic form	Using formal or specialist language can affect the quality of the answer.	Use formal language and specialist terms if the question concerns specific topics.	Formal Language: *What nursing interventions are recommended for patients with erectile dysfunction?* Informal Language: *How can a nurse help men with erection problems?*
Question precision	Clear and precise questions help to get more accurate answers.	Formulate questions clearly and precisely.	Precise question: *What are the recommendations for the care of patients with erectile dysfunction after pharmacological treatment?* Imprecise question: *How to care for patients with erectile dysfunction?*
Thematic area	The specificity of the topic can influence the level of detail in the answer.	Ask questions related to specific thematic areas for more detailed answers.	Specific Area: *What are the most common causes of erectile dysfunction in patients over the age of 50 and how can nurses assist them?* General Area: *What are the causes of erectile dysfunction?*

Table 3. Conyt.

Question aspect	Description	Recommendation	Example
Terminological specificity	Using specialised medical language can increase the accuracy of the answer.	Use specialised language when the question concerns a specific field.	Specialist language: *What are the nursing interventions for erectile dysfunction associated with diabetic neuropathy?* Simple language: *How to help patients with erection problems?*
Difficulty level of the issue	The difficulty level of the question can affect the quality of the answer.	Pose questions of moderate difficulty for optimal answer quality.	Moderate level: *What are the key steps in caring for a patient with erectile dysfunction?* Advanced level: *What are the advanced procedural techniques for treating erectile dysfunction?*
Expected response time	The expected response time can be short (immediate) or long (after a few minutes).	Expect a short response time by formulating clear and unambiguous questions.	Short response time: *What are the key nursing interventions for patients with erectile dysfunction?* Long response time: *What are the differences in nursing interventions for patients with erectile dysfunction depending on the cause of these dysfunctions, and what are the latest research findings in this field?*
Use of examples in prompts	Providing examples in questions can help to get better answers.	Add clinical examples in questions if appropriate.	With example: *How can a nurse support a patient with erectile dysfunction after prostate surgery?* Without example: *How can a nurse support patients with erectile dysfunction?*
Presence of additional prompt	The presence of additional questions can affect the quality of the information obtained.	Avoid asking multiple additional questions in one prompt.	Single question: *How can a nurse support patients with erectile dysfunction?* Multiple questions: *How can a nurse support patients with erectile dysfunction? What treatment options are available? What are the side effects?*
Prompt interactivity	Questions can be single or a sequence of questions (dialogue).	Prefer single questions to obtain more consistent answers.	Single question: *What are the key nursing interventions for patients with erectile dysfunction?* Sequence of questions: *What are the key nursing interventions for patients with erectile dysfunction? What are the most common causes?*
Expectation of sources	Requesting sources in responses can increase their reliability.	Clearly indicate if you expect sources in the answer.	With expectation of sources: *What are the key nursing interventions for patients with erectile dysfunction? Please provide sources.* Without expectation of sources: *What are the key nursing interventions for patients with erectile dysfunction?*

is strongly dependent on the quality of the prompt, highlighting the need for user education and the standardisation of interactions with language models.

Future research should focus on evaluating the effectiveness of these recommended strategies in clinical practice, examining the impact of linguistic and cultural biases in responses, and developing models integrated with real-time updated medical knowledge bases to improve their reliability and safety in healthcare applications.

Funding

No external funding.

Ethical approval

Not applicable.

Conflict of interest

The authors declare no conflict of interest.

References

1. Sarkar C, Das B, Rawat VS, Wahlang JB, Nongpiur A, Tiewsoh I, Lyngdoh NM, Das D, Bidarolli M, Sony HT. Artificial intelligence and machine learning technology driven modern drug discovery and development. Int J Mol Sci. 2023; 24(3): 2026.
2. Ossowska A, Kusiak A, Świetlik D. Artificial intelligence in dentistry-narrative review. Int J Environ Res Public Health. 2022; 19(6): 3449.
3. Nowy Marketing. Raport Digital 2024 Global Overview: Jaki był 2023 rok w Internecie i mediach społecznościowych [Internet]. 2024 [cytowane 2024 Lip 23]. Available at: https://nowymarketing.pl/raport-digital-2024-global-overview-jaki-byl-2023-rok-w-internecie-i-mediach-spolecznosciowych/
4. Kokka I, Mourikis I, Nicolaides NC, Darviri C, Chrousos GP, Kanaka-Gantenbein C, Bacopoulou F. Exploring the effects of problematic internet use on adolescent sleep: a systematic review. Int J Environ Res Public Health. 2021; 18(2): 760.
5. Biswas S, Dobaria D, Cohen HL. ChatGPT and the future of journal reviews: a feasibility study. Yale J Biol Med. 2023; 96(3): 415-420.
6. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023; 11(6): 887.
7. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023; 25: e48568.
8. Dergaa I, Chamari K, Zmijewski P, Saad HB. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport. 2023; 40(2): 615-622.
9. Gödde D, Nöhl S, Wolf C, Rupert Y, Rimkus L, Ehlers J, Breuckmann F, Sellmann T. A SWOT (strengths, weaknesses, opportunities, and threats) analysis of ChatGPT in the medical literature: concise review. J Med Internet Res. 2023; 25: e49368.
10. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009; 6(7): e1000097.
11. Ibrahim MT, Khaskheli SA, Shahzad H, Noordin S. Language-adaptive artificial intelligence: assessing ChatGPT's answer to frequently asked questions on total hip arthroplasty questions. J Pak Med Assoc. 2024; (Suppl. 4): 74.
12. Sallam M, Salim NA, Barakat M, Al-Tammemi AB. ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. Narra J. 2023; 3(1): e103.
13. Hernandez CA, Vazquez Gonzalez AE, Polianovskaia A, Sanchez RA, Arce VM, Mustafa A, Vypritskaya E, Gutierrez OP, Bashir M, Sedeh AE. The future of patient education: AI-driven guide for type 2 diabetes. Cureus. 2023; 15(11): e48919.
14. Alanzi TM, Alzahrani W, Albalawi NS, Allahyani T, Alghamdi A, Al-Zahrani H, Almutairi A, Alzahrani H, Almulhem L, Alanzi N, Al Moarfeg A, Farhah N. Public awareness of obesity as a risk factor for cancer in Central Saudi Arabia: feasibility of ChatGPT as an educational intervention. Cureus. 2023; 15(12): e50781.
15. Aliyeva A, Sari E, Alaskarov E, Nasirov R. Enhancing postoperative cochlear implant care with ChatGPT-4: a study on artificial intelligence (AI)-assisted patient education and support. Cureus. 2024; 16(2): e53897.
16. Kasapovic A, Ali T, Babasiz M, Bojko J, Gathen M, Kaczmarczyk R, Roos J. Does the information quality of ChatGPT meet the requirements of orthopedics and trauma surgery? Cureus. 2024; 16(5): e60318.
17. Anastasio AT, Mills IV FB, Karavan Jr MP, Adams Jr SB. Evaluating the quality and usability of artificial intelligence–generated responses to common patient questions in foot and ankle surgery. Foot Ankle Orthop. 2023; 8(4): 24730114231209919.
18. King RC, Samaan JS, Yeo YH, Peng Y, Kunkel DC, Habib AA, Ghashghaei R. A multidisciplinary assessment of ChatGPT's knowledge of amyloidosis: observational study. JMIR Cardio. 2024; 8: e53421.
19. Giorgino R, Alessandri-Bonetti M, Del Re M. Evaluating ChatGPT and Google Bard as educational tools for patients in orthopedics. Diagnostics. 2024; 14: 1253.
20. Dergaa I, Fekih-Romdhane F, Hallit S, Loch AA, Glenn JM, Fessi MS, Aissa MB, Souissi N, Guelmami N, Swed S, El Omri A, Bragazzi NL, Saad HB. ChatGPT is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry. 2024; 14: 1277756.
21. Sahin S, Erkmen B, Duymaz YK, Bayram F, Tekin AM, Topsakal V. Evaluating ChatGPT-4's performance as a digital health advisor for otosclerosis surgery. Front Surg. 2024; 11: 1373843.
22. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors. 2023; 10: e47564.
23. Xue E, Bracken-Clarke D, Iannantuono GM, Choo-Wosoba H, Gulley JL, Floudas CS. Utility of large language models for health care professionals and patients in navigating hematopoietic stem cell transplantation: comparison of the performance of ChatGPT-3.5, ChatGPT-4, and Bard. J Med Internet Res. 2024; 26: e54758.
24. Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J. Leveraging large language models for improved patient access and self-management: assessor-blinded comparison between

expert- and AI-generated content. J Med Internet Res. 2024; 26: e55847.
25. Fahy S, Niemann M, Böhm P, Winkler T, Oehme S. Assessment of the quality and readability of information provided by ChatGPT in relation to the use of platelet-rich plasma therapy for osteoarthritis. J Pers Med. 2024; 14: 495.
26. Gray M, Baird A, Sawyer T, James J, DeBroux T, Bartlett M, Krick J, Umoren R. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of ChatGPT: exploratory observational study. JMIR Med Educ. 2024; 10: e50705.
27. Sivarajkumar S, Gao F, Denny P, Aldhahwani B, Visweswaran S, Bove A, Wang Y. Mining clinical notes for physical rehabilitation exercise information: natural language processing algorithm development and validation study. JMIR Med Inform. 2024; 12: e52289.
28. Liao LL, Chang LC, Lai IJ. Assessing the quality of ChatGPT's dietary advice for college students from dietitians' perspectives. Nutrients. 2024; 16: 1939.
29. Dennstädt F, Hastings J, Putora PM, Vu E, Fischer GF, Süveg K, Glatzer M, Riggenbach E, Hà HL, Cihoric N. Exploring the capabilities of large language models in answering domain-specific questions related to radiation oncology. Adv Radiat Oncol. 2024; 9: 101400.
30. Thia I, Saluja M. ChatGPT: is this patient education tool for urological malignancies readable for the general population? Res Rep Urol. 2024; 16: 31-37.
31. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, Staubli SM. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res. 2023; 25: e47479.
32. Mondal H, Dash I, Mondal S, Behera JK. ChatGPT in answering queries related to lifestyle-related diseases and disorders. Cureus. 2023; 15: e48296.
33. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, Ayoub W, Yang JD, Liran O, Spiegel B, Kuo A. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023; 29(3): 721-732.
34. Alan R, Alan BM. Utilizing ChatGPT-4 for providing information on periodontal disease to patients: a DISCERN quality analysis. Cureus. 2023; 15: e46213.
35. Haver HL, Gupta AK, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology. 2023; 307(4): e230424.
36. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology. 2023; 307(5): e230922.
37. Seth I, Lim B, Xie Y, Cevik J, Rozen WM, Ross RJ, Lee M. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: an observational study. Aesthet Surg J Open Forum. 2023; 5: ojad084.
38. Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients' education for dermatological diseases: a pilot study. Indian Dermatol Online J. 2023; 14: 482-486.
39. Ponzo V, Goitre I, Favaro E, Merio FD, Mancino MV, Riso S, Bo S. Is ChatGPT an effective tool for providing dietary advice? Nutrients. 2024; 16: 469.
40. Drabiak K. Leveraging law and ethics to promote safe and reliable AI/ML in healthcare. Front Nucl Med. 2022; 2: 983340.
41. Baumgartner R, Arora P, Bath C, Burljaev D, Ciereszko K, Custers B, Ding J, Ernst W, Fosch-Villaronga E, Galanos V, Gremsl T, Hendl T, Kropp C, Lenk C, Martin P, Mbelu S, Dos Santos Bruss SM, Napiwodzka K, Nowak E, Roxanne T, Samerski S, Schneeberger D, Tampe-Mai K, Vlantoni K, Wiggert K, Williams R. Fair and equitable AI in biomedical research and healthcare: social science perspectives. Artif Intell Med. 2023; 144: 102658.

Address for correspondence

Filip M. Tkaczyk
Doctoral School
Collegium Medicum
Jan Kochanowski University
Kielce, Poland
E-mail: filip1193@onet.pl

Received: 10.07.2025
Accepted: 4.08.2025
Online publication: 11.08.2026