TABLE 1 Evaluations and explanations of generative AI for medical applications.

Section/Topic	Item no	Recommendation	Explanations
Title	1	Identify the report as an article related to the research that evaluates generative AI's applicability in medicine.	Provide the reader with an initial understanding of the nature of the text.
Abstract	2	State the purpose of the research, the generative AI model used and its version, the source of the questions, methods, results, and conclusions.	Lay the foundation for readers to quickly understand the study and facilitate other researchers to critically analyze the design and results of this research.
Introduction			
Justification	3	Review existing relevant information and explain the background of the study.	Enable readers to grasp the central theme of the article.
Objectives	4	State-specific objectives including the generative AI model used and its version, the training set used for generative AI, the source of the questions, the nature of research, and the limitations.	Provide the necessary framework for readers to understand the article.
Methods			
Question collection	5	Select the professional questions from guidelines, official examination question banks, and high-frequency issues found via search engines like Google, or drafted by experts, ensuring that the questions cover specific subfields of medicine.	For guidelines or question banks, questions can be either manually selected or extracted using software, while using an API to select questions can reduce subjective errors and make more sense for the entire data set. When selecting questions from search engines, researchers may opt for frequently occurring ones. If the questions are drafted by experts, the experts need to have authority and experience in the relevant field.
	6	Ensure the questions are representative in terms of difficulty, type, and professionalism.	Enhance the universality of the study.
	7	Describe how the questions were collected, the number of questions, whether the questions were pre-screened, the conditions of the screening, the modality of the input as well as the relevant format.	Input modes such as text, image, sound, video input, and so on, and related attributes (e.g., image resolution).
Agent	8	Record the model used, the version of the generative AI, and customized parameters such as temperature parameters, if applicable. State the strengths and weaknesses of the current version used and the rationale for assessing it.	The model used, and the version of the generative AI may have a significant influence on the result. Temperature is a parameter influencing text generation randomness. Higher temperatures yield more diverse and novel outputs, with increased unpredictability and potential inaccuracies. Lower temperatures produce consistent, predictable text closely aligned with training data but might lack creativity.
	9	If intend to report them as a functional series, it is recommended to report the relationship between model versions (e.g., whether it is a simple upgrade; if not, it is recommended to report the horizontal comparison results).	Reporting the relationship between model versions clarifies the evolution of the technology, helping users understand improvements or changes. Additionally, providing horizontal comparison results aids in comprehending the distinct capabilities and applications of each version.

(Continues)

TABLE 1 (Continued)

Section/Topic	Item no	Recommendation	Explanations
Questioning	10	Use a consistent prompt with identically formatted patterns and provide the full prompt in the article.	Reduce the objective differences introduced by different questioning methods and the impact of such differences on the quality of answers. Providing the full prompt in the article ensures that the study is transparent and reproducible.
	11	Ask the same question multiple times and record each response.	Generative AI, known for delivering varied responses to identical queries, necessitates repeated questioning to gauge its consistency.
	12	Indicate whether the question is open-ended or multiple-choice.	Subjective and objective questions are assessed differently.
	13	Initiate a new chat for each question.	Prevent generative AI from being affected by context.
	14	Record the data the responses were collected.	Reduce the impact of performance differences between AI versions and the timing of knowledge updates.
Accuracy	15	Describe any methods employed for scoring accuracy when dealing with subjective questions.	Accuracy refers to the degree to which the response reflects or corresponds to reality or truth.
	16	Compare with reference answers, record the number of correct responses to each question, and calculate the rate of correct answers if you asked objective questions.	The more times the generative AI model responds correctly, the more robust it is considered to be.
Integrity	17	Describe any methods used to assess the integrity between responses.	Integrity refers to whether the response is comprehensive, detailed, and covers relevant information.
Readability	18	Describe any methods used to assess the readability of responses.	Reflect on the ease with which a text can be read and understood (e.g., clarity of language, the organization of structure, and grammatical and spelling accuracy).
Reviewers	19	Clarify the composition of reviewers and the rationale for this composition, which is recommended to be more than two experts from varied fields like medicine, AI, and interdisciplinary areas, along with stakeholders from ethics, sociology, and user groups.	Ensure the fairness and effectiveness of the evaluation process.
	20	Pay special attention to assessing the impenetrability of responses.	
	21	Evaluate the consistency across responses to the same question to assess whether the generative AI can steadily provide consistent responses.	
	22	Assess the consistency and reliability of reviewer ratings, avoiding significant differences in the subjective scores among reviewers.	A way to effectively monitor model performance, helping detect if the model has erratic behavior.
Results			
Results selection	23	Describe the results of the search process, from the number of questions collected to the final results, ideally using a flow diagram.	Uncovering its performance in specific subdomains is critical to a deeper understanding of the value and limitations of AI applications in medicine.

TABLE 1 (Continued)

Section/Topic	Item no	Recommendation	Explanations
Study characteristics	24	State all studies included in the analysis and detail their characteristics.	
Results of individual studies	25	Present results for accuracy, completeness, and readability for each study, recommending the use of tables or charts for presentation.	
Results of syntheses	26	Present results of all statistical syntheses conducted and results of analyses conducted to explore possible causes of heterogeneity among study results.	
Discussion			
Interpretation	27	Analyze the results according to the study objectives.	Comprehensively analyze the performance of generative AI in terms of accuracy, completeness, and readability.
Strengths and limitations	28	Describe the advantages of the research.	To make the reader understand the importance of the study.
	29	Explore constraints of the research, acknowledging possible origins of partiality or inaccuracy.	Enhance the understanding of the scope, accuracy, and applicability of the research findings.
	30	Engage in rational discussion and reject exaggeration.	An honest and rational expression is necessary to maintain academic norms and advance knowledge.
Conclusion	31	Provide a condensed conclusion that summarizes the study's main findings, reiterates its importance, and indicates directions or recommendations for future research.	Provide direction for future research and help promote the further development and application of generative AI in the medical field.
Other information			
Funding and sponsorship	32	Provide the origin of financial support and the function of the sponsors for the current investigation, as well as for the initial research if relevant to the foundation of this article.	Maintain the objectivity and transparency of the research.

Abbreviations: AI, artificial intelligence; API, application programming interface.

recommend a flexible approach to ethical considerations, adapting to diverse contexts while firmly adhering to core principles like patient privacy and data security. This balance is key in a field where ethical challenges evolve as rapidly as technology.

Moreover, our checklist is designed to maintain a balance between methodological rigor and flexibility, crucial for the rapidly evolving field of AI in medicine. We have not provided a scoring rule, aiming to fuel the creativity of researchers. For example, one study explored the potential of large language models as tools against medical disinformation [22]. This approach encourages innovative applications and interpretations of AI technology, allowing for groundbreaking developments that might not be fully captured by a rigid scoring system. By embracing this open-ended approach, our framework fosters an environment where unconventional ideas can be tested and refined, thereby accelerating the advancement of AI in healthcare. This flexibility also permits the integration of new techniques and findings, ensuring that our guidelines remain relevant and effective as the field continues to grow and change.

Furthermore, the detailed explanation of each checklist item in our guidelines is not just about aiding comprehension. It also plays a vital role in minimizing subjective interpretation variances and bolstering the reproducibility of assessments. This guidance empowers researchers to identify and address potential challenges in their studies, which is instrumental in elevating the quality of research. By providing this comprehensive framework and guidance, we aim to pave the way for more nuanced, effective, and reliable use of generative AI in medical research. This approach is vital in ensuring that AI technologies not only advance in capability but