Application of ChatGPT as a content generation tool in continuing medical education: acne as a test topic

Luigi Naldi,[1,2] Vincenzo Bettoli,[2,3] Eugenio Santoro,[4] Maria Rosa Valetto,[5] Anna Bolzon,[1,6] Fortunato Cassalia,[1,6] Simone Cazzaniga,[2,7] Sergio Cima,[5] Andrea Danese,[8] Silvia Emendi,[5] Monica Ponzano,[6] Nicoletta Scarpa,[5] Pietro Dri[5]

[1]Dermatology Unit, Ospedale San Bortolo, Vicenza, Italy; [2]Centre of the Italian Group for Epidemiological Research in Dermatology, Bergamo, Italy; [3]Section of Dermatology and Infectious Diseases, Department of Medical Sciences, University of Ferrara, Italy; [4]Unit of Research in Digital Health and Digital Therapeutics, Department of Clinical Oncology, Mario Negri Institute for Pharmacological Research, Milan, Italy; [5]Zadig ltd Benefit Company, CME National Provider, Milan, Italy; [6]Unit of Dermatology, Department of Medicine, University of Padua, Italy; [7]Department of Dermatology, Inselspital University Hospital of Bern, Switzerland; [8]Unit of Dermatology, Department of Integrated Medical and General Activity, University of Verona, Italy

Correspondence: Luigi Naldi, Dermatology Unit, Ospedale San Bortolo, Vicenza, Italy; Study Centre of the Italian Group for Epidemiological Research in Dermatology (GISED), Bergamo, Italy.
E-mail: luiginaldibg@gmail.com

Key words: acne; artificial intelligence; ChatGPT; medical information; medical education; large language models.

Contributions: PD, LN, ES, MRV, study design; SCi, SE, NS, query sessions and data collection; AB, VB, FC, AD, MP, analysis of contents; SE, NS, analysis of reproducibility; SCa, statistical analysis and preparation of figures and tables; VB, LN, ES, MRV, manuscript drafting; PD, ES, critical revision of the manuscript; PD, supervision of each phase of manuscript preparation. All the authors have read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors declare that they have no competing interests.

Ethics approval and consent to participate: ethics approval is not applicable due to the methodological nature of the study and because it did not involve human or animal subjects, nor imply access to identifiable personal information.

Availability of data and materials: data supporting the findings of this study are available in the supplementary material of this article or from the corresponding author upon reasonable request.

Received: 10 September 2024.
Accepted: 5 November 2024.

Dermatology Reports 2025; 17:10138
doi:10.4081/dr.2024.10138

Abstract

The large language model (LLM) ChatGPT can answer open-ended and complex questions, but its accuracy in providing reliable medical information requires a careful assessment. As part of the AI-CHECK (Artificial Intelligence for CME Health E-learning Contents and Knowledge) study, aimed at evaluating the potential of ChatGPT in continuous medical education (CME), we compared ChatGPT-generated educational content to the recommendations of the National Institute for Health and Care Excellence (NICE) guidelines on acne vulgaris. ChatGPT version 4 was exposed to a 23-item questionnaire developed by an experienced dermatologist. A panel of five dermatologists rated the answers positively in terms of "quality" (87.8%), "readability" (94.8%), "accuracy" (75.7%), "thoroughness" (85.2%), and "consistency" with guidelines (76.8%). The references provided by ChatGPT obtained positive ratings for "pertinence" (94.6%), "relevance" (91.2%), and "update" (62.3%). The internal reproducibility was adequate both for answers (93.5%) and references (67.4%). Answers related to issues of uncertainty and/or controversy in the scientific community scored the lowest. This study underscores the need to develop rigorous evaluation criteria for AI-generated medical content and for expert oversight to ensure accuracy and guideline adherence.

Introduction

Developed by OpenAI, ChatGPT[1] is an advanced large language model (LLM) with numerous potential applications in healthcare information and education for both professionals and patients. Several benefits of ChatGPT have been envisaged. These include enhancing scientific writing, promoting equity and versatility in research, supporting medical research through efficient data analysis and reviews, improving healthcare practices, and advancing healthcare education and learning.[2-7] Drawbacks have also been pointed out for medical applications, including a lack of consideration of all the determinants that influence medical advice with ethical implications if patients experience harm.[3,4,8,9]

In medical education, ChatGPT demonstrates potential in several important areas. It can facilitate the development of academic and postgraduate training content, generate assessment questions to evaluate knowledge and skills, create interactive simulated clinical scenarios to enhance decision-making skills, support medical-patient communication through realistic dialogue generation, and aid in the development of interactive educational resources.[5-7]

ChatGPT's performance in terms of consistency, accuracy,

relevance, and reliability has been evaluated in a variety of clinical areas, obtaining non-univocal results.[4,7,10-15]

The AI-CHECK (Artificial Intelligence for CME Health E-learning Contents and Knowledge) study, focusing on acne, has been designed in three steps to explore the potential use of ChatGPT in continuous medical education. In the first step, we explored the strengths and limitations of ChatGPT in providing information on acne to the general population.[16] Here, we present the second step of the project, which aims to evaluate the materials produced by ChatGPT for a continuing medical education (CME) course targeting general practitioners and to compare them with the recommendations of the recent National Institute for Health and Care Excellence (NICE) guidelines on acne and pertinent bibliographic references.[17]

Materials and Methods

Choice of the topic

Acne vulgaris (hereinafter acne) has been chosen as a topic since it is a common condition that affects 9.4% of people globally, with management criteria that have not changed significantly in recent years.[17-20] Thus, this choice could overcome a possible updating bias when comparing information produced by ChatGPT.

ChatGPT interaction

ChatGPT version 4 (released on March 14, 2023) was used for data acquisition. For the study conducted from September 19 to 21, 2023, the version updated to September 2021 was used. No plugins were used to enable ChatGPT to browse the internet, ensuring that all responses generated were based solely on internal knowledge up to the training cutoff date, without access to updated information from the web. All activities were conducted in English, and all data were recorded and archived.

Assessment of agreement with guidelines

The information provided by ChatGPT on acne management was evaluated by comparison to the NICE guidelines "Acne vulgaris: management"[17] using a 23-item questionnaire developed by an experienced dermatologist (LN). The questionnaire addressed the main issues in managing acne, considering the most common questions posed by users in acne forums and also how acne management is typically presented in textbooks. Each question was assigned a score (correction factor) from 1 to 3, weighing the relevance of the question (1 for the lowest relevance, 3 for the highest relevance) based on both the strength of the available scientific evidence and the practical relevance for management. Furthermore, the expert matched the questions with the guidelines' recommendations (*Supplementary Table 1*).

The 23 questions were prompted three times by independent operators (NS, SE, SCi), recorded, and archived. The first set of answers provided by ChatGPT was independently evaluated by a panel of 5 dermatologists, including four residents (AB, FC, AD, MP) and one experienced dermatologist with a research focus on acne (VB), using a dedicated online spreadsheet. The answers were scored according to 5 domains: "quality", "readability", "accuracy", "thoroughness", and "consistency with guidelines" (the latter when applicable) using a 5-point Likert scale (from 1 "very poor" to 5 "very good").

In addition, the evaluators were allowed to enter a qualitative judgment for all answers to the questionnaire (*Supplementary Table 2*).

Assessment of internal reproducibility of contents

To assess ChatGPT's internal reproducibility (*i.e.,* the ability to consistently reproduce its answers under the same conditions), three independent operators (NS, SE, SCi) prompted the 23 questions three times in separate sessions. All the answers were recorded and archived for content comparison.

Two operators (NS, SE) independently evaluated the three sets of answers. Taking the first query session as the standard, they qualitatively judged the subsequent two sessions as having "complete overlap", "partial overlap," or "no overlap" of contents.

Assessment of references

During all three query sessions and following each question prompt, ChatGPT was asked to quote three references from the biomedical literature to support the answers provided.

To identify AI hallucinations (*i.e.,* wrong or out-of-context answers), each reference suggested during the first query session was verified based on the correctness of the quotation (authors, title, journal name, year of publication, issue, and pages) by comparison to PubMed database. After excluding AI hallucinations, the references provided during the first query session were evaluated by the panel of 5 dermatologists using a dedicated online spreadsheet. The answers were scored according to three criteria: "pertinence", "relevance", and "update", with a binary judgment ("Yes" or "No"). In addition, the evaluators were allowed to enter a qualitative judgment for each reference (*Supplementary Table 3*).

Assessment of internal reproducibility of references

The three sets of references were independently evaluated by two authors (NS, SE), assuming the first query session as the standard, and judged the subsequent two sessions as "complete overlap" (CO), "partial overlap" (PO), or "no overlap" (NO) of references.

Recording of unexpected or unpredictable events

Throughout all query sessions, query errors and data flow disruptions were recorded and documented.

Similarly, throughout all evaluation sessions, AI hallucinations were recorded and documented.

Statistical analysis

For descriptive purposes, median values and interquartile ranges (IQR) of evaluators' judgments of the answers were calculated. Total scores were presented as both crude and weighted values. Frequencies and percentages were calculated for positive evaluations ("Yes") related to the quality of references, reproducibility of questionnaire responses and cited sources, and the overall categorical classification of total scores. The inter-reviewer agreement (*i.e.,* the concordance between different dermatologists when evaluating the same set of answers) was measured using Gwet's AC2 with quadratic weights for scores assessment on an ordinal scale, and AC1 was used for reference judgment on a dichotomous scale and reported along with its 95% confidence interval (CI). Gwet's AC statistics were chosen because they provide more reliable agreement estimates than standard kappa statistics, particularly in cases of uneven category distributions. The interpretation of AC1-2 is similar to kappa and can be read as follows: <0.20 poor, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 good, 0.81-1.00 very good agreement. Statistical analysis was conducted with R software (version 4.1.1; R Project for Statistical Computing).

Results

Assessment of answers

Findings are reported in *Supplementary Table 1* and Figures 1 and 2.

The 23 answers from ChatGPT on acne obtained a total of 468 positive ratings out of the 555 available (84.3%). Pooling the negative ("poor" *plus* "very poor"), neutral ("acceptable"), or positive ("good" *plus* "very good") evaluators' judgments, the answers obtained 101/115 (87.8%) positive ratings for "quality", 109/115 (94.8%) positive ratings for "readability", 87/115 (75.7%) positive ratings for "accuracy", 98/115 (85.2%) positive ratings for "thoroughness", and 73/95 (76.8%) positive ratings for "consistency" (Figure 1).

Considering the single answers, median values below 4 were obtained from the answer to question 2 ("Can diet influence the appearance and severity of acne?") for "accuracy", "thoroughness", and "consistency" and from the answer to question 8 ("Are there physical acne therapies? If so, how should they be included in the therapeutic program?") for "accuracy" and "consistency" (Figure 2). The total inter-reviewer agreement was 0.82 (95% CI: 0.79-0.85). Within specific domains, it was 0.84 (95% CI: 0.79-0.89) for "quality", 0.90 (95% CI: 0.86-0.94) for "readability", 0.75 (95% CI: 0.68-0.82) for "accuracy", 0.82 (95% CI: 0.76-0.88) for "thoroughness", and 0.78 (95% CI: 0.68-0.87) for "consistency" (*Supplementary Table 1*).

Assessment of references

Findings are reported in *Supplementary Table 4* and Figure 3.

Based on the evaluators' judgments, the 69 references provided by ChatGPT obtained a total of 645 positive ratings out of the 780 total judgments (82.7%). As for the domains explored, the references obtained 246/260 (94.6%) positive ratings for "pertinence", 237/260 (91.2%) positive ratings for "relevance", and 162/260 (62.3%) positive ratings for "update" (*Supplementary Table 4*).

The total inter-reviewer agreement was 0.67 (95% CI: 0.59-0.75). Within specific domains, it was 0.89 (95% CI: 0.83-0.96) for "pertinence", 0.82 (95% CI: 0.73-0.91) for "relevance", and 0.14 (95% CI: 0.0-0.28) for "update".

Figure 1. Stacked bar chart of overall evaluators' judgments of questionnaire answers for each domain.

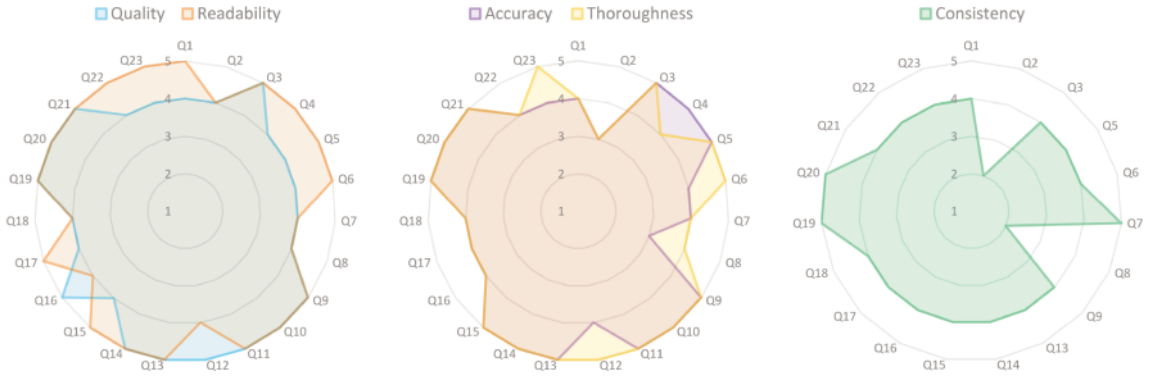

Figure 2. Radar chart of median evaluators' judgments of questionnaire answers for each domain. Questions not assessable due to the lack of or limited discussion in guidelines were removed from the domain "consistency".

Considering the single references, only those provided to question 8 ("Are there physical acne therapies? If so, how should they be included in the therapeutic program?") scored below 80% for "pertinence" and "relevance". Most references scored below 80% for "update".

Only one reference appeared more than 3 times (*Supplementary Table 5*).

Internal reproducibility

The internal reproducibility of the answers was judged adequate (CO+PO) in 43/46 (93.5%) comparisons.

The internal reproducibility of the references was judged adequate (CO+PO) in 31/46 (67.4%) comparisons (*Supplementary Table 6*).

Unexpected or unpredictable events

A data flow disruption was recorded. No query error was recorded (*Supplementary Table 7*).

Seventeen AI hallucinations were recorded, all related to citing references (*Supplementary Table 4*) with errors in quoting authors, titles, journals, year of publication, numbers, or pages, or a combination of these.

Discussion

The information provided by ChatGPT for the implementation of a CME course on acne targeting general practitioners was evaluated by comparison with the NICE guidelines "Acne vulgaris: management"[17] using 23 answers generated by ChatGPT. The GPT-4-based ChatGPT demonstrates potential as a resource for professional dermatology CME, producing appropriate responses in terms of quality (87.8%) and thoroughness (85.2%), with very high readability (94.8%). However, the responses were sometimes inaccurate or inconsistent with NICE guidelines, indicating areas for improvement. For instance, ChatGPT did not mention the dose dependency of isotretinoin's cutaneous side effects (Q6) and incorrectly stated that prolonged UV exposure induces overproduction of sebum (Q11). Additionally, while it repeatedly cited the American Guidelines on acne therapy,[20] failed to cite the European Guidelines for the treatment of acne,[18] published in the same year. Certain questions, such as the role of diet in influencing the appearance and severity of acne and the inclusion of physical acne therapies in therapeutic programs, were not answered precisely, reflecting ongoing debates and a lack of evidence in the dermatology community. This suggests a default bias towards providing answers rather than acknowledging the absence of a definitive response, a flaw that could potentially spread health misinformation.

The total inter-reviewer agreement was high (0.82), with higher concordance within the domains of quality (0.84), readability (0.90), and thoroughness (0.82), indicating acceptable agreement among the evaluators. The references provided by ChatGPT were positively accepted (82.7%), especially in terms of pertinence (94.6%) and relevance (91.2%). However, issues with the currency of references suggest gaps in the availability of papers on which ChatGPT is trained. The internal reproducibility of the answers and references was judged adequate. The importance of adopting rigorous evaluation criteria for health responses provided by LLMs is crucial to ensure safe and effective use in healthcare contexts. However, no validated and unified evaluation criteria and metrics for LLMs are currently available. There is a need to develop and implement comprehensive metrics specifically designed to evaluate their performance, covering aspects such as accuracy and reliability. In this evolving scenario lacking adequate evaluation metrics, our study has adopted a robust set of criteria capable of exploring the reliability of content. The evaluation process relied on the consensus and independent judgment of several experts and on the Comparing our results with those of previous studies on acne or other skin diseases,[11,21-24] we found better accuracy in ChatGPT's responses, likely due to more precise questions and prompts. While ChatGPT is a useful tool for generating content in the continuing medical education setting, human expert scrutiny remains essential to identify incomplete or inconsistent information. Moreover, as a part of the AI-CHECK study, we have previously assessed[16] the accuracy and completeness of ChatGPT's answers to questions about acne commonly posed by the public. ChatGPT answers were evaluated using a modified version of the Ensuring Quality Information for Patients (EQIP) tool,[25] a validated 36-item method for evaluating online written health information. Despite the overall positive performance, the study identified several inaccuracies and errors in ChatGPT responses, including incomplete or inaccurate data on treatment side effects and disease management and mistakes in terminology. These findings emphasized a significant risk in depending solely on artificial intelligence for medical information available to the general public, highlighting the necessity for expert review to prevent the spread of misinformation.

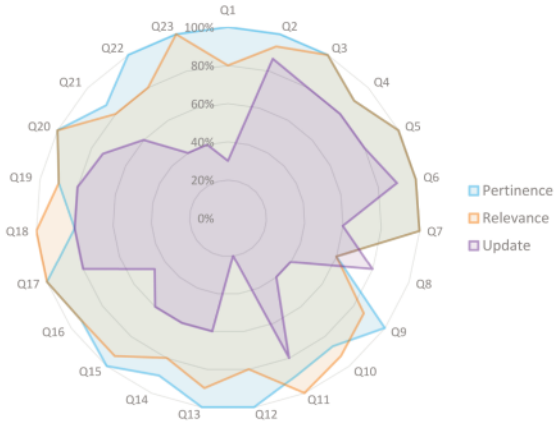

Figure 3. Radar chart of overall positive evaluators' judgments of questionnaire references provided for each answer and for each domain investigated.

This study has some limitations that need to be overcome in the next steps of the AI-CHECK Study. A research question that remains to be answered is how to compare the content provided by ChatGPT with that developed by human experts for CME. Moreover, the impact of the content produced by ChatGPT needs to be verified in terms of its capacity to modify (improve or even worsen) the knowledge and skills of potential learners who will use it.

Conclusions

Given the current performance of ChatGPT, it is essential for dermatologists to remain involved in developing clinical and patient-facing AI tools. These AI-based medical resources should be trained with evidence-based sources. Other LLMs (MedPalm2, Meditron) are already trained with medical datasets and linked to PubMed to provide more accurate and up-to-date information. Ethical concerns specific to dermatology have recently been raised, including data security and privacy, the risk of misdiagnosis and inaccurate responses, and uncertainty about the impact of AI implementation in clinical practice.[9] These issues should be thoroughly assessed on a case-by-case basis rather than being treated as general principles.

References

1. OpenAI. ChatGPT. https://chat.openai.com/chat
2. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science 2023;381:187-92.
3. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6:120.
4. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595.
5. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023;9:e50945.
6. Eysenbach G. The role of ChatGPT, Generative Language Models, and Artificial Intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9:e46885.
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11:887.
8. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of Large Language Models in medicine. JAMA 2023;330:866-9.
9. Gordon ER, Trager MH, Kontos D, et al. Ethical considerations for artificial intelligence in dermatology: a scoping review. Br J Dermatol 2024;190:789-97.
10. Chen S, Kann BH, Foote MB, et al. Use of Artificial Intelligence chatbots for cancer treatment information. JAMA Oncol 2023;9:1459-62.
11. Ferreira AL, Chu B, Grant-Kels JM, et al. Evaluation of ChatGPT dermatology responses to common patient queries. JMIR Dermatol 2023;6:e49280.
12. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023;6:e2336483.
13. Lam Hoai XL, Simonart T. Comparing meta-analyses with ChatGPT in the evaluation of the effectiveness and tolerance of systemic therapies in moderate-to-severe plaque psoriasis. J Clin Med 2023;12:5410.
14. Rossettini G, Cook C, Palese A, et al. Pros and cons of using Artificial Intelligence chatbots for musculoskeletal rehabilitation management. J Orthop Sports Phys Ther 2023;53:1-7.
15. Temsah O, Khan SA, Chaiah Y, et al. Overview of early ChatGPT's presence in medical literature: insights from a hybrid literature review by ChatGPT and human experts. Cureus 2023;15:e37281.
16. Bettoli V, Naldi L, Santoro E, et al. ChatGPT and acne: Accuracy and reliability of the information provided-The AI-check study. J Eur Acad Dermatol Venereol 2024.
17. National Institute for Health and Care Excellence (NICE) Acne vulgaris: management. NICE guideline 198. Available from: https://www.nice.org.uk/guidance/ng198/resources/acne-vulgaris-management-pdf-66142088866501.
18. Nast A, Dréno B, Bettoli V, et al. European evidence-based (S3) guideline for the treatment of acne - update 2016 - short version. J Eur Acad Dermatol Venereol 2016;30:1261-8.
19. Reynolds RV, Yeung H, Cheng CE, et al. Guidelines of care for the management of acne vulgaris. J Am Acad Dermatol 2024;90:1006.e1-1006.e30.
20. Zaenglein AL, Pathy AL, Schlosser BJ, et al. Guidelines of care for the management of acne vulgaris. J Am Acad Dermatol 2016;74:945-73.e33.
21. Lakdawala N, Channa L, Gronbeck C, et al. Assessing the accuracy and comprehensiveness of ChatGPT in offering clinical guidance for atopic dermatitis and acne vulgaris. JMIR Dermatol 2023;6:e50409.
22. Cirone K, Akrout M, Abid L, Oakley A. Assessing the utility of multimodal Large Language Models (GPT-4 Vision and Large Language and Vision Assistant) in identifying melanoma across different skin tones. JMIR Dermatol 2024;7:e55508.
23. Reynolds K, Tejasvi T. Potential use of ChatGPT in responding to patient questions and creating patient resources. JMIR Dermatol 2024;7:e48451.
24. O'Hagan R, Poplausky D, Young JN, et al. The accuracy and appropriateness of ChatGPT responses on nonmelanoma skin cancer information using zero-shot chain of thought prompting. JMIR Dermatol 2023;6:e49889.
25. Charvet-Berard AI, Chopard P, Perneger TV. Measuring quality of patient information documents with an expanded EQIP scale. Patient Educ Couns 2008;70:407-11.

Online Supplementary Material:
Supplementary Table 1. Median values of evaluators' judgments of the questionnaire answers, in total and by specific domain.
Supplementary Table 2. Qualitative judgments of evaluators for all answers to the questionnaire.
Supplementary Table 3. Qualitative judgments of evaluators for each reference.
Supplementary Table 4. Numbers and percentages of overall positive evaluators' judgments of the questionnaire references provided for each answer, in total, and by specific domain. References identified as ChatGPT hallucinations were excluded from assessment. The numbers and percentages of hallucinations over the total references provided are also reported.
Supplementary Table 5. Recurrence of the references.
Supplementary Table 6. Internal reproducibility of the references.
Supplementary Table 7. Unexpected or unpredictable events during query sessions.