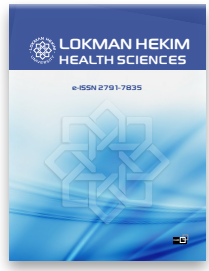
LOKMAN HEKIM HEALTH SCIENCES

DOI: 10.14744/lhhs.2026.81418
Lokman Hekim Health Sci 2026;6(2):323–333



lokmanhekimhs.com

SYSTEMATIC REVIEW

A Systematic Review and Meta-Analysis of Diagnostic Performance Comparison between DeepSeek and Physicians

Jianwen Zeng[1], Xule Zhu[1], Xin Liu[1], Shiying Shen[1], Sixie Li[1], Shihua Cao[1,2]

[1]School of Public Health and Nursing, Hangzhou Normal University, Hangzhou, China
[2]Key Engineering Research Center of Mobile Health Management System, Ministry of Education, Hangzhou, China

Abstract

Introduction: Since the release of DeepSeek, it has attracted substantial global attention and has increasingly been explored as a tool for medical diagnosis, showing promising potential for clinical applications. To comprehensively evaluate the effectiveness, potential, and limitations of DeepSeek in medical diagnosis, thereby informing future research and real-world implementation and supporting the development of AI-assisted diagnostic care.

Methods: We searched Web of Science Core Collection, Embase, MEDLINE, Scopus, IEEE Xplore, and medRxiv from inception to August 8, 2025. Two authors independently screened studies, extracted data according to predefined inclusion and exclusion criteria, and assessed study quality using the Prediction model Risk of Bias Assessment Tool.

Results: Twenty-four studies were included, evaluating 6 DeepSeek model variants; DeepSeek-R1 was the most frequently assessed. Quality appraisal indicated a high risk of bias in 13 studies (54%). DeepSeek's performance varying across medical specialties. Overall performance did not differ significantly between DeepSeek and physicians (p=0.07); however, DeepSeek did not reach physician-level performance, with diagnostic accuracy 7.7% points lower than physicians.

Discussion and Conclusion: DeepSeek demonstrated no statistically significant difference compared with physicians, yet it remained below physician performance. At present, it should not replace expert clinicians. Nevertheless, it may serve as a valuable adjunct in non-specialist settings and as an educational tool for medical trainees.

Keywords: DeepSeek; Diagnosis; Large language model; Meta-analysis; Systematic review

Large language models (LLMs) have revolutionized the field of artificial intelligence (AI) by demonstrating impressive capabilities in natural language understanding and reasoning. These models are rapidly becoming transformative tools in the medical field, showing potential across various clinical applications, including personalized health consultations, research, clinical decision support, surgical planning assistance, and telemedicine promotion.[1] Their ability to process and understand complex medical information presents opportunities for improving clinical decision-making, automating administrative tasks, and enhancing patient care.[2–4] As AI technology matures, these models are expected to become valuable aids in navigating the expanding domain of medical knowledge and improving healthcare services.

Cite this article as: *Zeng J, Zhu X, Liu X, Shen S, Li S, Cao S. A Systematic Review and Meta-Analysis of Diagnostic Performance Comparison between DeepSeek and Physicians. Lokman Hekim Health Sci 2026;6(2):323–333.*

Correspondence: Shihua Cao, M.D. School of Public Health and Nursing, Hangzhou Normal University, Hangzhou, China
E-mail: csh@hznu.edu.cn **Submitted:** 08.05.2026 **Revised:** 15.05.2026 **Accepted:** 16.05.2026 **Available Online:** 11.06.2026

The application of LLMs in medicine has led to growing research attention toward their diagnostic capabilities. Studies have extensively explored these models' performance in interpreting clinical data, understanding patient histories, and even suggesting potential diagnoses.[5,6] Medical diagnosis is a challenging task, but with comprehensive medical knowledge, these models serve as diagnostic support tools through natural language interaction,[7] showing promising potential even for diagnosing complex clinical cases.[8] LLMs' accuracy, speed, and efficiency in processing vast amounts of medical literature and patient information underscore their value as tools in medical diagnostics.

In late 2024, DeepSeek introduced two open-source models, DeepSeek-V3 and DeepSeek-R1,[9,10] which quickly garnered global attention. These models slightly outperform generative pre-trained transformer (GPT)-4o and GPT-o1 in performance while reducing computational costs by an order of magnitude.[11] Furthermore, the open nature of the DeepSeek models has fostered the development of a collaborative ecosystem, enabling researchers and developers worldwide to experiment with, refine, and adapt these models for various applications. This collective effort has accelerated DeepSeek's adoption in the healthcare sector.[12] The scalable DeepSeek-R1 architecture (1.5B–671B parameters) has driven the development of medical LLMs,[13–15] prompting investigations into whether the model can compete with proprietary models in clinical decision tasks, including medical diagnosis, and whether enhanced reasoning abilities can benefit clinical workflows.

Despite the growing body of research on DeepSeek's use in medical diagnosis, a significant gap remains in the literature: the lack of a comprehensive meta-analysis of the model's diagnostic capabilities and comparisons with physician performance. Such comparisons are crucial for understanding the real-world significance and effectiveness of DeepSeek in clinical settings. For other models, such as ChatGPT, LLaMA, and Gemini, comprehensive analyses of their performance in medical diagnostics have been conducted.[16] Therefore, there is a need for a systematic review and meta-analysis of DeepSeek's diagnostic performance to draw more reliable conclusions.

This study aims to fill this gap by systematically evaluating the diagnostic capabilities of DeepSeek in the medical field. Our focus is to provide a comprehensive assessment of DeepSeek's diagnostic performance and compare it with physician performance. By synthesizing results from various studies, we seek to gain deeper insights into DeepSeek's effectiveness, potential, and limitations in medical diagnosis. This analysis aims to provide a foundational reference for future research and practical applications in the field, ultimately advancing the development of AI-assisted diagnostics in healthcare.

Methods

Protocol and Registration

This systematic review was prospectively registered with PROSPERO (CRD420251125959). Our study adhered to the relevant sections of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for diagnostic accuracy studies (Supplementary Material 1 S3).[17,18] All stages of the review (title and abstract screening, full-text screening, data extraction, and bias assessment) were independently conducted by two reviewers (JZ and WQ), with discrepancies resolved through discussion with a third independent reviewer (SC).

Search Strategy and Study Selection

We conducted searches in the Web of Science Core Collection, Scopus, Embase, Medline, IEEE Xplore, and Medrxiv to identify studies validating DeepSeek for diagnostic tasks. A search strategy was developed, incorporating variants of DeepSeek and diagnostic-related terms, with the full search strategy provided in Supplementary Material 1 S1. Two authors (JZ and WQ) independently screened the titles and abstracts of the retrieved studies using the search strategy to identify studies meeting the inclusion and exclusion criteria (Textbox 1). Full-text assessments were then performed, with disagreements resolved through discussion with a third author. The search covered the period from database inception to August 8, 2025. To ensure comprehensiveness, we also reviewed the reference lists of relevant studies and citations. Literature management and duplicate removal were conducted using EndNote software.

Data Extraction

Before full-text screening, title and abstract screening were performed. A data extraction form was created using Microsoft Excel for data extraction, which was independently conducted by two reviewers. Any discrepancies between the reviewers were resolved through discussion. Information was extracted from each study, including the first author, model and version, model task, type of test dataset (internal, external, or unknown),[19] medical specialty, accuracy, sample size, and publication

status (preprint or peer-reviewed) for the meta-analysis of DeepSeek performance. Based on the relationship between the model's training and test data, three types of tests were defined.[20] Internal tests were defined as cases where the test data came from the same source or distribution as the training data but was appropriately separated from the training set using standard methods such as cross-validation or random splitting. External tests were defined as cases where the test data were collected after the training data cutoff, or when the model was tested on private data. Unknown tests were defined as cases where the test data were collected before the training data cutoff, and the data were publicly available. This distinction was made because the complete training datasets for the companies developing these models have not been made public. Additionally, when both model and physician diagnostic performance were presented in the same paper, both were extracted for meta-analysis. When a single model used multiple prompts and individual performances were available for each, the original prompt's performance was selected over the performance derived from iterative prompts. When multiple languages were used, the diagnostic performance based on the language of the test questions was selected.

Quality Assessment

We used Prediction model Risk of Bias Assessment Tool (PROBAST) to assess the risk of bias and applicability of the studies.[19] This tool utilizes signaling questions across four domains (participants, predictors, outcomes, and analysis) to provide both an overall and detailed assessment. Some PROBAST signaling questions were not included, as they are not relevant to generative AI models. We made modifications to PROBAST based on the study by Takita et al.,[16] with details of the modifications provided in Supplementary Material 1 S4.

Statistical Analysis

Heterogeneity was assessed using the I2 statistic. A random-effects model was applied when the I2 value exceeded 50%; otherwise, a fixed-effects model was used. Diagnostic accuracy was reported with 95% confidence intervals (CIs). Statistical significance was set at $p<0.05$. The pooled accuracy values for DeepSeek and physicians were calculated, and the overall accuracy of the model was compared with that of physicians. Subgroup analyses were also conducted to compare the model's performance across different specialties. To evaluate the impact of overall risk of bias, we performed a subgroup analysis limited to studies with low bias risk. Publication bias was assessed using funnel plots and Egger's regression test to evaluate its effect on the comparison of diagnostic performance between the model and physicians. Additionally, heterogeneity was analyzed in the full dataset and the low-bias-risk subgroup to assess its impact. All statistical analyses were conducted using R version 4.4.0.

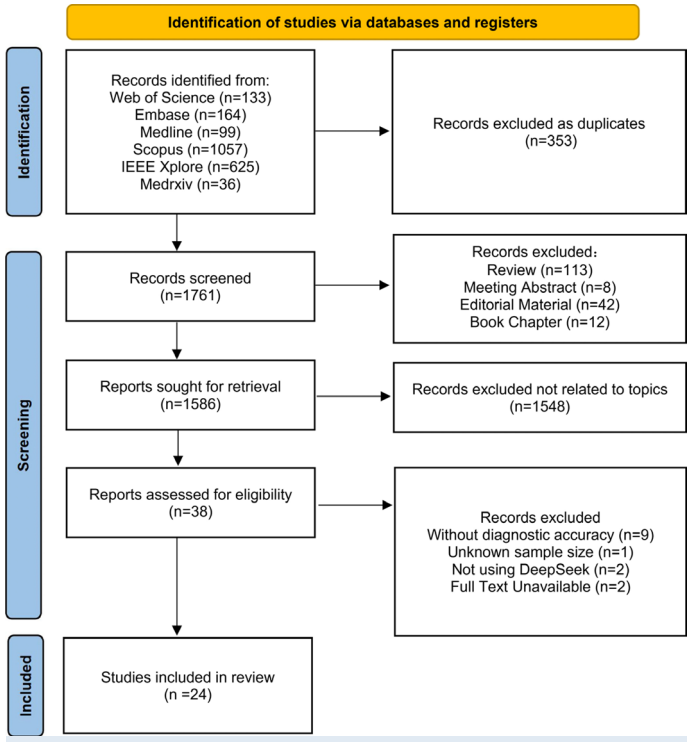


Figure 1. Literature screening process.

Results

Study Selection and Characteristics

A total of 2114 studies were identified across six databases, of which 353 were duplicates. After excluding reviews, conference abstracts, editorials, and book chapters, the titles and abstracts of 1586 studies were screened. From these, 1546 studies were excluded, and 39 studies underwent full-text review. Nine studies were excluded for lacking diagnostic accuracy, two for not using DeepSeek, two for inability to access the full text, and one for lacking sample size, resulting in the inclusion of 24 studies.[21–44] The detailed screening process is shown in Figure 1.

Among the 24 studies, all were published in 2025 and involved 11 countries or regions (Fig. 2). The United States contributed the most studies (6), followed by China (4) and Canada (3). The most frequently evaluated model was DeepSeek-R1 (16 studies), which includes other parameters or versions such as DeepSeek-R1-70B and

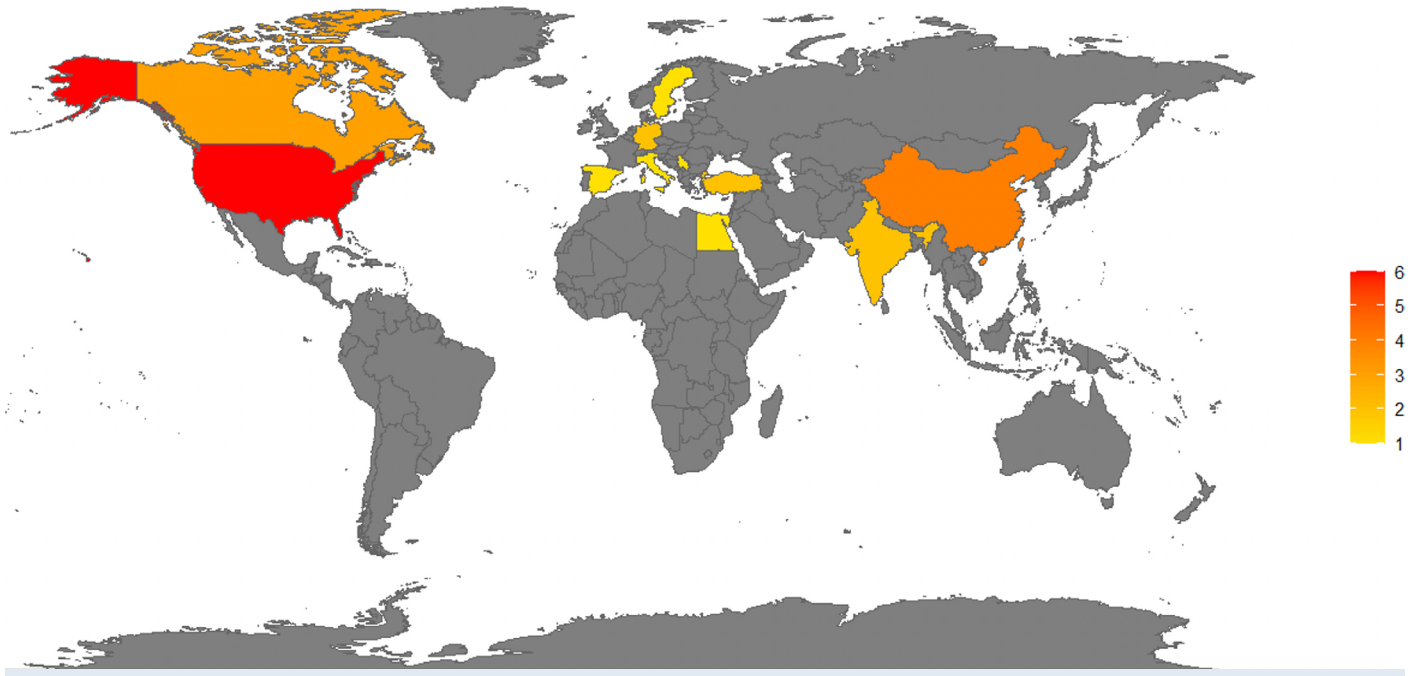


Figure 2. Distribution by country or region.

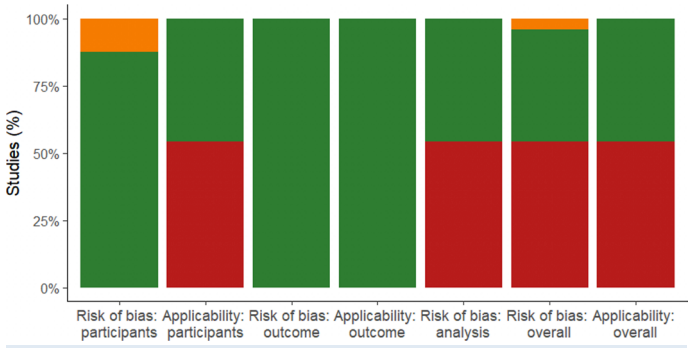


Figure 3. Summary of Prediction Model Study Risk of Bias Assessment Tool risk of bias.

DeepSeek-R1-Distill-LLaMA-70B; these were grouped under the DeepSeek-R1 category. The next most evaluated model was DeepSeek-V3 (6 studies), while DeepSeek Medical and DeepSeek VL2 Tiny (1B) were each represented by one study. Detailed information on each model can be found in Appendix A: Supplementary Material 1 S2.

This review covers a wide range of medical specialties, with ophthalmology (5 studies) and general medicine (4 studies) being the most common. Other specialties represented include oral medicine (3), pediatrics (2), and emergency medicine (2). Additionally, there was representation from gastroenterology, oncology, neurology, sleep medicine, radiology, otolaryngology, critical care medicine, and rheumatology, with one study each. Regarding model tasks, free-text tasks were the most common (20 studies), followed by classification tasks (4 studies). In terms of test dataset type, 11 studies involved external testing, while 13 studies had unknown test dataset types due to the unavailability of training data for generative AI models. Among the included studies, 17 were peer-reviewed articles, 6 were preprints, and 1 was a letter. The study characteristics are summarized in Table 1 and the Supplementary Material 2. Nine studies compared DeepSeek's performance with that of physicians.[21,25,27–30,39,40,44] Of these, 6 studies used DeepSeek-R1, 2 used DeepSeek-V3, and 1 used DeepSeek Medical.

Quality Assessment

The risk of bias assessment using the PROBAST indicated that participant and outcome adjudication were generally associated with a low risk of bias, while analysis was associated with a high risk of bias. Overall, 13 studies (54%) were assessed as having a high risk of bias, 10 studies (42%) were assessed as having a low risk of bias, and 1 study had an unknown risk of bias. In terms of applicability, there were high concerns regarding the applicability of participants, while there were low concerns regarding the applicability of outcomes. In total, 13 studies (54%) were rated as having high concerns regarding generalizability, and 11 studies (46%) were rated as having low concerns regarding generalizability[19] (Fig. 3). The primary factors contributing to these concerns were studies evaluating models with small test datasets and research where external validation could not be substantiated due to the unknown training data of generative AI models. Detailed results are provided in Supplementary Material 2.

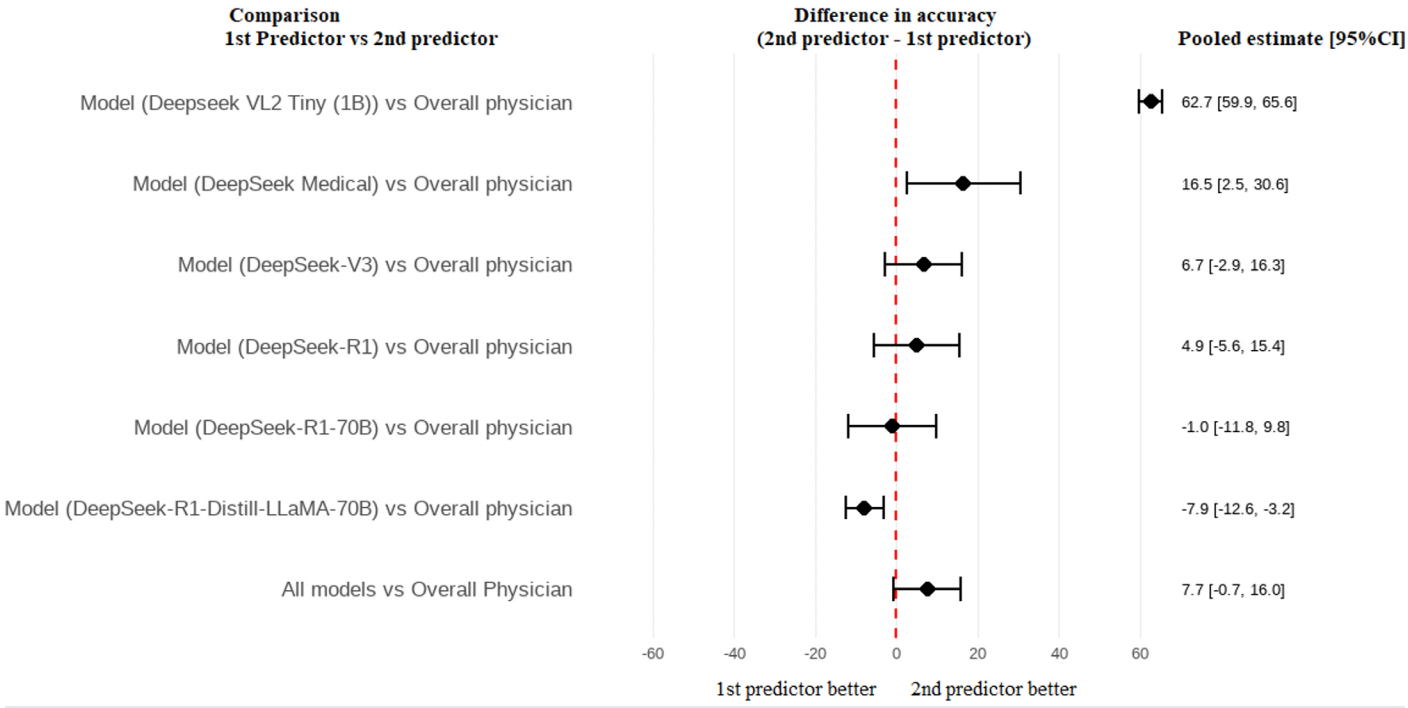


Figure 4. Comparison results between models and physicians.

Meta-Analysis

The comparison results between the model and physicians are shown in Figure 4. Meta-analysis revealed no significant difference between DeepSeek and physicians overall (physician accuracy was 7.7% higher [95% CI: -0.7–16.0%], p=0.07). DeepSeek VL2 Tiny (1B) exhibited significantly lower diagnostic performance than physicians (p<0.001), and DeepSeek Medical also performed worse than physicians (p<0.05). Although DeepSeek-V3, DeepSeek-R1, and DeepSeek-R1-70B showed no significant differences compared to physicians, DeepSeek-V3 and DeepSeek-R1 had lower diagnostic accuracy, while DeepSeek-R1-70B slightly outperformed physicians. Interestingly, DeepSeek-R1-Distill-LLaMA-70B performed significantly better than physicians.

In our meta-regression (Fig. 5), we found no significant differences in model performance between general medicine and specialties such as rheumatology, radiology, oral medicine, otolaryngology, and critical care medicine. Significant differences were observed when comparing model performance with specialties like gastroenterology, neurology, pediatrics, and ophthalmology (p<0.001). Model performance also showed some significance when compared with sleep medicine and oncology (p<0.05), with general medicine performing worse than these specialties. In contrast, a significant difference was observed when compared with emergency medicine (p<0.001), but the model performed significantly better in general medicine than in emergency medicine. In the low-bias risk subgroup analysis, DeepSeek's performance overall did not significantly differ from that of physicians (p=0.514). Meta-regression analysis of bias risk and publication status revealed no significant differences for bias risk (p=0.708) or publication status (p=0.214). In the heterogeneity analysis, the R^2 value (representing the degree of explained heterogeneity) was 45.4% for all studies and 18.5% for low-bias risk studies, indicating a moderately low level of explainable variation. We assessed publication bias through regression analysis and quantified the asymmetry of the funnel plot (Supplementary Material 1 S5). The results indicated no significant risk of publication bias (p=0.200).

Discussion

In this systematic review and meta-analysis, we analyzed the diagnostic performance of DeepSeek and physicians. To our knowledge, this is the first analysis comparing the performance of all versions of DeepSeek in medical diagnosis, alongside physicians. Initially, we identified 2114 studies, and 24 studies were ultimately included for the systematic review and meta-analysis. This study covers multiple versions of the DeepSeek model and a wide range of medical specialties, with DeepSeek-R1 being the most frequently evaluated model and Ophthalmology being the most commonly represented medical specialty. Quality

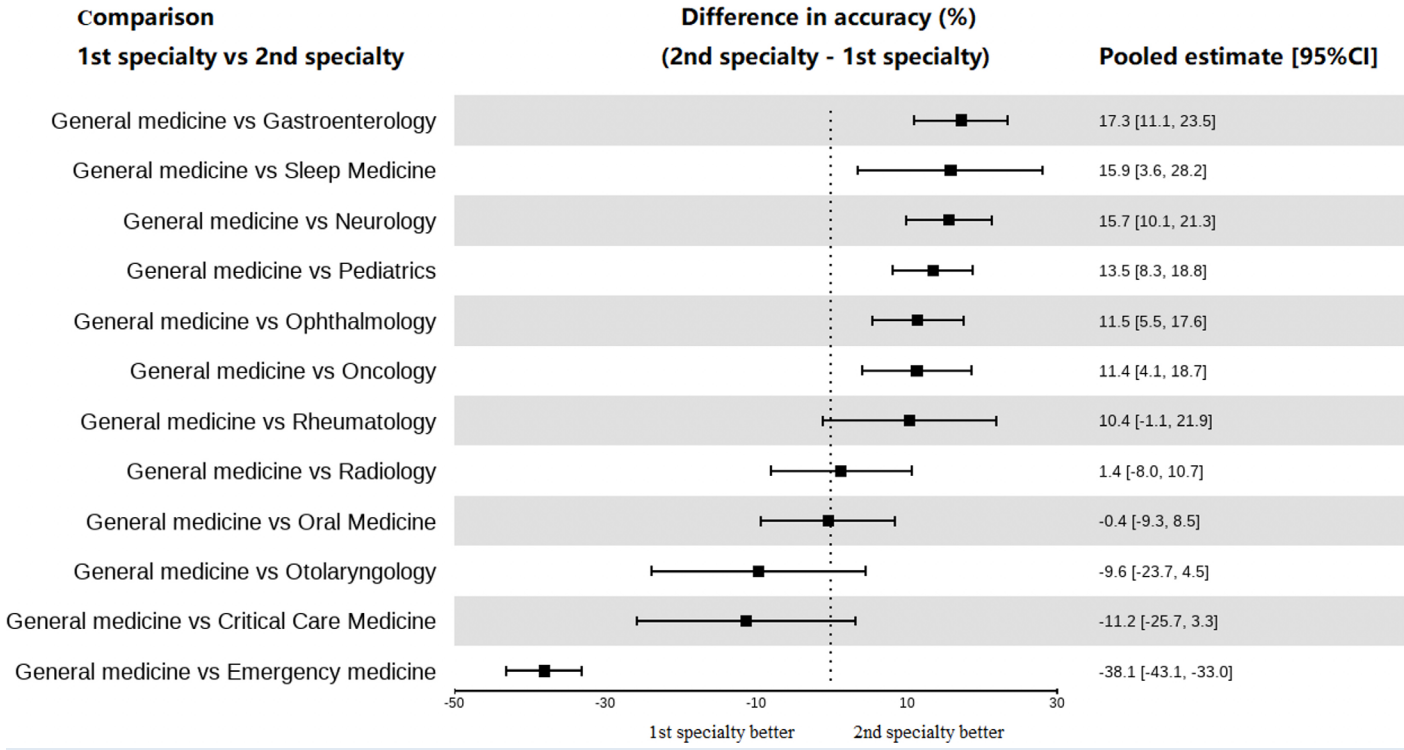


Figure 5. DeepSeek performance among specialties.

assessment revealed that most studies had a high risk of bias. Meta-analysis showed that the average diagnostic accuracy of the DeepSeek models was 76% (95% CI: 67–83%). Some models performed similarly to physicians, with no significant performance differences (accuracy difference: 7.7% [95% CI: -0.7–16.0%]; p=0.07). Furthermore, our analysis also found that DeepSeek's performance varied across most medical fields. This comprehensive study highlights the varied capabilities and limitations of generative AI in medical diagnostics.

LLMs have the potential to improve patient care by enhancing clinical decision-making and automating administrative tasks.[45] However, adhering to data privacy regulations and medical device compliance poses challenges for proprietary LLM implementations.[46,47] Open-source LLMs present a potential alternative for clinical applications. The release of the open-source DeepSeek model has garnered widespread attention.[48,49] including in the medical field. The meta-analysis of DeepSeek models in healthcare provides critical insights into its clinical implications. With an accuracy of 76%, DeepSeek demonstrates near-excellent performance, confirming its substantial potential for application in certain clinical scenarios. Some models, including DeepSeek-V3, DeepSeek-R1, and DeepSeek-R1-70B, showed comparable performance to physicians, and even DeepSeek-R1-Distill-LLaMA-70B outperformed physicians. Although the analysis of DeepSeek-R1-70B and DeepSeek-R1-Distill-LLaMA-70B is based on a single study, it is undeniable that DeepSeek can assist in the provision of medical services or serve as a preliminary diagnostic tool in resource-limited environments.[50] Furthermore, compared to proprietary models, DeepSeek offers local deployment capabilities and customization potential, making it highly suitable for various medical research and clinical tasks.[51] It is expected to be deployed at scale in hospitals.[52] This large-scale application marks the expansion of AI beyond diagnostic support to areas such as hospital management, research facilitation, and patient management.[53]

There were notable differences in the performance of DeepSeek across various medical specialties. However, most specialties included only one study, which limits the generalizability of these findings and calls for further investigation. Additionally, among the models included, all except DeepSeek VL2 Tiny (1B) do not support direct image analysis or interpretation. Consequently, image-related tasks were typically excluded during model testing or converted into text for testing purposes. This limitation may restrict the use of DeepSeek in specialties such as Ophthalmology and Oral Medicine, but if users describe visual features of conditions in text form—such as appearance, location, color, associated

Table 1. Study characteristics

Citation	First author	Model	Model task	Test type	Specialty	Comparison group	Eligible	Preprint	Overall ROB	Overall applicability
[21]	Aminan	DeepSeek-R1	Free text	Unknown	Ophthalmology	Physicians	11	Preprint	High	Low
[26]	Hoyt	DeepSeek-R1	Free text	Unknown	General medicine	NA	162	Preprint	Low	Low
[27]	Hussain	DeepSeek-V3	Free text	External	Ophthalmology	Physicians	53	Preprint	High	High
[36]	Mondillo	DeepSeek-R1	Choice	Unknown	Pediatrics	NA	500	Preprint	Low	Low
[37]	Mruthyunjaya	DeepSeek-R1-70B	Free text	External	Rheumatology	NA	50	Preprint	High	Low
[38]	Naliyatthaliyazchayil	DeepSeek-R1	Free text	External	Emergency medicine	NA	300	Preprint	Low	High
[22]	Chan	DeepSeek-R1	Free text	Unknown	General medicine	NA	100	Peer-reviewed	Low	High
[24]	Diniz-Freitas	DeepSeek-R1	Free text	Unknown	Oral medicine	NA	36	Peer-reviewed	High	High
[25]	Hassanein	DeepSeek-V3	Free text	Unknown	Oral medicine	Physicians	80	Peer-reviewed	High	High
[28]	Ilic	DeepSeek Medical	Free text	External	Pediatrics	Physicians	45	Peer-reviewed	High	High
[29]	Cheng	DeepSeek-R1	Free text	Unknown	Ophthalmology	Physicians	20	Peer-reviewed	High	Low
[30]	Kang	DeepSeek-R1	Free text	External	Oncology	Physicians	159	Peer-reviewed	Low	High
[31]	Kaygisiz	DeepSeek-V3	Free text	External	Oral medicine	NA	16	Peer-reviewed	High	High
[32]	Kurz	Deepseek-VL2	Choice	Unknown	Emergency medicine	NA	1012	Peer-reviewed	Low	Low
[34]	Mikhail	DeepSeek-R1	Choice	Unknown	Ophthalmology	NA	158	Peer-reviewed	Low	Low
[35]	Moël	DeepSeek-R1	Choice	Unknown	General medicine	NA	100	Peer-reviewed	Low	Low
[39]	Pan	DeepSeek-R1	Free text	External	Neurology	Physicians	338	Peer-reviewed	Low	High
[40]	Patel	DeepSeek-R1	Free text	Unknown	Sleep medicine	Physicians	32	Peer-reviewed	High	Low
[41]	Spitzl	DeepSeek-V3	Free text	External	Radiology	NA	111	Peer-reviewed	Unclear	High
[42]	Tordjman	DeepSeek-R1	Free text	Unknown	General medicine	NA	50	Peer-reviewed	High	High
[43]	Camalan	DeepSeek-V3	Free text	External	Otolaryngology	NA	50	Peer-reviewed	High	High
[44]	Wu	DeepSeek-R1	Free text	External	Critical care medicine	Physicians	48	Peer-reviewed	High	Low
[33]	Li	DeepSeek-V3	Free text	Unknown	Ophthalmology	NA	22	Letter	High	Low
[23]	Dai	DeepSeek-R1-Distill	Free text	External	Gastroenterology	NA	198	Peer-reviewed	Low	High

ROB: Risk of bias.

symptoms, and duration—DeepSeek may still provide relevant information,[24] thus reducing this limitation. However, studies have found that despite ChatGPT-4o's ability to directly analyze images, adding images to case descriptions did not improve diagnostic accuracy.[54] While it is unclear whether the lack of image analysis capabilities affects DeepSeek's diagnostic accuracy, its technical limitations in multimodal data integration hinder its ability to provide comprehensive diagnostic and treatment recommendations. A potential solution to this issue is the implementation of hierarchical attention mechanisms (e.g., cross-modal attention weight allocation between images and text), combined with time-series modeling to capture the dynamic features of disease progression.[55]

Research comparing DeepSeek and physician performance also offers an intriguing perspective in the context of medical education.[56] Currently, DeepSeek's performance is comparable to that of physicians, and the model's ability to be deployed locally and its cost-effectiveness[13] suggest opportunities for integrating it into medical training. Since studies comparing models to physicians do not specify the physicians' experience levels, we assume the doctors involved in the included studies are non-experts, making it impossible to conduct further comparisons between expert and non-expert physicians and the model. However, DeepSeek can be used as an educational tool for medical students or residents, especially when simulating non-expert scenarios, where its performance is on par with healthcare professionals.[57] This integration can enhance the learning experience, especially when DeepSeek is locally deployed, allowing training with complex or rare cases using local data, providing diverse clinical case studies, and offering targeted instruction to students or residents, while also promoting self-assessment and feedback. Although the model's performance does not significantly differ from physicians, its diagnostic accuracy has yet to reach that of physicians, emphasizing the irreplaceable value of human judgment and experience in medical decision-making. One key advantage of models like DeepSeek-R1 is its reasoning-centered design, making the decision-making process more interpretable and transparent. However, this feature does not eliminate the fundamental issue of AI-generated "hallucinations," where models generate seemingly reasonable but actually erroneous medical information, which poses significant risks in patient care, especially when AI output influences diagnosis, treatment recommendations, or research conclusions.[58] Furthermore, DeepSeek-R1's self-reflection capability could be used to bypass its safety constraints, raising concerns about models generating diagnoses that deviate from established medical guidelines or rationalizing incorrect treatments.[59] Therefore, external safeguards, such as rule-based reinforcement filters, human-in-the-loop validation, and continuous real-world verification, are crucial for safe deployment.[60] In this context, we propose a hybrid system where physicians ask the model for differential diagnoses of clinical conditions, which may include diagnoses not previously considered by the physician. This collaborative model can leverage the strengths of both parties and provide more comprehensive diagnostics.

To examine the impact of overall bias risk, we conducted a subgroup analysis of studies with low bias risk. The results of studies with low bias risk did not significantly differ from those of the full dataset. Therefore, the higher proportion of studies with high bias risk does not substantively affect our findings or their generalizability. The training data specifics of the DeepSeek series models have not been fully disclosed, but the transparency of training data and its collection period is crucial. Without this transparency, we cannot determine if the test dataset is an external dataset, which could introduce bias. Transparency ensures the fairness, interpretability, and scientific integrity of the model, helps identify potential biases, and promotes independent replication and validation. As AI technology continues to evolve, data transparency will become increasingly important. Thus, establishing strict data collection and management standards to ensure transparency and fairness should be one of the core tasks in LLMs research and application.

Study Limitations

Although the methodology of this study is comprehensive, it has certain limitations. The generalizability of our findings requires careful consideration. Heterogeneity analysis revealed that the explained variability was at a moderately low level, indicating that our meta-regression model did not fully account for the differences between studies. Other factors not addressed in the analysis may influence DeepSeek's performance, including case complexity, model hyperparameter settings, and prompt design. Furthermore, the included studies did not clearly specify the languages used, but it can be confirmed that all were non-Chinese, with only one study utilizing bilingual input (Chinese and English).[39] The impact of using native languages (e.g., Chinese) on the model's performance remains uncertain. Finally, we did not compare DeepSeek with other models, and further research is needed to extend this analysis.

Conclusion

This meta-analysis provides a detailed understanding of the capabilities and limitations of DeepSeek in medical diagnosis. With an overall diagnostic accuracy of 76%, DeepSeek shows no significant difference in performance compared to physicians. However, it has not yet reached the level of physicians and cannot replace expert clinicians. It may, however, serve as a valuable auxiliary tool in non-specialized settings and as an educational tool for medical students. Furthermore, it is important to emphasize the need for continued advancement and specialization in model development, as well as the necessity for rigorous, external validation studies to address the widespread high risk of bias. The transparency of training data should also be prioritized to ensure that DeepSeek can effectively integrate into clinical practice.

Ethics Committee Approval: Ethical approval was not required for this study since this is a review article.

Conflict of Interest: The author declare that there is no conflict of interest.

Financial Disclosure: The authors declared that this study received no financial suppor.

Use of AI for Writing Assistance: Not declared.

Authorship Contributions: JZ, XZ, XL, SS, SC; Design: JZ, XZ, XL, SS, SL; Supervision: JZ, XZ, XL, SS, SC; Data Collection and/or Processing: JZ, XZ, XL, SS, SL; Analysis and/or Interpretation: JZ, XZ, XL, SS, SL; Literature Search: JZ, XZ, XL, SS, SL; Writing: JZ, XZ, XL, SS, SL; Critical Reviews: JZ, XZ, XL, SS, SL.

Peer-review: Double blind peer-reviewed.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-40. [CrossRef]
2. Bellini V, Bignami EG. Generative Pre-trained Transformer 4 (GPT-4) in clinical settings. Lancet Digit Health 2025;7(1):e6-7. [CrossRef]
3. Boussina A, Krishnamoorthy R, Quintero K, Joshi S, Wardi G, Pour H, et al. Large language models for more efficient reporting of hospital quality measures. NEJM AI 2024;1(11). [CrossRef]
4. McCoy TH, Perlis RH. Applying large language models to stratify suicide risk using narrative clinical notes. J Mood Anxiety Disord 2025;10:100109. [CrossRef]
5. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA 2023;330(1):78-80. [CrossRef]
6. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology 2023;308(1):e231040. [CrossRef]
7. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024;7(10):e2440969. [CrossRef]
8. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun 2024;15(1):2050. [CrossRef]
9. Liu A, Feng B, Wang B, Wang B, Liu B, Zhao C, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv 2024;240504434.
10. Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv 2025;250112948.
11. Deng Z, Ma W, Han Q-L, Zhou W, Zhu X, Wen S, et al. Exploring DeepSeek: a survey on advances, applications, challenges and future directions. Ieee-Caa Journal of Automatica Sinica 2025;12(5):872-93. [CrossRef]
12. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 2025;17(2):e79221. [CrossRef]
13. Gibney E. Scientists flock to DeepSeek: how they're using the blockbuster AI model. Nature 2025. doi: 10.1038/d41586-025-00275-0. Epub ahead of print. PMID: 39881178. [CrossRef]
14. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025;638(8049):13-4. [CrossRef]
15. Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. Nature. 2025;638(8050):300-1. [CrossRef]
16. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med 2025;8(1):175. [CrossRef]
17. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM; the PRISMA-DTA Group; et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. JAMA 2018;319(4):388-96. [CrossRef]
18. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535. [CrossRef]
19. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51-8. [CrossRef]
20. Walston SL, Seki H, Takita H, Mitsuyama Y, Sato S, Hagiwara A, et al. Data set terminology of deep learning in medicine: a historical review and recommendation. Jpn J Radiol 2024;42(10):1100-9. [CrossRef]
21. Aminan M, Darnell SS, Delsoz M, Nabavi A, Wright C, Jerkins B, et al. GlaucoRAG: A Retrieval-Augmented Large Language Model for Expert-Level Glaucoma Assessment. medRxiv 2025;2025.07.03.25330805. [CrossRef]

22. Chan LN, Xu XJ, Lv KY. DeepSeek-R1 and GPT-4 are comparable in a complex diagnostic challenge: a historical control study. International Journal of Surgery 2025;111(6):4056-9. [CrossRef]
23. Dai J, Kim MY, Sutton RT, Mitchell JR, Goebel R, Baumgart DC. Comparative analysis of natural language processing methodologies for classifying computed tomography enterography reports in Crohn's disease patients. NPJ Digit Med 2025;8(1):324. [CrossRef]
24. Diniz-Freitas M, Diz-Dios P. DeepSeek: Another step forward in the diagnosis of oral lesions. J Dent Sci 2025;20(3):1904-7. [CrossRef]
25. Hassanein FEA, El Barbary A, Hussein RR, Ahmed Y, El-Guindy J, Sarhan S, et al. Diagnostic Performance of ChatGPT-4o and DeepSeek-3 Differential Diagnosis of Complex Oral Lesions: A Multimodal Imaging and Case Difficulty Analysis. Oral Dis 2025;31(12):3361-71. [CrossRef]
26. Hoyt RE, Knight D, Haider M, Bajwa M. Evaluating a Large Reasoning Model's Performance on Open-Ended Medical Scenarios. medRxiv 2025. DOI: 10.1101/2025.04.29.25326666. Epub ahead of print. [CrossRef]
27. Hussain ZS, Delsoz M, Elahi M, Jerkins B, Kanner E, Wright C, et al. Performance of DeepSeek, Qwen 2.5 MAX, and ChatGPT Assisting in Diagnosis of Corneal Eye Diseases, Glaucoma, and Neuro-Ophthalmology Diseases Based on Clinical Case Reports. medRxiv 2025. doi: 10.1101/2025.03.14.25323836. Epub ahead of print. PMID: 40166547. [CrossRef]
28. Ilić N, Marić N, Cvetković D, Bogosavljević M, Bukara-Radujković G, Krstić J, et al. The artificial intelligence-assisted diagnosis of skeletal dysplasias in pediatric patients: a comparative benchmark study of large language models and a clinical expert group. Genes (Basel) 2025;16(7):762. [CrossRef]
29. Jiao C, Rosas E, Asadigandomani H, Delsoz M, Madadi Y, Raja H, et al. Diagnostic performance of publicly available large language models in corneal diseases: a comparison with human specialists. Diagnostics (Basel) 2025;15(10):1221. [CrossRef]
30. Kang C, Li J, Yang X, Ren G, Zhang L, Wang W, et al. Performance of large language models in the differential diagnosis of benign and malignant biliary stricture. Front Oncol 2025;15:1613818. [CrossRef]
31. Kaygisiz ÖF, Teke MT. Can deepseek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health 2025;25(1):638. [CrossRef]
32. Kurz CF, Merzhevich T, Eskofier BM, Kather JN, Gmeiner B. Benchmarking vision-language models for diagnostics in emergency and critical care settings. NPJ Digit Med 2025;8(1):423. [CrossRef]
33. Li X, He J, Xie JS, Sharma RA. Comment: Diagnosing Neuro-Ophthalmology Diseases Based on Case Reports: DeepSeek vs ChatGPT. J Neuroophthalmol 2025;45(3):e261-2. [CrossRef]
34. Mikhail D, Farah A, Milad J, Nassrallah W, Mihalache A, Milad D, et al. Performance of DeepSeek-R1 in ophthalmology: an evaluation of clinical decision-making and cost-effectiveness. Br J Ophthalmol 2025;109(9):976-81. [CrossRef]
35. Moëll B, Sand Aronsson F, Akbar S. Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. Front Artif Intell 2025;8:1616145. [CrossRef]
36. Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Masino M. Comparative Evaluation of Advanced AI Reasoning Models in Pediatric Clinical Decision Support: ChatGPT O1 vs. DeepSeek-R1. medRxiv 2025;01.27.25321169. [CrossRef]
37. Mruthyunjaya P, Verma S, Agarwal A, Maharana U, Mandal M, Ahmed S. Right Diagnoses But Wrong Reasoning: Current Large-Language Model-Based Agentic Frameworks Have Flawed Clinical Reasoning Despite High Diagnostic Accuracy. SSRN 2025. doi: 10.2139/ssrn.5339074. Epub ahead of print. [CrossRef]
38. Naliyatthaliyazchayil P, Muthyala R, Gichoya JW, Purkayastha S. Evaluating the Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification: Zero-Shot Prompting Approach. J Med Internet Res 2025;27:e74142. [CrossRef]
39. Pan Y, Tian S, Guo J, Cai H, Wan J, Fang C. Clinical feasibility of AI Doctors: Evaluating the replacement potential of large language models in outpatient settings for central nervous system tumors. Int J Med Inform 2025;203:106013. [CrossRef]
40. Patel A, Ruoff C, Helgeson SA, Carvalho DZ, Castillo PR, Cheung J. Diagnostic performance of Large Language Models (LLMs) compared with physicians in sleep medicine. Sleep Med 2025;134:106677. [CrossRef]
41. Spitzl D, Mergen M, Braren R, Endrös L, Eiber M, Steinhelfer L. LLM-powered breast cancer staging from PET/CT reports: a comparative performance study. Int J Med Inform 2025;204:106053. [CrossRef]
42. Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med 2025;31(8):2550-5. [CrossRef]
43. Vural Camalan B, Doluoglu S, Taraf NH, Gunay MM, Ozlugedik S. ChatGPT versus DeepSeek in head and neck cancer staging and treatment planning: guideline-based study. Eur Arch Otorhinolaryngol 2025;282(9):4815-24. [CrossRef]
44. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. Crit Care 2025;29(1):230. [CrossRef]
45. Quer G, Topol EJ. The potential for large language models to transform cardiovascular medicine. Lancet Digit Health 2024;6(10):e767--1. [CrossRef]
46. de Hond A, Leeuwenberg T, Bartels R, van Buchem M, Kant I, Moons KG, van Smeden M. From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Health 2024;6(7):e441-3. [CrossRef]
47. Ong JCL, Chang SY-H, William W, Butte AJ, Shah NH, Chew LST, et al. Medical ethics of large language models in medicine. NEJM AI 2024;1(7). [CrossRef]
48. Poo M-m. Reflections on DeepSeek's breakthrough. National Science Review. 2025;12(3):nwaf044. [CrossRef]
49. Sandmann S, Eils R. Open-source LLM DeepSeek on a par with proprietary models in clinical decision making. Nature Medicine. 2025;31:2496-7. [CrossRef]

50. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health 2018;3(4):e000798. [CrossRef]
51. MohanaSundaram A, Sathanantham ST, Ivanov A, Mofatteh M. DeepSeek's Readiness for Medical Research and Practice: Prospects, Bottlenecks, and Global Regulatory Constraints. Ann Biomed Eng 2025;53(7):1754-6. [CrossRef]
52. Ye H. Key Players Overlooked in the Rapid Deployment of DeepSeek To China's Hospitals. Journal of Medical Systems 2025;49. [CrossRef]
53. Chen J, Miao C. DeepSeek Deployed in 90 Chinese Tertiary Hospitals: How Artificial Intelligence Is Transforming Clinical Practice. J Med Syst 2025;49(1):53. [CrossRef]
54. Diniz-Freitas M, Lago-Méndez L, Limeres-Posse J, Diz-Dios P. Challenging ChatGPT-4V for the Diagnosis of Oral Diseases and Conditions. Oral Dis 2025;31(2):701-6. [CrossRef]
55. Liang W, Chen P, Zou X, Lu X, Liu S, Yang J, et al. DeepSeek: the "Watson" to doctors-from assistance to collaboration. J Thorac Dis 2025;17(2):1103-5. [CrossRef]
56. Preiksaitis C, Rose C. Opportunities, Challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023;9:e48785. [CrossRef]
57. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3(1):141. [CrossRef]
58. Wang X, Zhang NX, He H, Nguyen T, Yu K-H, Deng H, et al. Safety challenges of AI in medicine in the era of large language models. arXiv 2024;2409.18968
59. Mercer S, Spillard S, Martin DP. Brief analysis of DeepSeek R1 and its implications for Generative AI. arXiv 2025;250202523. [CrossRef]
60. Pantha N, Ramasubramanian M, Gurung I, Maskey M, Ramachandran R. Challenges in guardrailing large language models for science. arXiv 2024;241108181.

S1 Search strategy

Database	Query
WoS CC	(ALL=("DeepSeek")) AND ALL=("medic*" OR "healthcare" OR "diagnosis" OR "diagnostic" OR "diagnose" OR "quiz" OR "exam*" OR "vignette")
Embase	(ALL=('Deepseek')) AND ALL=('medic*' OR 'healthcare' OR 'diagnosis' OR 'diagnostic' OR 'diagnose' OR 'quiz' OR 'exam*' OR 'vignette')
Scopus	(ALL(DeepSeek) AND ALL(medic* OR healthcare OR "diagnosis" OR "diagnostic" OR "diagnose" OR "quiz" OR "exam*" OR "vignette"))
Medline	(TS=("DeepSeek")) AND TS=("medic*" OR "healthcare" OR "diagnosis" OR "diagnostic" OR "diagnose" OR "quiz" OR "exam*" OR "vignette")
MedRxiv	(DeepSeek) AND (medical OR medicine OR healthcare OR diagnosis OR diagnostic OR diagnose OR quiz OR examination OR vignette)
IEEE Xplore	(Full text only=("DeepSeek")) AND Full text only=("medic*" OR "healthcare" OR "diagnosis" OR "diagnostic" OR "diagnose" OR "quiz" OR "exam*" OR "vignette")

S2 Models' description

Model	**Description**
DeepSeek-V3	Parameters: 671B; Based on a mixture-of-experts (MoE) architecture, supporting up to 128K context length, suitable for mathematical, programming, and Chinese tasks. Its performance rivals that of GPT-4o, with a training cost below 6 million USD.
DeepSeek-R1	Parameters: 671B; Built upon DeepSeek-V3, enhanced with reinforcement learning for reasoning capabilities, supporting chain-of-thought (CoT), excelling in mathematics, programming, and reasoning tasks.
DeepSeek-R1-70B	Parameters: 70B; Distilled from DeepSeek-R1, based on the Llama3.3-70B-Instruct architecture, suitable for reasoning and code generation tasks.
DeepSeek-R1-Distill-LLaMA-70B	Parameters: 70B; Based on the Llama3.3-70B-Instruct architecture, distilled from DeepSeek-R1 for reasoning, applicable to a wide range of natural language processing tasks.
DeepSeek Medical	Parameters: Not disclosed; Fine-tuned from DeepSeek-R1 for the medical domain, supporting diagnostic and clinical decision-making tasks, with chain-of-thought (CoT) and advanced reasoning capabilities.
Deepseek VL2 Tiny (1B)	Parameters: 1B; A vision-language model based on MoE architecture, suitable for visual question answering, document understanding, and visual localization tasks, employing dynamic tiling visual encoding strategies.

S3 PRISMA 2020 checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Introduction
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Introduction
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Search strategy and study selection
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Search strategy and study selection
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Supplementary Table 1
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Figure 1
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Protocol and Registration, Data extraction
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Data Extraction, Table1, Supplement ary Material 2
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Data Extraction, Table1, Supplement ary Material 2
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Quality Assessment
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the	Statistical

Section and Topic	Item #	Checklist item	Location where item is reported
		synthesis or presentation of results.	Analysis
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Statistical Analysis
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Statistical Analysis
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Statistical Analysis
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Statistical Analysis
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Statistical Analysis
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Statistical Analysis
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Quality Assessment, Statistical Analysis
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Quality Assessment, Statistical Analysis
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Figure 1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Figure 1
Study characteristics	17	Cite each included study and present its characteristics.	Table 1, Supplement ary Material 2
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Quality Assessment
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Meta analysis
Results of	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing	Meta analysis

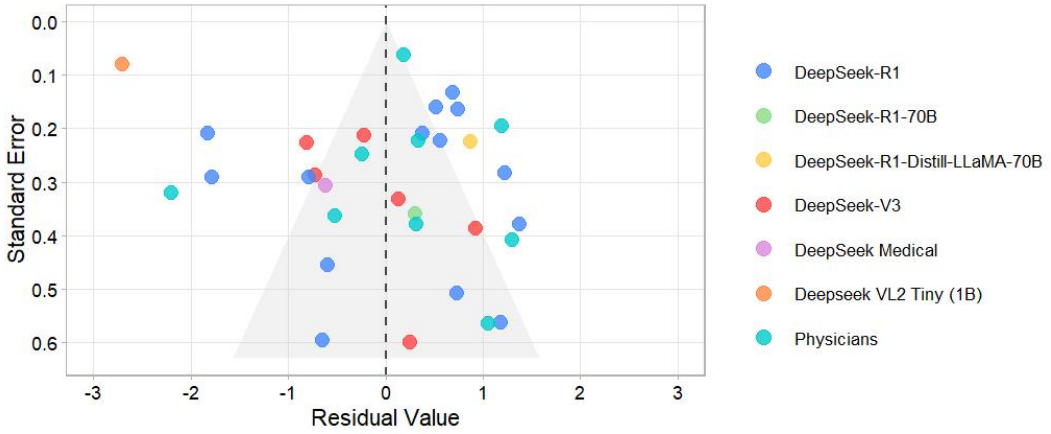
Section and Topic	Item #	Checklist item	Location where item is reported
syntheses		studies.	
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Meta analysis
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Meta analysis
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Meta analysis
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Meta analysis
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Meta analysis
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion
	23b	Discuss any limitations of the evidence included in the review.	Discussion
	23c	Discuss any limitations of the review processes used.	Discussion
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Protocol and Registration in the Methods
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Protocol and Registration in the Methods
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Protocol and Registration in the Methods
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Acknowled gements
Competing interests	26	Declare any competing interests of review authors.	Competing Interests
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data availability

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

S4 PROBAST modifications

PROBAST Items	Modifications
Domain 1: Participants	No changes (Refer to participant data for diagnosis)
Domain 2: Predictors	N/A–removed from scoring
Domain 3: Outcome	Items 3.3, 3.5, and 3.6 N/A
Domain 4: Analysis	Items 4.5, 4.6, and 4.9 N/A
Domain 5: Overall	No changes

S5 Funnel plot.



The funnel plot illustrates the distribution of the residuals of the fitted values corresponding to their standard errors in the meta-regression. The Egger test results z-value = 1.3085 and p = 0.200, which indicates no significant publication bias.

Supplementary material 2

Year	Study	Model	Version	Database	Referebce standard	ROB of Participants	Applicability of participants	ROB of Outcome	Applicability of outcome	ROB of analysis
2025	Aminan[21]	DeepSeek-R1	Not written	The case reports from EyeRounds	Expert consensus	Low	Low	Low	Low	High
2025	Hoyt[26]	DeepSeek-R1	Not written	Medical scenarios from MMLU-Pro question bank	Answer	Low	Low	Low	Low	Low
2025	Hussain[27]	DeepSeek V3	Not written	Case reports from university-accessible databases		Low	High	Low	Low	High
2025	Mondillo[36]	DeepSeek-R1	Not written	Questions from the MedQA dataset	Answer	Low	Low	Low	Low	Low
2025	Mruthyunjaya[37]	DeepSeek-R1-70B	Not written	Cases in knowledge base	Expert consensus	Low	Low	Low	Low	High
2025	Naliyatthaliyazchayil[48]	DeepSeek-R1	Not written	Patient cases in MIMIC-IV Notes database	Expert consensus	Low	High	Low	Low	Low
2025	Chan[22]	DeepSeek-R1	Not written	Cases from the NEJM	Expert consensus	Low	High	Low	Low	Low
2025	Dai[23]	DeepSeek-R1-Distill-LLaMA-70B	Not written	CTE reports of patients with CD and controls collected in IBD patient registry	Expert consensus	Low	High	Low	Low	Low
2025	Diniz-Freitas[24]	DeepSeek-R1	Not written	The quiz “Image Challenge” in the NEJM	Answer	Low	High	Low	Low	High
2025	Hassanein[25]	DeepSeek-3	Not written	Case reports retrieved in PubMed	Answer	Low	High	Low	Low	High
2025	Ilic[28]	DeepSeek Medical (2024)	Not written	Vignettes obtained from medical records	Expert consensus	Low	High	Low	Low	High
2025	Cheng[29]	DeepSeek-R1	Not written	The case reports from EyeRounds	Expert consensus	Low	Low	Low	Low	High
2025	Kang[30]	DeepSeek-R1	Not written	Cases from the hospital	Expert consensus	Low	High	Low	Low	Low
2025	Kaygisiz[31]	DeepSeek-V3	DeepSeek-2025-02-18	Clinical case scenarios made by the authors	Expert consensus	Unclear	High	Low	Low	High

2025	Kurz[32]	Deepseek VL2 Tiny (1B)	Not written	The quiz “Image Challenge” in the NEJM	Answer	Low	Low	Low	Low	Low
2025	Mikhail[34]	DeepSeek-R1	Not written	Cases from StatPearls	Answer	Low	Low	Low	Low	Low
2025	Moëll[35]	DeepSeek-R1	Not written	Clinical vignettes from the MedQA dataset	Answer	Low	Low	Low	Low	Low
2025	Pan[39]	DeepSeek-R1	Not written	Outpatient cases from hospital	Expert consensus	Low	High	Low	Low	Low
2025	Patel[40]	DeepSeek-R1	Not written	Cases from the Case Book of Sleep Medicine, Third Edition (2019)	Answer	Low	Low	Low	Low	High
2025	Spitzl[41]	DeepSeek-V3	Not written	Fictitious PET/CT reports generated by physicians	Answer	Unclear	High	Low	Low	Low
2025	Tordjman[42]	DeepSeek-R1	Not written	Case challenges in the NEJM	Answer	Low	High	Low	Low	High
2025	Camalan[43]	DeepSeek-V3	Not written	clinical scenarios designed by two otorhinolaryngologists	Answer	Unclear	High	Low	Low	High
2025	Wu[44]	DeepSeek-R1	Not written	Cases published in the NEJM, Mayo Clinic Proceedings, CHEST, Neurology	Answer	Low	Low	Low	Low	High
2025	Li[33]	DeepSeek-V3	Not written	The case reports from EyeRounds	Expert consensus	Low	Low	Low	Low	High