Content available at: https://www.ipinnovative.com/open-access-journals

Journal of Contemporary Orthodontics

Journal homepage: https://www.jco-ios.org/

Review Article

Trends and directions of ChatGPT in healthcare- A systematic review

Gyanda Mishra[1], Shubhnita Verma[1], Prasad Chitra[1]*

[1]Dept. of Orthodontics, Army College of Dental Sciences, Secunderabad, Telangana, India.

Abstract

Background: ChatGPT (Contextual Hierarchical Attention Graphic Based Programming tool) has garnered enormous attention of healthcare practitioners, researchers and academicians. This systematic review aims to provide a critical evidence-based evaluation of the current state of ChatGPT including its applications, merits, limitations and future prospects within this domain.

Materials and Methods: A systematic review was conducted following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Comprehensive searches were performed in PubMed via Medline, Science Direct, Cochrane Library, Cochrane Central Register of Controlled Trials (CENTRAL), LILACS, SciELO, ClinicalTrials.gov and Opengrey.eu without any time or language barriers.

Results: Of the 712 studies screened, 44 met the inclusion criteria. Quantitative studies (n=26) were evaluated for methodological rigour, results, coherence and relevance using the AXIS (Appraisal tool for cross-sectional studies) tool while qualitative studies (n=18) were assessed using the GRADE-CERQual (Confidence in the Evidence from Reviews of Qualitative Research) approach.

Conclusions: The findings indicate a rapidly evolving body of low to moderate level evidence regarding applications of ChatGPT in healthcare including medical practice, education, scientific research, scientific writing, clinical decision making and virtual consultation. Limitations such as hallucination, plagiarism, lack of contextual understanding and ethical concerns related to bias, data privacy, scientific misconduct, human autonomy and authorship were evident, highlighting need for further research in generative AI, coupled with development of policy guidelines and robust ethical and legal frameworks.

Keywords ChatGPT, Healthcare, Systematic review, Artificial Intelligence, Natural language processing, Large language model

Received: 17-04-2024; **Accepted:** 10-05-2025; **Available Online**: 27-05-2025

1. Introduction

ChatGPT's resounding success has taken the world by storm, reaching 100 million users within just two months of its launch, setting a record for the fastest-growing consumer application.[1] To put its popularity in perspective, platforms like Instagram and TikTok took 2.5 years and 9 months, respectively, to achieve the same milestone.[1] Developed by OpenAI, ChatGPT (Contextual Hierarchical Attention Graphic Based Programming tool) is a general, non-domain-specific Large Language Model (LLM).[2] LLMs are a distinct class of Artificial Intelligence (AI) algorithms trained to predict word sequences based on contextual information, unlike the previous class of AI models, i.e., Deep Learning (DL) models, which focus on recognizing patterns in data. With sufficient training on vast amounts of text data, LLMs can generate innovative word sequences that represent credible human language.[2]

ChatGPT was trained on a diverse set of data (approximately 570 GB) up to the year 2021 and continuously learns and upgrades itself through interactive learning. It belongs to the GPT 3.5 model family, one of the largest LLMs with over 175 billion parameters.[2] Another example of Generative AI by OpenAI is Dali and Eve (DALL-E), which functions similarly to ChatGPT but generates digital images as outputs.[3,4]

Appreciating the significance of ChatGPT's contributions in the healthcare sector requires an understanding of its origins and development. By generating human-like text in response to natural language prompts, it can be utilized for various language tasks like content

*Corresponding author: Prasad Chitra
Email: prasadchitra24@gmail.com

https://doi.org/10.18231/j.jco.2025.023

generation, summarization, translation, and question answering, across industries such as healthcare, finance, and customer service.[5] Operating on the 'Transformer architecture', it employs self-attention mechanisms and combines unsupervised pre-training with supervised fine-tuning to produce coherent and relevant responses.[4] **Figure 1** depicts the step by step working of the ChatGPT model.

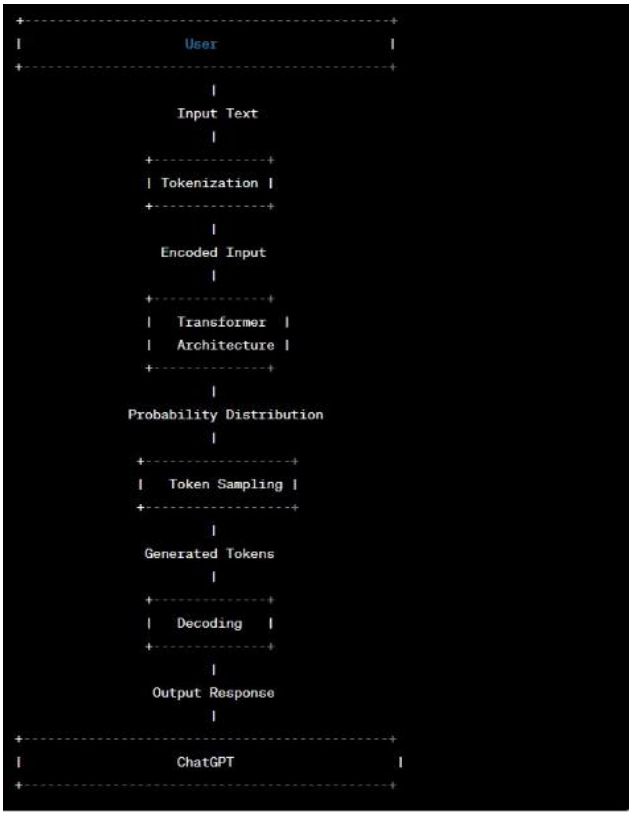

Figure 1: Chat GPT model architecture and workflow (figure created by ChatGPT)

The evolution of ChatGPT is a consequence of significant milestones and enhancements. The development of the Transformer architecture laid the foundation for efficient and scalable language models. Subsequently, the GPT series showcased the potential of AI language models in diverse applications like text generation, translation, and summarization. ChatGPT's release built on its predecessors, enhancing accuracy, context understanding, and adaptability. GPT-4's release on March 14, 2023, marked another stride in LLM advancement. It has been integrated into ChatGPT, offering reliability, creativity, and improved handling of nuanced instructions. However, limited technical details were provided by OpenAI due to competitive and safety concerns. **Figure 2** gives a brief overview of ChatGPT evolution.

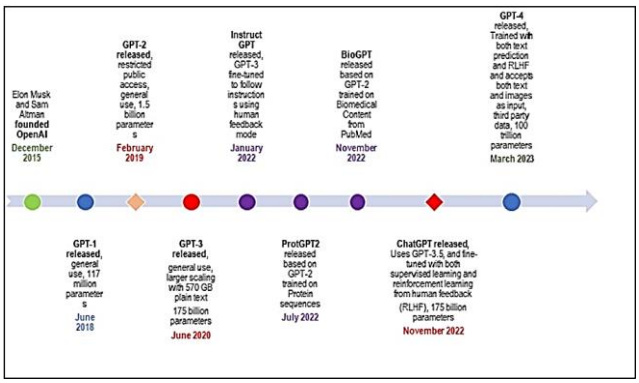

Figure 2: Evolution of ChatGPT

Despite the healthcare community adopting AI with much enthusiasm, its practical applications have remained rather restricted. However, ChatGPT has emerged as a game-changer, thanks to its advanced language processing capabilities and ability of deductive reasoning and chain of thought.[6] Within the healthcare sector, this tool can be used for tasks such as natural language processing, clinical decision support, education, research and patient engagement, providing a more comprehensive and integrated approach to healthcare, thereby potentially revolutionizing service delivery and improving patient outcomes and treatment efficiency. (**Figure 3**)

ChatGPT has demonstrated academic prowess in medicine as well as in law and business. It has also been credited as an author by a few scientific publications and has even authored a complete paper entirely on its own.[4,6 -10] As a testament to the extra-ordinary attention given by scientific and academic communities to research and development on ChatGPT, till June 2023, more than 6000 articles, blogs and news reports have been published in various journals, conferences, newspapers and social media platforms focussed on ChatGPT.

The phenomenal functional utility of the ChatGPT model has disrupted the AI community and has spawned a number of competitive models with rapidly changing technological features, each more advanced than the previous one.[11-13] While it presents novel opportunities, it has also sparked debates and concerns among academics and researchers. Thus, this paper aims to systematically review the evidence regarding various applications, merits, limitations, potential impact and future prospects of ChatGPT in healthcare and allied fields. For this qualitative research, we used the SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research) tool to formulate the research questions as under:

Figure 3: Diverse domains of application of ChatGPT in healthcare

1. RQ1: What are the domains of application of ChatGPT in the healthcare sector (including medicine, dentistry, nursing, pharmacy, alternate or indigenous medicine)?
2. RQ2: What are the stated merits of using ChatGPT in the healthcare sector?
3. RQ3: What are the challenges in using ChatGPT in the healthcare sector?
4. RQ4: What are the lacunae in literature that should be investigated in future to streamline the use of ChatGPT models in healthcare?

2. Aim

To review literature regarding ChatGPT in the healthcare sector.

3. Objectives

1. To identify the domains of applications of ChatGPT in the healthcare sector.
2. To enlist the merits and limitations of ChatGPT in the healthcare sector.
3. To explore future research areas and prospects of ChatGPT in the healthcare sector.

4. Materials and Methods

This systematic review followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.[14] It was registered in PROSPERO, the

International Prospective Register of Systematic Reviews (CRD42023442259).

3.1 Eligibility criteria

Studies were included if they reported on ChatGPT within the healthcare context—covering its applications, benefits, and limitations—using quantitative, qualitative, or mixed-methods approaches, without language or time restrictions. Editorials, commentaries, opinions, and reviews; studies on LLMs, chatbots, or NLP tools not specific to ChatGPT; non-healthcare domains; non-academic sources (e.g., blogs, news, social media); inaccessible full texts; and preprints were excluded.

3.2. Information sources

A literature search was conducted by reviewers GM and SV using English and non-English databases and grey literature, without time or language restrictions, including PubMed (Medline), ScienceDirect, Cochrane Library, CENTRAL, LILACS, SciELO, ClinicalTrials.gov, and Opengrey.eu.

3.3 Search strategy

Search strategy was developed for PUBMED/Medline using the Medical Subject Heading (MeSH) terms, keywords and

Boolean operators (Appendix I) which was then modified for other databases. The search strategies were independently verified by two reviewers (GM and SV).

3.4 Selection process

Two reviewers (GM and SV) independently screened titles and abstracts, selecting eligible articles for full-text review.

Disagreements were resolved through discussion with a third reviewer (PC). Citations of included studies were also cross-checked to ensure completeness.

3.5 Data collection process and data items

Data were extracted using a standardized form capturing study characteristics, ChatGPT model details, key findings, limitations, and future recommendations. Two reviewers (GM and SV) independently performed data extraction to minimize bias; discrepancies were resolved through discussion with a third reviewer (PC). Final data were tabulated digitally in Microsoft Excel.

3.6 Study risk of bias assessment

AXIS (Appraisal tool for cross sectional studies) tool was used to assess the worth, relevance and quality of the included studies.[15] Two investigators independently assessed and recorded 20 items on a two-point Likert Scale, including 'yes' (score: 1) and 'no' (score: 0). This tool assesses report quality (7 items), study design quality (7 items) and the possible introduction of biases (6 items). Finally, AXIS rates the quality of studies at three levels: high (70%–100%), fair (60%–69.9%) and low (0%–59.9%).

3.7 Synthesis methods

A thematic synthesis or narrative analysis of the findings to identify common themes and reported advantages, challenges and recommendations for future studies was performed. Studies were categorized according to the type of research and a keywords analysis was performed using Vos Viewer software.[16]

3.8 Certainty assessment

Qualitative studies were assessed using the GRADE-CERQual approach, which rates confidence in evidence as high, moderate, low, or very low based on methodological limitations, coherence, data adequacy, and relevance.[17]

5. Results

4.1 Study selection

A total of 937 records were imported into the Rayyan application (https://rayyan.qcri.org).[18] After removing duplicates (n=225), the remaining records were screened based on predefined eligibility criteria, resulting in 44 studies selected for full-text assessment (**Figure 4**).

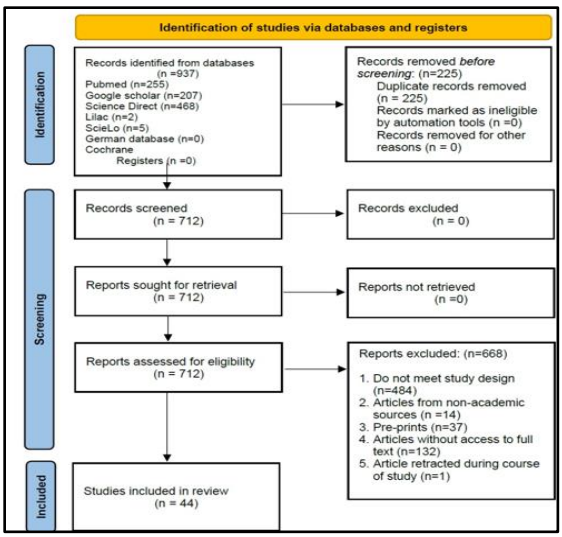

Figure 4: PRISMA flowchart 2020 depicting identification and screening of studies for inclusion in this systematic review

4.2 Study characteristics

Forty-four original research articles explored ChatGPT's role across healthcare domains using pre-defined methodologies. **Table 1** summarizes their study designs, core themes, limitations, and future recommendations. Notably, 35 studies (79.5%) utilized ChatGPT for tasks like study design, literature review, manuscript drafting, or language editing. However, only one listed ChatGPT as a co-author.

Intriguingly, some articles engaged in dialogue with ChatGPT[19-22] while some challenged ChatGPT against various examinations such as United States Medical Licence Examination (USMLE),[6] Radiology board style examinations,[23] University level medical biochemistry questions,[24] United Kingdom Bio-Medical Admissions Tests,[25] Assessment for membership of Royal College of Obstetricians and Gynaecologists,[26] Membership of the Royal College of General Practitioners Applied Knowledge Test,[27] Parasitology exam by Korea Health Personnel Licensing Examination Institute,[28] European Exam in core cardiology,[29] and Ophthalmic Knowledge Assessment Program (OKAP) examination.[30]

Of these, 26 quantitative studies were critically appraised using the AXIS tool and the remaining 18 qualitative studies were evaluated using the GRADE-CERQual approach. **Figure 5** highlights the most frequent keywords, reflecting the diverse range of ChatGPT-related topics in healthcare.

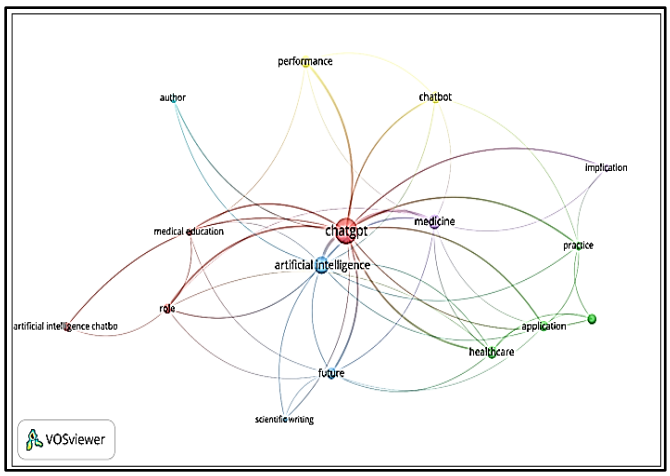

Figure 5: Bibliometric analysis of keywords most frequently associated with publications focussed on ChatGPT applications in healthcare; created by VOS viewer

An analysis of the countries of origin of the included works reveals the global interest in ChatGPT in healthcare, with the USA, India, and China leading in publication count as shown in **Figure 6**.

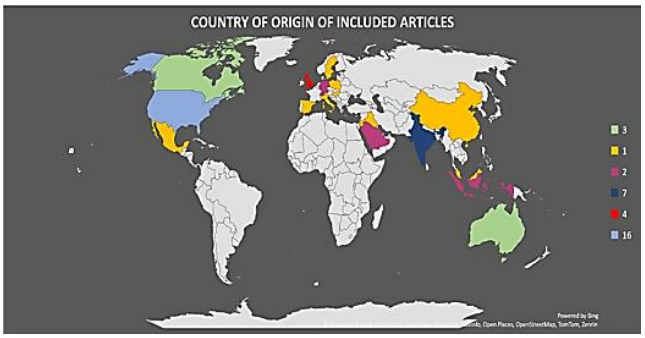

Figure 6: Geographical distribution map based on the number of publications on ChatGPT in healthcare from different countries.

4.3 Risk of bias in studies

All 26 studies appraised had a low risk of bias. (Appendix II) The majority of studies suffered from methodological limitations; 17 studies did not justify sample size, one study did not clearly report the statistical analysis used,[31] nine studies had selection bias,[23,24,32-38] four studies did not discuss their limitations,[29,30,36,39] 10 studies did not mention ethical clearance[6,23,25,30,32,35,36,40-42] and one did not disclose conflicts of interest.[33]

4.4 Thematic synthesis of studies

A narrative synthesis of the evidence was performed based on the core themes of applications, limitations and future recommendations mentioned in the included studies.

4.4.1 Clinical decision support

ChatGPT has been used for generating electronic health records, preauthorization, computer-aided diagnoses, differential diagnosis, treatment recommendations, patient triage, and postoperative support.[43–45] Ma et al. showed its ability to summarize patient charts and gather diagnostic data in gastrointestinal pathology.[46] While ChatGPT can suggest differential diagnoses from patient data, its diagnostic and treatment suggestions are often non-specific.[47–49] However, Zhou et al. reported that it can offer personalized treatment, dietary, and follow-up advice.[50] It can also interpret medical imaging in near real-time and efficiently summarize large medical records, minimizing chart review time.[45]

4.4.2 Lab reporting

Tested ChatGPT's ability to draft radiology reports, finding that while it scored well on report quality, it struggled with technical medical terminology. Bosbach et al.[39] concluded that the ChatGPT translated reports were patient-friendly, using common language instead of medical jargon, and integrated information comprehensively for better patient understanding. Lyu et al.[36] They also noted that when the same radiological report is used as a prompt on several occasions, the translated responses are inconsistent and potentially misleading due to oversimplification or information loss. This is consistent with the findings of Zhou et al who also observed that though ChatGPT can produce coherent, comprehensive, clinically relevant and accurate medical reports, the responses were sometimes inconsistent due to language model uncertainty, which could compromise translation quality.[50] In addition, Cadamuro et al reported that ChatGPT's interpretation of medical test results are superficial and do not offer any suggestions for follow up diagnostics.[51] It is worth noting that discerning whether the report was generated by a human or an AI tool proved to be difficult.[50,52]

4.4.3 Patient communication and education

When assessed by experts, ChatGPT responses were adjudged as trustworthy and valuable for patients.[32,40,49,53] ChatGPT answers were rated similar to non-AI answers with regard to ease of understanding, scientific adequacy and user rating in terms of satisfaction and acceptance.[32,35,37,54,55] Furthermore, the experts could hardly discern between ChatGPT and non-AI suggestions.[35,36] ChatGPT offered practical and comprehensive advice to patients and caregivers, regardless of the level of difficulty of the inquiry.[56,57] According to Hopkins et al, ChatGPT's responses were friendlier and more comforting to patients compared to Google search results.[55]

Table 1: Study characteristics

S No	Authors	Year	RQ 1: Domain of application	RQ 2: Merits of using ChatGPT in healthcare	RQ 3: Challenges in using ChatGPT in healthcare	RQ 4: Future recommendations to streamline use of ChatGPT in healthcare
1.	Kung et al.[6]	2023	Medical education.	1. Performed at or near the passing threshold for all three parts of USMLE without any specialized training or reinforcement. 2. Demonstrates high level of concordance and insight in its explanations.	1. Responses reflect cognitive human bias.	1. Better training of model. 2. Open science infrastructure for democratization of resources and access.
2.	Almazyad M et al.[20]	2023	Scientific writing.	1. Can analyse qualitative data to extract key themes, patterns and insights. 2. Can perform text summarization without loss of meaning. 3. Can suggest research topics and Frame research proposals.	1. Propagates bias and inaccuracies. 2. Limited training data. 3. Over-dependence on AI may. hamper critical thinking in HCWs.	1. Further research warranted to streamline its use in different healthcare scenarios.
3.	Bhayana et al.[23]	2023	Medical Education.	1. Can Pass board style radiology examinations without images.	1. Struggles with higher order thinking questions involving description of imaging findings, calculation and classification, and application of concepts.	1. Applications with radiology- specific pre-training should be explored.
4.	Ghosh et al.[24]	2023	Medical Education.	1. Correctly answers basic and higher-order questions in medical biochemistry.	1. Lacks knowledge of recent advances in medical biochemistry.	1. Continuous and updated domain specific training of model. 2. Training students to optimally utilise ChatGPT for supplementing education.
5.	Giannos et al.[25]	2023	Medical education.	1. Performance and score in section 1 of BMAT comparable to average human scores. 2. Effective as educational tool for exam practice.	1. Difficulty with higher order thinking and difficult questions in section 2 of BMAT with significantly low accuracy. 2. lack of contextual understanding. 3. Ethical concerns regarding potential for academic misconduct.	1. Domain specific training 2. Better algorithms to improve the model.

6.	Li et al.[26]	2023	Medical education.	1. Outscores human candidates in mock objective structured clinical examination simulating assessment for membership of the Royal College of Obstetricians and Gynaecologists.	1. Lacks in-depth understanding of medical terms.	1. Call to action for academic medicine to evolve its understanding and conception of evaluations and assessments with regard to AI based tools.
7.	Thirunavukarasu et al.[27]	2023	Medical education.	1. Could partially answer simulated questions from the Membership of the Royal College of General Practitioners Applied Knowledge Test, although mean score of ChatGPT was below that of qualified primary care physicians.	1. Inaccuracies in responses limit autonomous application of the tool.	1. Expert validation of training data for better models.
8.	Huh S.[28]	2023	Medical education.	1. Can correctly answer questions from Korean parasitology examination, however, it scored lower than medical students.	1. Limited contextual and in-depth understanding limit its use for medical education.	1. Better domain specific model training. 2. Updating medical. curriculum to incorporate AI based tools. 3. Increasing knowledge and awareness of medical faculty and students with regard to utilising AI based tools.
9.	Skalidis et al.[29]	2023	Medical Education.	1. Accurately answered and surpassed passing threshold in European Exam in Core Cardiology.	1. Lacks critical thinking ability.	1. Research to explore role of AI tools in medical standardized Examinations.
10.	Antaki et al.[30]	2023	Medical education.	1. Accuracy of ChatGPT in answering questions from OKAP only slightly below historical human averages.	1. Demonstrates difficulty with higher order thinking and difficult questions. 2. Cannot process image-based data.	1. Multimodal integration with other systems. 2. Domain specific training.
11.	Athaluri et al.[31]	2023	Scientific writing.	1. Can generate research proposals	1. Provides incorrect referencing (hallucination).	1. Improving the training inputs by including diverse, accurate, and contextually relevant data sets along with frequent updates.

12.	Van Bulck et al.[32]	2023	Patient communication and education.	1. Responses to virtual patient questions were rated trustworthy and valuable advice by experts.	1. Incomplete and occasionally misleading information. 2. Hallucination. 3. Sensitive to slight changes in input prompt and may produce confusing responses.	1. More research to improve accuracy and reliability.
13.	Gao et al.[33]	2023	Scientific writing.	1. Can generate abstracts which are difficult to distinguish from abstracts written by humans.	1. Potential for scientific misconduct by paper mills who may take advantage to falsify research.	1. Improved plagiarism detector software's with embedded watermarks to identify AI generated content.
14.	Juhi et al.[34]	2023	Pharmacology.	1. Effectively predicts and explains drug interactions.	1. Missing information.	1. Further research to improve the model.
15.	Lee et al.[35]	2023	Patient communication and education.	1. Answers to common questions about colonoscopy are easy to understand, scientifically adequate and satisfy the user. 2. AI generated response cannot be differentiated from human responses.	1. Propagates inherent bias. 2. Inconsistent responses, i.e., different response with same input prompt on different occasions.	1. Improving the model through medical domain specific training.
16.	Lyu et al.[36]	2023	1. Medical Practice. 2. Patient communication.	1. Can translate radiology report into common language.	1. Occasional over simplification and missing information. 2. Inconsistent responses, i.e., different response with same input prompt on different occasions. 3. Does not have a built-in template for the generated report translations.	1. Further developments of model through improvement in prompt engineering. 2. Additional evaluations of performance characteristics for purposes of regulatory review and user adoption.
17.	Shahsavar et al.[37]	2023	Patient communication and education.	1. Provides self-diagnosis and treatment recommendations in user friendly language with high user acceptance.	Not mentioned.	1. Policy and ethical guidelines for using ChatGPT in healthcare. 2. Medical domain specific training of the model. 3. Increased awareness among users about the utility and limitations of the model.

18.	Sinha et al.[38]	2023	Medical education.	1. Can provide answers to high order reasoning questions in pathology with high relational level of accuracy.	1. Missing information due to limited training data. 2. Lacks in-depth understanding of medical terms.	1. Validation of newer models by training on updated data and continuous updates. 2. Curriculum modification and training of students as well as teachers to adapt to AI tools in education and assessment.
19.	Bosbach et al.[39]	2023	Scientific writing.	1. Capable of generating radiology reports. 2. Can provides language and editing support. 3. Capability to adjust output files in response to minor changes in input command files.	1. Faces difficulty with technical/medical terms.	1. future applications should explore potential for mass manipulation to increase productivity of radiology reporting.
20.	Chervenak et al[40].	2023	Patient communication and education.	1. Can provide relevant, meaningful responses to fertility related clinical queries of lay persons comparable to established sources.	1.Unreliable citation of sources (hallucination). 2. Lack of medical domain specific Knowledge. 3. Propagates biases due to 'black box' nature of algorithms.	1. Medical domain specific training to improve output.
21.	Johnson et al.[41]	2023	Patient communication and education.	1. Provides accurate and relevant responses about common cancer myths and misconceptions, irrespective of difficulty of query.	1. Responses may be incomplete and occasionally misleading 2. lack of in-depth understanding of technical medical terms	1. Domain specific training. 2. Supervision by human experts.
22.	Lahat et al.[42]	2023	Research.	1. Can conduct literature search and identify research questions.	1. Lacks novelty	1. Further research warranted to update its training data set.
23.	Balas et al[43]	2023	Medical practice.	1. Can provide correct diagnosis and differential diagnosis. 2. Accuracy of response can be improved through reinforcement learning.	1. Provides incorrect referencing (hallucination). 2. Lacks medical domain specific Knowledge.	1. Further research warranted into capabilities and limitations of the model and ways of integrating it into clinical practice.

24.	Kim HY.[47]	2023	1. Medical Practice. 2. Scientific writing.	1. Can analyse large datasets of patient data to provide differential diagnosis and treatment recommendations. 2. Speed up clinical documentation, work flow and augment patient care. 3. Can draft a case report, when provided with relevan information. 4. Provides language support and feedback.	1. Does not provide specific diagnostic or treatment recommendations. 2. Ethical concerns regarding medical decisions made by AI tools. 3. Incorrect referencing (hallucination).	1. Research to determine long term dependability and reliability of AI intervention in healthcare. 2. Integration into other clinical models.
25.	Sonntagbauer et al.[48]	2023	1. Medical practice. 2. Scientific writing.	1. Can assist the practitioner by providing computer aided diagnosis. 2. Can assist with documentation of medical records.	1. Propagates inaccuracies and bias. 2. Factually correct but contextually incorrect or irrelevant response (hallucination). 3. Ethical concerns regarding authorship, Plagiarism and scientific misconduct.	1. Research to examine possible applications and challenges of ChatGPT in healthcare
26.	Xie et a.l[49]	2023	Patient communication and education.	1. Able to provide coherent, comprehensive response to patient questions on aesthetic plastic surgery in a natural language 2. Admits its lack of ability to provide medical advice.	1. Lacks in-depth understanding of medical terms. 2. Cannot provide detailed or specific treatment advice.	1. Further research to explore the potential of ChatGPT in digital clinical guidance and the broader healthcare context through better training data.
27.	Zhou Z.[50]	2023	1. Medical practice. 2. Scientific Writing.	1. Can produce coherent, comprehensive, and clinically relevant, accurate medical reports 2. Provided tailored recommendations addressing identified concerns and abnormalities including lifestyle modifications, follow up medical treatment options and specialist consultation.	1. Lacks multimodal processing capabilities and cannot interpret typical X-rays, MRIs, or other medical images. 2. cannot provide quantitative suggestions on medication details.	1. Research to integrate AI tools like ChatGPT in the healthcare sector to expand patient-centred care delivery
28.	Cadamuro et al.[51]	2023	Scientific writing.	1. Can interpret laboratory test results and generate laboratory reports.	1. Interpretation of laboratory test results is superficial and lacks coherence. 2. Does not suggest follow up diagnostics.	1. Medical domain specific training may improve output.

29.	Altmae et al.[52]	2023	1. Scientific Writing 2. Research.	1. Can design research studies 2. Can conduct literature search and identify lacunae in literature 3. Can generate abstract, title and parts of manuscript.	1. Provides incorrect referencing (hallucination) 2. Can pass plagiarism detection tools. 3. Inaccuracies in response. 4. Propagates misinformation.	1. Requires continuous training on domain specific data.
30.	Samaan et al. [53]	2023	Patient communication and education.	1. Can provide accurate, reproducible and comprehensive responses to common questions related to bariatric Surgery.	1. Bias. 2. Missing information.	1. Further research towards use of AI for improving patient experience and outcomes in bariatric surgery.
31.	Hopkins et al.[55]	2023	Patient communication and education.	1. Can act as a virtual assistant to patients and formulate concise and relevant responses to basic fact based and complex clinical questions. 2. Its responses are less alarming than those of Google's features snippet.	1. Incomplete information due to limited training data and lack of domain specific knowledge. 2.References provided may not be correct and true (hallucination). 3. Inconsistent responses, i.e., the response differs with same prompt on different occasions. 4. Lacks access to other databases and models through internet.	1. Policy development by regulators for minimum standards. 2. Raising awareness among patients.
32.	Sallam et al.[57]	2023	Patient communication and education.	1. Can provide clear and concise source of medical information for patients on COVID-19 related controversies.	1. Missing information. 2. Bias.	1. Medical domain specific training of the model. 2. Further studies needed to quantitatively assess the implications of this new technology as a reliable source of information.
33.	Webb et al.[59]	2023	Medical education	1. Can simulate a patient for training of communications skills and teaching empathy.	1. Inaccuracies and inconsistency in responses due to limited training data. 2. Inadequate emotional/ sentient Capability.	1. Training of the model based on communication theories and up to date data. 2. Improved prompt engineering.

34.	Amri et al.[60]	2023	1. Medical education 2. Scientific writing 3. Patient interaction	1. Can simulate patient interaction. 2. Can assist with writing parts of manuscript, abstracts. 3. Can provide language and editing Support. 4. Can suggest research topics. 5. Can generate assignments and exam practice.	1. Difficult to differentiate generated content from that written by a human. 2.Propagates bias. 3.Threatens data privacy. 4.potential for academic misconduct.	1. Policy guidelines need to be developed for ethical use of Chat GPT.
35.	Hisan et al.[61]	2023	1. Medical education. 2. Scientific writing.	1. Can generate abstracts difficult to distinguish from abstracts written by humans. 2. Can partially answer Indonesian medical doctor examination questions.	1. Potential for academic and scientific misconduct. 2. confidently provides factually incorrect answer (hallucination) 3. Lacks contextual understanding for medical questions.	1. Ethical safeguards to ensure responsible use.
36.	Yeo et al.[62]	2023	Patient communication and education.	1. Exhibited extensive knowledge of patient questions related to liver cirrhosis and hepatic cell carcinoma, particularly related to basic. knowledge, lifestyle, and treatment 2. Provided practical and multifaceted advice to patients and caregivers regarding the next steps and adjusting to a new diagnosis.	1. Lacks detailed and comprehensive medical knowledge of domains of diagnosis and preventive medicine. 2. Lacks knowledge of regional guidelines variations, such as HCC screening criteria. 3. Outdated/incorrect response due to limited training data.	1. Further research to improve the model's accuracy.
37.	Mijwil et al.[63]	2023	1. Scientific writing. 2. Research.	1. Can generate title for manuscripts 2. Can create content for parts of manuscript 3. Can conduct literature searches	1. Ethical concerns regarding plagiarism. 2. Lacks academic style of writing	1. Should be used only as a supplementary tool with human supervision at all stages.
38.	Lal Vallath et al.[66]	2023	1. Scientific writing. 2. Research.	1. Can write parts of a manuscript for case reports. 2. Can conduct literature review. 3. Can correct grammatical errors.	1. Risk of inherent bias. 2. Lack of context. 3. Incorrect referencing (hallucination). 4. Unintentional plagiarism as the content is sourced from the training data.	1. Further research for rapid integration of AI tools in healthcare.
39.	Raxwal et al.[68]	2023	Scientific writing.	1. Can generate title, abstract, introduction and conclusion sections of a case report.	1. Lacks originality. 2. Lacks context. 3. Incorrect referencing and citations (hallucination). 4. Lacks ability to critically think and analyse.	1. Should be used only as a supplementary tool with human supervision at all stages.

40.	Hegde et al.[69]	2023	1. Medical practice. 2. Scientific writing.	1. Can generate relevant output in response to a medical query. 2. Can draft parts of manuscript.	1. Responses lack of specificity and in- depth understanding of medical issues. 2. Provides incorrect referencing (hallucination). 3. May affect human researchers' autonomy. 4. Unsupervised written text may be misleading and distort facts.	1. Exercise caution while using ChatGPT and all outputs should be supervised by humans
41.	Zamarud et al.[71]	2023	Scientific writing.	1. Can draft parts of a manuscript for case report	1. Incorrect referencing (hallucination). 2. Lacks of domain specific knowledge. 3. Not integrated with domain specific applications like PubMed.	Not mentioned
42.	Schuppe et al.[72]	2023	Scientific writing.	1. Can create outline and draft parts of manuscript of a case report. 2. Provides language support.	1. Incorrect referencing (hallucination).	1. Vetting of ChatGPT output by human researcher.
43.	Chu et al.[73]	2023	Scientific writing.	1. Can draft a manuscript.	Not mentioned.	Not mentioned.
44.	Babl et al.[74]	2023	Scientific writing.	1. Can draft abstracts.	1. Provides incorrect referencing (hallucination).	1. Guidelines for publishing, authorship and plagiarism.

*RQ= Review Question, HCW= health care worker, OKAP=Ophthalmic Knowledge Assessment Program, UK-BMAT= United Kingdom Biomedical Admission test, AI= Artificial Intelligenc

Table 2: Summary of qualitative findings

S. No.	Summary of review finding	Studies contributing to the review finding	CERQual assessment of confidence in the evidence	Explanation of CERQual assessment
	Scientific writing - ChatGPT can be utilized for scientific writing.			
1.1	Case report writing – ChatGPT can write case reports when prompted to do so if relevant data are provided.	47,66,68,69, 71,72,73	Very low confidence.	Serious concerns regarding methodological limitations and coherence and moderate concerns regarding adequacy of data. No concerns regarding relevance of any study.
1.2	Generate abstract, titles and parts of manuscript - ChatGPT can generate abstract, title and parts of manuscript when prompted to do so if relevant data are provided.	52,60,61, 63,74	Low confidence.	Moderate concerns regarding methodological limitations and coherence and serious concerns regarding adequacy of data. No concerns regarding relevance of any study.
1.3	Text summarization - ChatGPT can summarize text, reduce word count, provide editing, language and grammar support.	48,63,66, 70,72	Low confidence.	Moderate concerns regarding methodological limitations, coherence and adequacy of data from a single study. Moderate methodological limitations due to researcher reflexivity and insufficient data collection. Moderate concerns about coherence due to nonspecific aim and incomplete data. No concerns regarding relevance of any study.
1.4	Generate laboratory report - ChatGPT can generate laboratory report when prompted to do so if relevant data are provided.	50	Moderate confidence.	Minor concerns regarding methodological limitations, coherence and adequacy of data from a single study. No concerns regarding relevance of any study.
2.	Clinical Decision Support - ChatGPT can assist healthcare workers with Clinical Decision Support.			
2.1	Differential diagnosis - ChatGPT can provide a list of differential diagnosis when prompted to do so if relevant data are provided.	47, 48	Low confidence.	Moderate concerns about methodological limitations, coherence and adequacy of data. No concerns regarding relevance of any study.
2.2	Treatment planning - ChatGPT can formulate treatment plan when prompted to do so if relevant data are provided.	47, 49	Moderate confidence.	Moderate concerns about adequacy of data. No concerns regarding relevance of any study.

3.	Research - ChatGPT can assist healthcare workers with research			
3.1	Study design, literature review, statistical analysis - ChatGPT can design a study, perform literature review and provide support for statistical analysis.	[52]	Moderate confidence.	Moderate methodological limitations and serious concerns about adequacy of data from a single study. No concerns regarding relevance of any study.
3.2	Suggest research topic - ChatGPT can identify lacunae in literature and suggest topics for research.	[60]	Low confidence.	Serious concerns about methodology, coherence and adequacy of data from a single study. Serious concerns about adequacy of data due to small sample size (single experiment) and lack of detailed data Serious methodological limitations due to researcher reflexivity, method of data collection not discussed and insufficient data presented.
4	Education - ChatGPT can be useful in healthcare education.			
4.1	Answer medical questions and Assist with Assignments - ChatGPT can support healthcare students by answering medical questions and assisting with assignments.	[61]	Very low confidence.	Moderate concerns about adequacy of data from a single study. Moderate concerns about adequacy of data due to small sample size and lack of details. No concerns regarding relevance of any study.
4.2	Patient stimulate and role play - ChatGPT can augment healthcare education by simulating patients and role-play.	[59, 60]	Low confidence.	Moderate concerns regarding methodology, coherence and adequacy of data. No concerns regarding relevance of any study.
4.3	Generate exam practice - ChatGPT can augment healthcare education by generating questions/ clinical vignettes for exam practice.	[60]	Low confidence.	Serious concerns regarding methodology, coherence and adequacy of data. No concerns regarding relevance of any study.
5.	Patient communication - ChatGPT can provide support to patients by answering their health-related questions and acting as a virtual consultant.	[55, 49]	Moderate confidence.	Minor methodological limitations and moderate concerns about adequacy of data. No concerns regarding relevance of any study.
6.	Advantages/ Benefits of using ChatGPT - Using ChatGPT in healthcare can be advantageous/ beneficial.			
6.1	Natural language - ChatGPT created text is similar to natural language.	[20,48,49, 50, 52,55,59, 60, 61, 63, 74]	Low confidence.	Moderate concerns regarding methodology, coherence and adequacy of data. No concerns regarding relevance of any study.
6.2	Low index of similarity - ChatGPT created text has low index of similarity.	[52, 63, 69]	Moderate confidence.	Moderate concerns regarding methodology and adequacy of data and minor concerns about coherence. No concerns regarding relevance of any study.

6.3	Time saving / expedites work /optimize health care worker time and resources - Using ChatGPT saves time by expediting clinical and administrative work, optimizes health care workers' time and resources	47,50, 63, 66	Moderate confidence.	Moderate concerns regarding methodology and adequacy of data and minor concerns about coherence. No concerns regarding relevance of any study.
6.4	Can work with imperfect user input - ChatGPT can work with imperfect user input in terms of spelling and grammatical errors and incomplete sentences	50	Low confidence.	Data available from only a single study. No concerns regarding relevance of any study.
6.5	Democratic access to education material - ChatGPT has the potential to democratize access to education material irrespective of training/classroom location / time	50	Low confidence.	Data available from only a single study Very minor methodological limitations due to researcher reflexivity. No concerns regarding relevance of any study.
6.6	Recognizes its own limitations giving medical advice - ChatGPT acknowledges and cautions against its own limitations regarding giving medical advice to patients	49, 55	Moderate confidence.	Minor concerns regarding methodology, moderate concerns about adequacy of data and no concerns about coherence and relevance. No concerns regarding relevance of any study.
6.7	Relevant output - ChatGPT output is relevant to the context inputted in the prompt	47, 48, 69	Low confidence.	Moderate concerns regarding methodology and coherence and serious concerns about adequacy of data. No concerns regarding relevance of any study.
7.	Limitations - ChatGPT has some limitations when used for healthcare related work			
7.1	Hallucination - ChatGPT suffers from 'hallucination', i.e. experiences the apparent perception of something not present.	47,48,52, 55, 63,66,68, 69, 72,74	Very low confidence.	Moderate concerns regarding methodology and coherence and serious concerns about adequacy of data. No concerns regarding relevance of any study.
7.2	Lacks in depth understanding / critical / contextual difficulty - ChatGPT lacks in- depth and contextual understanding of medical terminology and lacks the ability to critically analyze questions.	49,61,63, 66, 68, 69	Low confidence.	Moderate concerns regarding methodology and coherence and serious concerns about adequacy of data. No concerns regarding relevance of any study.
7.3	Bias - ChatGPT responses reflect and propagate inherent biases.	20,48,60, 63, 66, 69	Very low confidence.	Moderate concerns regarding methodology and coherence and serious concerns about adequacy of data. No concerns regarding relevance of any study.
7.4	Inconsistency in response on different instances of same prompt - ChatGPT output is limited by inconsistency in responses on different instances of same prompt.	52,55, 59, 63	Moderate confidence.	Moderate concerns regarding methodology and adequacy of data and minor concerns about coherence. No concerns regarding relevance of any study.

7.5	Non – academic writing style - ChatGPT generated text reflects non – academic writing style.	[63]	Very low confidence.	Serious concerns regarding methodology, and adequacy of data and moderate concerns about coherence. serious concerns about adequacy of data due to small sample size (single experiment) and lack of detailed data Moderate concerns about coherence due to incomplete data analysis and discussion Serious methodological limitations due to researcher reflexivity, inadequately described methodology and insufficient data collection.
7.6	ChatGPT lacks access to other database/websites /CDM tools as it is not connected to internet.	[55, 71]	Very low confidence.	Serious concerns regarding methodology and adequacy of data and moderate concerns about coherence. No concerns regarding relevance of any study.
7.7	ChatGPT cannot give quantified prescription advice.	[50]	Low confidence.	Data available from only a single study. Minor concerns about adequacy of data as the number of researchers and method of comparison are not clear. No concerns regarding relevance of any study.
7.8	ChatGPT lacks multi modal processing capabilities, i.e. it cannot work with visual/graphic/ audio prompts.	[50]	Low confidence.	Data available from only a single study. Minor concerns about adequacy of data as the number of researchers and method of comparison are not clear. No concerns regarding relevance of any study.
7.9	ChatGPT cannot recommend follow up examination.	[48, 50]	Low confidence.	Moderate concerns regarding methodology and adequacy of data and minor concerns about coherence. No concerns regarding relevance of any study.
8.	Ethical concerns - Use of ChatGPT in healthcare applications raises many ethical concerns.			
8.1	ChatGPT created content lacks originality and exhibits “unintentional plagiarism”.	[48,61, 63, 68]	Very low confidence.	Serious concerns regarding adequacy of data and moderate concerns about methodology and coherence. No concerns regarding relevance of any study.
8.2	Academic misconduct - Use of ChatGPT in healthcare education provides opportunity for academic misconduct, e.g., cheating in exams, projects/assignments written by ChatGPT.	[60, 61]	Very low confidence.	Serious concerns regarding adequacy of data and moderate concerns about methodology and coherence. No concerns regarding relevance of any study.

8.3	Scientific misconduct - Use of ChatGPT in healthcare research provides opportunity for scientific misconduct, e.g. plagiarism.	[48, 60, 61]	Very low confidence.	Serious concerns regarding adequacy of data and moderate concerns about methodology and coherence. No concerns regarding relevance of any study.
8.4	Threat to human autonomy - Use of ChatGPT can pose a threat to human autonomy and stifle critical thinking in healthcare practitioners and researchers.	[20, 47, 69]	Very low confidence.	Serious concerns regarding adequacy of data and moderate concerns about methodology and coherence. No concerns regarding relevance of any study.
8.5	Roles as author - Giving authorship credits to ChatGPT is unethical as the application cannot take responsibility and accountability for the work produced.	[48, 52, 61]	Very low confidence.	Serious concerns regarding adequacy of data and moderate concerns about methodology and coherence. No concerns regarding relevance of any study.
8.6	Patient privacy and informed consent - Use of ChatGPT in healthcare practice endangers patient privacy and data rights.	[20, 60]	Very low confidence.	Serious concerns regarding methodology and adequacy of data and moderate concerns about coherence. Serious methodological limitations due to researcher reflexivity, method of data collection not discussed and insufficient data presented. No concerns regarding relevance of any study.
9.	Future recommendations- Future recommendations for using ChatGPT in healthcare.			
9.1	There is a need to improved models through better AI algorithms, prompt engineering, Domain specific continual training, Integration into other clinical models, etc.	[47, 52]	Moderate confidence.	Moderate concerns regarding methodology and adequacy of data. No concerns regarding relevance of any study.
9.2	Formulation of ethical and policy guidelines for use of ChatGPT in healthcare practice, education and research.	[61, 74]	Very low confidence.	Moderate concerns regarding methodology and coherence and serious concerns about adequacy of data. No concerns regarding relevance of any study.
9.3	There is need to spread awareness among regulators, health care workers and patients regarding advantages and limitations of using ChatGPT for healthcare related work.	[55]	Moderate confidence.	Moderate concerns regarding methodology and adequacy of data. Moderate concerns about adequacy of data due to small sample size and lack of detailed data. Moderate methodological limitations due to researcher reflexivity, credibility of data not discussed, changes in prompt for five different responses not discussed/disclosed. No concerns regarding relevance of any study.

It effectively predicted and explained drug-drug interactions to patients in a simple, easy-to-read language.[34] However, some limitations were reported such as missing information, vagueness, and lack of patient-centred language.[32,34,57,58] Although Samaan et al found it to be a

helpful resource for patients regarding bariatric surgery, it cannot provide quantitative medication details.[32,53]

4.4.4 Education

Chat GPT could solve higher order reasoning questions in pathology, thereby enabling students to improve their clinical decision-making abilities.[38] It assisted students in navigating e-learning resources and improving engagement with curricula, leading to better learning outcomes and student experiences.[59] It was able to simulate patient-doctor interactions, allowing students to practice diagnostic and communication skills in a controlled setting with real-time feedback.[59-61] It could generate assignments individualised for specific learning objectives and provided exam practice with feedback.[60]

Kung et al assessed ChatGPT's performance on the US Medical Licensing Examination (USMLE) which consists of three standardized tests. After removing the image-based questions, ChatGPT's accuracy approached or exceeded the 60% passing threshold with a high degree of concordance and insight.[6] Paradoxically, in the same study, ChatGPT outperformed PubMed GPT, a counterpart LLM with similar neural structure, but trained exclusively on biomedical domain literature. It also demonstrated a high level of accuracy in solving higher-order reasoning questions in pathology, with approximately 80% correct answers.[38] In another investigation, ChatGPT's answers on cirrhosis and hepatic cell carcinoma were mostly correct (79.1% and 74%, respectively) although more than half the answers lacked comprehensiveness.[62] They also stated that the model performed better in areas of basic knowledge, lifestyle, and therapy when compared to diagnosis and preventative medicine.

Despite lacking radiology-specific pretraining, ChatGPT almost passed a radiology board-style exam without images, excelling in lower-order thinking and clinical management questions.[23] However, it struggled with higher-order thinking questions in biochemistry, ophthalmology, radiology image classification and concept application.[23,24,30]Although ChatGPT outperformed human candidates in a virtual obstetrics and gynaecology examination, it occasionally provided confidently phrased inaccurate responses.[26] In parasitology, ChatGPT's performance was inferior to that of a Korean medical student and it had only partial reliability in pharmacology when it came to predicting and explaining drug-drug interactions.[28,34] Similarly, ChatGPT scored lower on the Membership of the Royal College of General Practitioners Applied Knowledge Test compared to qualified primary care physicians and displayed significantly low accuracy in UK BMAT.[25,27]

4.4.5 Research

ChatGPT demonstrated the ability to handle various research tasks such as designing search strategy,[52,54] identifying research gaps,[42,60,63] framing review questions,[31,42,64] and conducting a review of literature with detailed referencing.[64-67] It proved capable of providing statistical and programming language support to assist researchers with analyses of the results of a study.[54] Taking a novel approach, Almazyad and colleagues applied ChatGPT to analyse transcribed data from a panel discussion on paediatric research. The model successfully analysed qualitative data, extracting key themes, patterns and insights from the participants' conversations. Additionally, it generated a concise summary of the discussions, highlighting main points and conclusions reached by the participants.[20]

4.4.6 Scientific writing

When given the appropriate inputs, ChatGPT proved capable of drafting sections of a manuscript and generating titles and abstracts.[45,47,60,63,66,68-74] It could correct grammatical errors and paraphrase for improved clarity.[47,60,66,70,72] The text generated was described as relevant, structured, concise and linguistically coherent.[63,69,70] Further, the model exhibited the ability to improve its performance through its reinforcement learning model by interacting with and obtaining feedback of human users.[45] On the downside, ChatGPT's style of writing lacked academic quality and tone and the quality of output was inconsistent.[35] Many authors noted that the responses lacked specificity and in-depth understanding of context.[54,68-76]

Despite the successful generation of an output that is factually and grammatically correct, a major limitation of ChatGPT was a high degree of unintentional plagiarism;[42,64,68] and incorrect or fictitious references, which at first glance appear to be legitimate.[52,55,64-76] Athaluri et al reported that out of the 178 references cited by ChatGPT, 69 references did not have a valid DOI.[31] The DOIs cited either belonged to another publication or were outright fabrications termed as 'hallucination'. Temsah et al addressed this issue by carefully analysing ChatGPT's output and cross-referencing it with original articles.[65] In an unorthodox study, Zhavoronkov collaborated with ChatGPT, to discuss the potential applications of Rapamycin based on Pascal's Wager philosophical argument commonly utilized to justify the belief in God.[9] ChatGPT succinctly provided the pros and cons for the use of Rapamycin considering the preclinical evidence of potential life extension in animals.

In a study by Gao et al, human evaluators could only correctly identify 68% of abstracts generated by ChatGPT as being generated, and 86% of original abstracts as being original.[33] Ironically, generated abstracts performed better on

anti-plagiarism tools compared to original abstracts. Both human evaluators and AI detectors struggled to distinguish perfectly between original and ChatGPT-generated abstracts, with some real abstracts being misclassified as generated.[33] Other studies back up the conclusion that ChatGPT can deceive professionals and experts.[60-63]

Many case reports were written with the assistance of ChatGPT.[47,68-73] The chatbot assisted with literature search for writing review of literature, enlisted differential diagnosis, suggested investigations, drafted discussion and created titles and abstracts. It effectively edited text and improved language but struggled with providing references and citations. The erroneous references had surprisingly plausible titles, journal names and authors.[72] It could also assist researchers by identifying and suggesting potentially suitable journals and potential reviewers, as tested and validated by Xames et al.[77] Hosseini and Horbach demonstrated the feasibility of using ChatGPT to write grammatically correct and coherent peer reviews, benefiting researchers, especially non-native speakers of supported languages.[78] It can veritably summarize text, with clear description of the main goal and conclusions of the paper and offer constructive criticism.[75]

4.4.7 Limitations and challenges

In addition to the issues of plagiarism, hallucination, difficulty with higher order thinking, and lack of depth of understanding, the other reported limitations of ChatGPT are bias,[20,35,48,60,66] lack of contextual understanding of medical terms,[25,28,38,49] difficulty with technical medical terms,[39,40,49,61,79] inability to read or generate visual content,[30,50] limited domain-specific expertise,[30,38,40,43,54,55,71] inability to interact with external systems,[30,50,71] missing information,[34,44,53,55,57] misinformation,[32,52,56] academic misconduct by students,[25,60] scientific misconduct by researchers and nefarious paper mills,[33,61] lack of critical thinking and analysis,[20,29,68] limited word count, vast and expensive resource requirements for training and deployment.[52,80] Ethical considerations which were discussed include crediting ChatGPT as an author, transparency, accountability, equity and access, environmental impact, data security, data privacy and loss of human autonomy.[25,47,60,80]

4.4.8 Recommendations for future

To improve the ChatGPT model, further research is suggested to create better algorithms and update its knowledge base to improve contextual understanding, accuracy, reliability and relevance.[6,20,25-29,32,34,43,49,53,61] Improvements in training input data validated by experts along with frequent updates is expected to solve the problem of missing information, mis-information and hallucination.[24-27,31,38,41-43,59] Enhancements in prompt engineering, language understanding, personalization, and multilingual capabilities are also anticipated to improve the outcome.[36,59]

Antaki et al recommend multimodal integration with other technologies.[30] Continual learning with healthcare domain specific training data is suggested to improve contextual understanding and critical thinking ability of the chatbot.[23,30,35,37,40,51,52,57] Collaboration with human experts is envisioned for better contextual understanding and empathetic response.[49] Exploring the potential integration of AI chatbots with other digital health interventions may improve fidelity of AI response and optimize patient care and outcomes.[37,59] Bosbach et al recommend that appropriate research may be taken up to increase the productivity and efficiency of the bot for mass manipulation.[39]

The need for research to develop bias mitigation measures and promote fairness is imperative.[27,30,57] Webb and Webb recommend training the model on communication theories to improve user experience.[59] To facilitate the use of AI-generated content for education while reducing the risk of academic dishonesty, enhanced plagiarism detector software with incorporated watermarks that can recognise AI generated content will be advantageous. Simultaneously, medical curriculum must be updated and adapted to incorporate such AI models.[25,28,33,38] Li et al. contend that the medical academic committee must update its conception of evaluations and assessments in order to remain relevant in an era of instantly accessible information.[26] Kung et al. also advocated looking into generative AI's potential to build unique assessment and evaluation methodologies.[6] With the development of technology, AI chatbots are expected to play a larger role in educational settings, giving students the resources they need to achieve their academic objectives and succeed in their chosen fields while also allowing teachers to play a more facilitative role. Better teacher and student training must be envisaged to enable this transformation.[25,38]

Updated policy guidelines and ethical safeguards are strongly recommended.[37,55,61,74] Setting up ethical education and research boards along with guiding principles for authors, researchers, reviewers and publishers should help mitigate the threat of unfair practices. Authorship, ethical and disclosure guidelines should be updated and modified to include use of AI based tools.[33,48,74] Health care professionals, academics, researchers, and publishers as well as general public need to be made aware of the benefits and limitations of ChatGPT and other similar AI tools.[43,45,55] Standardizing human-AI interactions in health care will also require the development of an open, democratic, and environmentally responsible science research infrastructure with freely accessible resources.[6]

4.5. Certainty of evidence

Table II depicts the summary of qualitative findings based on Grade-CERQual approach.

6. Discussion

Touted as a disruptive innovation, ChatGPT has rapidly caught the attention of healthcare workers and general public alike. Its ability to dynamically interact with the users and continuously evolve is ushering in a paradigm shift in the way medicine is practiced, taught and sought. Through this systematic review, we attempted to synthesize current scientific literature on ChatGPT to shed light on trends, opportunities and challenges it presents along-with implications for the future.

The results of this review highlight the global and widespread use of ChatGPT in the healthcare sector at a break neck speed. Researchers, practitioners and academicians world-wide have shown a keen interest in exploring the possibilities of ChatGPT in healthcare. A flood of level 1 publications (editorials, letters to the editor, short communications, opinions, viewpoints, perspective articles, reviews, etc) has created an info emic on ChatGPT use which simultaneously thrills and worries researchers. Most of them acknowledge and embrace integration of ChatGPT into the healthcare sector but are also cautious about the challenges it poses and ethical concerns it raises.

It is evident that ChatGPT has found applications in various aspects of healthcare, including clinical practice, research, patient communication, education and scientific writing. ChatGPT is able to perform several intricate tasks relevant to handling complex medical and clinical information encompassing a wide array of medical fields including gynaecology, cardiology, general surgery, hepatology, neurology, gastroenterology, paediatrics, dermatology, radiology, ophthalmology, general medicine, radiology, psychiatry, rheumatology, plastic surgery, emergency medicine, dentistry, pharmacology and nursing.

Nevertheless, it lacks the ability to read or generate visual content, such as images, videos, or graphs, limiting its applicability in multimedia content creation and visual communication tasks. The model tends to propagate in-built biases which tarnishes the credibility and veracity of its responses. As the model has been trained on non-medical data limited to 2021, it lacks in-depth contextual understanding of technical medical terminology and struggles with higher order reasoning and analytical thinking. It cannot be integrated into other systems and may be suitable for individual laboratory test interpretation but not for providing an overall diagnostic picture. It currently cannot replace critical thinking and creativity, which are essential attributes for healthcare professionals. Ethical concerns like plagiarism, its role as an author, lack of transparency and accountability, data privacy and data governance issues are some of the challenges that need to be addressed. Safeguards are necessary to prevent inaccurate and biased responses to health-related queries from language models like ChatGPT.

5.1. Limitations of the Evidence included in the review

Most studies were of a descriptive nature which excluded the possibility of meta-analysis and formal synthesis of evidence. Although cross-sectional studies had a low risk of bias, they had some methodological limitations like sample size, unclear statistical methods, selection bias and lack of adequate discussion. A few studies also failed to mention their limitations and were unclear on ethical clearance and disclosure of conflict of interest. The qualitative evidence on utility of ChatGPT in scientific writing had a low or very low confidence level while its role in clinical decision support is backed by low to moderate confidence levels. Application of ChatGPT in research and education is also supported by low to very low level of evidence whereas moderate confidence can be placed in ChatGPT acting as a virtual consultant for general population. Research on AI and LLM is new and does not fit into the traditional hierarchical model of research. This makes developed checklists and models for assessing risk of bias and evidence quality insufficient. In view of this deficiency, there is a need to modify the traditional system of research for defining and assessing the quality of research conducted on AI.

5.2. Strengths and limitations of this review

This systematic review improved upon previous systematic reviews exploring the utility ChatGPT in healthcare.[81,82] In those studies, search was limited to only two data bases, PubMed/MEDLINE and Google Scholar with time and language constraints. The inclusion of pre-prints that are not peer-reviewed along with date and language filters introduces inherent bias, loss of data and limits the generalisability of results. Our search was more extensive and had no restrictions. Also, even though they are described as systematic reviews, the PRISMA checklist was not followed and no formal synthesis of evidence was offered, whereas, we have attempted qualitative and quantitative evidence synthesis in addition to a narrative synthesis of evidence.

Nevertheless, the results of the current review should be viewed carefully in light of the following limitations (1) Included records were highly variable qualitatively, thus, eliminating the possibility of a rigorous data synthesis (2) the exclusion of several records that could not be accessed could have resulted in missing relevant data (3) the rapidly growing body of literature addressing ChatGPT applications/risks make imperative the need for further studies and reviews considering that the search in this review was concluded in June 2023 (4) this review focussed solely on ChatGPT which is but a subset of a larger family of LLMs which find applications in the healthcare sector (5) advanced versions of ChatGPT were launched during the course of this study that may limit the generalisability of the results to future models of the application. (6) The included studies could not meet the criteria set by most contemporary appraisal tools and

therefore, a standardized, rigorous data synthesis was not possible.

7. Conclusion

The widespread use of Chat GPT and other LLMs in health care is imminent necessitating revised policy and ethical standards and regulations to prevent misuse of generative AI. This technology harbours the potential to transform healthcare delivery, enhance accessibility, and respond to escalating demands for conversational AI, but extensive real-world testing with updated models would be required before its widespread implementation. As newer sets of evidence on Chat GPT and other AI based tools are being generated, there exists an opportunity to improve research practices to enable standardized appraisal and synthesis of such evidence to promote translation of technological innovation into impactful health care reforms.

6.1. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to create **Figure 1** and for language and editing support. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

8. Funding Sources

This work did not receive any grants from funding agencies in the public, commercial, or not-for-profit sectors.

9. Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
2. Introducing ChatGPT. Available from; 2025. https://openai.com/index/chatgpt/
3. Sahoo S, Kumar S, Abedin MZ, Lim WM, Jakhar SK. Deep learning applications in manufacturing operations: A review of trends and ways forward. *J Enterp Inf Manag*. 2022;36(1):221-51.
4. Biswas S. ChatGPT and the Future of Medical Writing., Radiology, 2023;307(2):e223312.
5. Shen Y, Heacock L, Elias J. ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology*. 2023;307(2): e230163.
6. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198.
7. O'Connor S, ChatGPT. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Educ Pract*. 2023;66(11):103537.
8. Gpt Generative Pretrained Transformer, Almira Osmanovic Thunström, Steinn Steingrimsson: Can GPT-3 write an academic paper on itself, with minimal human input?. (hal-03701250) p.1-8.
9. ChatGPT Generative Pre-trained Transformer, Zhavoronkov A. Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective. Oncoscience. 2022:9:82-4.
10. Else H. Abstracts written by ChatGPT fool scientists. *Nature*. 2023;613(7944):423.
11. Friedman C, Hripcsak G. Natural language processing and its future in medicine. *Acad Med J Assoc Am Med Coll*. 1999;74(8):890-5.
12. Wang J, Deng H, Liu B. Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years: Bibliometric Study on PubMed. *J Med Internet Res*. 2020;22(1): e16816.
13. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform*. 2006;39(6):589-99.
14. Page MJ, McKenzie JE, Bossuyt PM. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;371-2.
15. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). *BMJ Open*. 2016;6(12):e011458.
16. VOSviewer - Visualizing scientific landscapes. 2025. Available from: https://www.vosviewer.com/
17. Lewin S, Booth A, Glenton C. Applying Grade-Cer Qual to qualitative evidence synthesis findings: Introduction to the series. *Implement Sci*. 2018;13(1):1-10.
18. Rayyan: AI-Powered Systematic Review Management Platform. 2025. Available from: https://www.rayyan.ai/
19. Corsello A, Santangelo A. May Artificial Intelligence Influence Future Pediatric Research? - The Case of ChatGPT. *Children*. 2023;10(4):757
20. Almazyad M, Aljofan F, Abouammoh NA. Enhancing Expert Panel Discussions in Pediatric Palliative Care: Innovative Scenario Development and Summarization With ChatGPT-4. *Cureus*. 2023:15(4): e38249
21. Shahriar S, Hayawi K. Let's Have a Chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. *Artif Intell Appl*. Published online 2023;2(1):11-20.
22. Ciaccio EJ. Use of artificial intelligence in scientific paper writing. *Inform Med Unlocked*. 2023;41:101253.
23. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. *Radiology*. 2023;307(5):e230582.
24. Ghosh A, Bir A. Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry. *Cureus*. 2023;15(4):e37023.
25. Giannos P, Delardas O. Performance of ChatGPT on UK Standardized Admission Tests: Insights From the BMAT, TMUA, LNAT, and TSA Examinations. *JMIR Med Educ*. 2023;9:e47737
26. Li SW, Kemp MW, Logan SJS. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am J Obstet Gynecol*. *Am J Obstet Gynecol*, 2023;229(2):172.
27. Thirunavukarasu AJ, Hassan R, Mahmood S. Trialling a Large Language Model (ChatGPT) in General Practice With the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care. *JMIR Med Educ*. 2023;9:e46599
28. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof*. 2023;20:1:1-6.
29. Skalidis I, Cagnina A, Luangphiphat W. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? *Eur Heart J Digit Health*. 2023;4(3):279-81.
30. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings. *Ophthalmol Sci*. 2023:100324.
31. Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence

Hallucination in Scientific Writing Through ChatGPT References. *Cureus*. 2023:15(4):e37432.
32. Bulck LV, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. *Eur J Cardiovasc Nurs*. 2023;23(1):95-8.
33. Gao CA, Howard FM, Markov NS. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *Npj Digit Med*. 2023;6(1):75.
34. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The Capability of ChatGPT in Predicting and Explaining Common Drug-Drug Interactions. *Cureus*. 2023;15(3):e36272.
35. Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT Answers Common Patient Questions About Colonoscopy. *Gastroenterology*. 2023;165(2):509-11.
36. Lyu Q, Tan J, Zapadka ME. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art*. 2023;6(1):1-9.
37. 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.
38. Sinha RK, Roy AD, Kumar N, Mondal H. Applicability of ChatGPT in Assisting to Solve Higher Order Problems in Pathology. *Cureus*. 2023;15(2):e35237
39. Bosbach WA, Senge JF, Nemeth B. Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier. *Curr Probl Diagn Radiol*. 2023.53(1):102-10.
40. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertil Steril*. 2023;120(3):575-83.
41. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr*. 2023;7(2):15.
42. Lahat A, Shachar E, Avidan B, Shatz Z, Glicksberg BS, Klang E. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Sci Rep*. 2023;13(1):4164.
43. Balas M, Ing EB. Conversational AI Models for ophthalmic diagnosis: Comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator. *JFO Open Ophthalmol*. 2023;1:100005.
44. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc*. 2023;30(7):1237-45.
45. Brown C, Nazeer R, Gibbs A, Page PL, Mitchell AR. Breaking Bias: The Role of Artificial Intelligence in Improving Clinical Decision-Making. *Cureus*. 2023;15(3):e36415.
46. Ma Y. The potential application of ChatGPT in gastrointestinal pathology. *Gastroenterol Endosc*. 2023;1(3):130-1.
47. Kim HY. A Case Report on Ground-Level Alternobaric Vertigo Due to Eustachian Tube Dysfunction With the Assistance of Conversational Generative Pre-trained Transformer (ChatGPT). *Cureus*. 2023;15(3): e36830.
48. Sonntagbauer M, Haar M, Kluge S. Artificial intelligence: How will ChatGPT and other AI applications change our everyday medical practice? *Med Klin - Intensivmed Notfallmedizin*. 2023;118(5):366-71.
49. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic Surgery Advice and Counseling from Artificial Intelligence: A Rhinoplasty Consultation with ChatGPT. *Aesthetic Plast Surg*. 2023;47(5):1985-93.
50. Zhou Z. Evaluation of ChatGPT's Capabilities in Medical Report Generation. *Cureus*. 2023;15(4):e37589
51. Cadamuro J, Cabitza F, Debeljak Z. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). *Clin Chem Lab Med*. 2023;61(7):1158-66.
52. Altmäe S, Sola-Leyva A, Salumets A. Artificial intelligence in scientific writing: a friend or a foe? *Reprod Biomed Online*. 2023;47(1):3-9.
53. Samaan JS, Yeo YH, Rajeev N. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. *Obes Surg*. 2023;33(6):1790-6.
54. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. *Hepatol Commun*. 2023;7(4):e0097.
55. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr*. 2023;7(2):1-10.
56. Johnson D, Goodman R, Patrinely J. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res Sq*. 2023;3:2566942.
57. Sallam M, Salim NA, Al-Tammemi AB. ChatGPT Output Regarding Compulsory Vaccination and COVID-19 Vaccine Conspiracy: A Descriptive Study at the Outset of a Paradigm Shift in Online Search for Information. *Cureus*. 2023;15(2): e35029.
58. Zhou Z, Wang X, Li X, Liao L. Is ChatGPT an Evidence-based Doctor? *Eur Urol*. Published online 2023;84(4):355-6.
59. Webb JJ, Webb JJ. Proof of Concept: Using ChatGPT to Teach Emergency Physicians How to Break Bad News. *Cureus*. 2023;15(5): e38755.
60. Amri MM, Hisan UK. Incorporating AI Tools into Medical Education: Harnessing the Benefits of ChatGPT and Dall-E. *J Nov Eng Sci Technol*. 2023;2(2):34-9.
61. Hisan UK, Amri MM. ChatGPT and Medical Education: A Double-Edged Sword. *J Pedagogy Educ Sci*. 2023;2(01):71-89.
62. Yeo YH, Yang JD. Correspondence on Letter regarding "Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma". *Clin Mol Hepatol*, 2024;30(1):124–5.
63. Mijwil MM, Hiran KK, Doshi R, Dadhich M., ChatGPT and the Future of Academic Integrity in the Artificial Intelligence Era: A New Frontier. *Al-Salam J Eng Technol*. 2023:2(2):116-27.
64. Aydın Ö, Karaarslan E. OpenAI ChatGPT Generated Literature Review: Digital Twin inHealthcare. In Ö. Aydın (Ed.), Emerging Computer Technologies. İzmir Akademi Dernegi*(PDF) OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare*. 2022;p.22-31.
65. Temsah O, Khan SA, Chaiah Y. Overview of Early ChatGPT's Presence in Medical Literature: Insights From a Hybrid Literature Review by ChatGPT and Human Experts. Cureus. 2023;15(4):e37281
66. Vallath AL, Sivasubramanian BP, Chatterjee A, Erva S, Ravikumar DB, Dasgupta I. Ventricular Septal Rupture and Artificial Intelligence (AI)-Assisted Healthcare. *Cureus*. 2023;15(3):1-12.
67. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? *Diabetes Metab Syndr Clin Res Rev*.2023:245;102744.
68. Raxwal B, Baisla P, Nath J. A Collaborative Case Report Utilizing ChatGPT AI Technology of Traumatic Right Coronary Artery Dissection Resulting in Inferior Wall ST-Elevation Myocardial Infarction. *Cureus*. 2023;15(3):e35894.
69. Hegde A, Srinivasan S, Menon G. Extraventricular Neurocytoma of the Posterior Fossa: A Case Report Written by ChatGPT. *Cureus*.2023;15(3):e35894.
70. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. 2023;15(2):e35179.
71. Zamarud A, Marianayagam N, Sekar V, Chang SD, Meola A. Treatment Outcomes of Leiomyosarcoma Metastasis Affecting the Brachial Plexus: A Comparative Case Report Using Chat Generative Pre-trained Transformer (ChatGPT). *Cureus*. 2023;15(3):e36715.

72. Schuppe K, Burke S, Cohoe B, Chang K, Lance RS, Mroch H. Atypical Nelson Syndrome Following Right Partial and Left Total Nephrectomy With Incidental Bilateral Total Adrenalectomy of Renal Cell Carcinoma: A Chat Generative Pre-Trained Transformer (ChatGPT)-Assisted Case Report and Literature Review. *Cureus*. 2023;15(3): e36042.
73. Chu ECP, Huang KHK, Cheung G, Ng G, Lin A. Delayed Skin Allergy to Glucosamine Chondroitin Supplement. *Cureus*. 2023;15(3):e36310.
74. Babl FE, Babl MP. Generative artificial intelligence: Can ChatGPT write a quality abstract?. *EMA - Emerg Med Australas*. *Emerg Med Aust.* 2023;35(5):809-11.
75. Donker T. The dangers of using large language models for peer review. *Lancet Infect Dis*. 2023;23(7):781.
76. Hill-Yardin EL, Hutchinson MR, Laycock R, Spencer SJ. A Chat (GPT) about the future of scientific publishing. *Brain Behav Immun*. 2023;110:152-4.
77. Xames MD, Shefa J. ChatGPT for research and publication: Opportunities and challenges. *J Appl Learn Teach*. 2023;6(1):390-5.
78. Hosseini M, Horbach SPJM. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res Integr Peer Rev*. 2023;8(1):1-9.
79. Johnson TR, Bernstam EV. Why is biomedical informatics hard? A fundamental framework. *J Biomed Inform*. 2023:140:104327.
80. Kim SG. Using ChatGPT for language editing in scientific articles. *Maxillofac Plast Reconstr Surg*. 2023;45(1):13.
81. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthc Basel Switz*. 2023;11(6):887.
82. Muftić F, Kadunić M, Mušinbegović A, Almisreb AA. *Southeast Europe J Soft Comp Explor Med Breakthroughs.* 2023:12(1):13-41

Cite this article: Mishra G, Verma S, Chitra P. Trends and directions of chatgpt in healthcare- A systematic review. *J Contemp Orthod* 2025;9(2):153-176.