Cureus
Part of SPRINGER NATURE

Open Access Original Article

Large Language Models Perform at Chance Level in the Diagnosis of Pediatric Pneumonia Using Chest Radiographs

Review began 08/21/2025
Review ended 09/08/2025
Published 09/17/2025

DOI: 10.7759/cureus.92596

Justin Gillette [1], Michelle Lu [2], Thomas F. Heston [3, 4]

1. Medical Education and Clinical Sciences, Elson S. Floyd College of Medicine, Washington State University, Spokane, USA 2. Internal Medicine, University of Washington School of Medicine, Seattle, USA 3. Medical Education and Clinical Sciences, Washington State University, Spokane, USA 4. Family Medicine, University of Washington, Spokane, USA

Corresponding author: Thomas F. Heston, theston@uw.edu

Abstract

Introduction

Pneumonia remains a significant cause of morbidity and mortality in children globally. Chest radiographs (CXRs) are widely used to diagnose pediatric pneumonia; however, distinguishing between bacterial and viral etiologies on imaging is a diagnostically challenging task. Large language models (LLMs), particularly those integrated with vision capabilities, have shown promise in preliminary studies for interpreting CXR findings. However, the diagnostic performance of general-purpose LLMs without specialized medical training or add-ons remains poorly understood. This study examined whether such LLMs could independently and reliably distinguish between bacterial, viral, and normal CXRs in pediatric patients.

Methods

We evaluated four publicly available LLMs, such as ChatGPT o3, Claude 3.7 Sonnet, Gemini 2.5 Pro, and Grok 3, on a dataset of 44 pediatric CXRs confirmed by human readers to show bacterial pneumonia (n = 17), viral pneumonia (n = 13), or no abnormality (n = 14). Each image was analyzed twice by each LLM using a standardized prompt, resulting in a total of eight readings per image. Diagnostic accuracy was assessed relative to human expert consensus. Internal consistency was measured by comparing repeated interpretations. A prespecified adaptive stopping rule was employed based on performance futility criteria. Sample size calculations and statistical analyses were conducted using G*Power.

Results

Across all models and CXR types, the average diagnostic accuracy was 31%, consistent with chance-level performance in a three-choice classification task. Accuracy was highest for viral pneumonia (54%) and lowest for normal CXRs (18%). Internal consistency ranged from 46% to 71% across models, indicating unreliable performance. Concordance with human expert interpretation did not exceed 49% for any of the models. Futility criteria were met after 44 cases, prompting early termination of data collection.

Conclusion

General-purpose LLMs currently available to the public are not reliable diagnostic tools for pediatric pneumonia on chest radiographs. Their accuracy is low, particularly in ruling out disease, and their responses lack internal consistency. These findings highlight the risks associated with deploying such models in unsupervised clinical or consumer-facing settings. Future research should focus on purpose-built radiologic AI tools trained on diverse, clinically representative datasets and integrated with clinician oversight to ensure the safe and effective use of these tools.

Categories: Radiology, Internal Medicine, Pediatrics
Keywords: ai in medical imaging, chatgpt, chest-radiography, diagnostic accuracy of ai, large language models, pediatric pneumonia

Introduction

Pneumonia is a leading cause of morbidity and mortality in pediatric populations worldwide, making prompt diagnosis essential [1]. Chest radiographs (CXRs) have been effective in ruling out pneumonia in children, which avoids unnecessary treatment with antibiotic therapy [2]. However, distinguishing between bacterial and viral pneumonia on chest X-rays has proven to be highly challenging, leading some studies to recommend that all children with radiologically confirmed pneumonia receive antibiotic treatment [3]. Even with these findings, clinicians continue to rely on CXRs to differentiate between bacterial and viral pneumonia, thereby guiding their treatment [4]. Since CXRs have a significant impact on the clinical management of pneumonia, it is crucial to enhance diagnostic accuracy.

How to cite this article
Gillette J, Lu M, Heston T F (September 17, 2025) Large Language Models Perform at Chance Level in the Diagnosis of Pediatric Pneumonia Using Chest Radiographs. Cureus 17(9): e92596. DOI 10.7759/cureus.92596

Large language models (LLMs) like ChatGPT, with an “X-ray interpreter” add-on, have shown potential in identifying pathologies such as atelectasis, effusion, emphysema, pneumothorax, pneumonia, and masses on CXR [5]. ChatGPT with the add-on yielded varying results in identifying each pathology, but pneumonia showed the highest pathology accuracy at 91.0%, with a sensitivity of 76.2% and a specificity of 98.7% [5]. There has even been the development of radiology-dedicated LLMs, such as CXR-LLaVA, which have outperformed ChatGPT-4-vision and Gemini-Pro-Vision [6]. There will inevitably be an increase in the use of LLMs to help identify pathologies on CXRs. On October 29, 2024, even Elon Musk encouraged users to submit medical images to Grok for analysis [7]. Both clinicians and patients will increasingly have access to tools that can assist in diagnosis or validate existing findings. This highlights the importance of understanding the capabilities of AI models as well as the differences between them.

Independent evidence shows that general-purpose LLMs struggle with radiologic anatomy; in Part 1 radiologic-anatomy examinations for the Fellowship of the Royal College of Radiologists, ChatGPT-4 performed poorly, indicating significant limitations in recognizing normal radiological anatomy [8]. While prior studies have shown promising results using LLMs enhanced with specialized add-ons or radiology-specific tools, we aim to assess whether baseline, general-purpose models can independently provide valuable diagnostic insights without the need for additional modifications or integrations. To evaluate the efficacy of LLMs in diagnosing pneumonia from CXRs, we evaluated four leading LLMs, such as ChatGPT o3, Claude 3.7 Sonnet, Gemini 2.5 Pro, and Grok 3-for their ability to analyze pediatric CXRs categorized as showing bacterial pneumonia, viral pneumonia, or no abnormalities. The popularity of these LLMs increases the likelihood that patients could upload their CXRs to have a second opinion on their diagnoses. The LLMs will be evaluated for the accuracy of their diagnoses compared to those of human readers and their consistency when shown the exact CXR multiple times.

This study utilizes images from a public domain dataset that includes pediatric CXRs, which either display bacterial pneumonia, viral pneumonia, or a normal CXR [9,10]. These CXRs were obtained from retrospective cohorts of pediatric patients aged one to five years from Guangzhou Women and Children’s Medical Center in Guangzhou, China. These CXRs were interpreted to confirm a diagnosis and make treatment referrals [10].

The objective of this study was to evaluate the diagnostic accuracy and consistency of four leading general-purpose LLMs in classifying pediatric chest radiographs as bacterial pneumonia, viral pneumonia, or normal. The results of this study can help inform clinicians on how LLMs can be utilized or improved to enhance diagnostic accuracy. Additionally, patients should be aware of the potential for misdiagnosis when uploading their imaging and of which LLMs are the most consistent.

Materials And Methods

Study population

From a public dataset of 5,856 CXRs, 44 were randomly sampled and analyzed prior to meeting the adaptive futility boundary. Inclusion criteria, including patients aged one to five years, anterior-posterior image view, sufficient quality, and expert-confirmed diagnosis, were predefined by the public domain dataset curators. These images were evenly distributed across three diagnostic categories: bacterial pneumonia (n = 17), viral pneumonia (n = 13), and normal (n = 14). Each image was independently analyzed twice by four general-purpose LLMs, resulting in 352 total diagnostic interpretations (Figure *1*). As each CXR was interpreted eight times, with non-unanimous predictions being common, we did not assign a single predicted diagnosis per image and did not compute standard cross-tabulation matrices. All models received the same standardized prompt, without prefix, suffix, or role assignment, consistent with prior studies demonstrating that prompt structure significantly affects diagnostic output from LLMs [8]. The full prompt was: “Look at this image carefully. Analyze it thoroughly. This is a CXR of a pediatric patient with suspected pneumonia. Based on the CXR, is the pneumonia bacterial or viral? Or is the CXR normal? Respond in one word: bacterial, viral, or normal.”

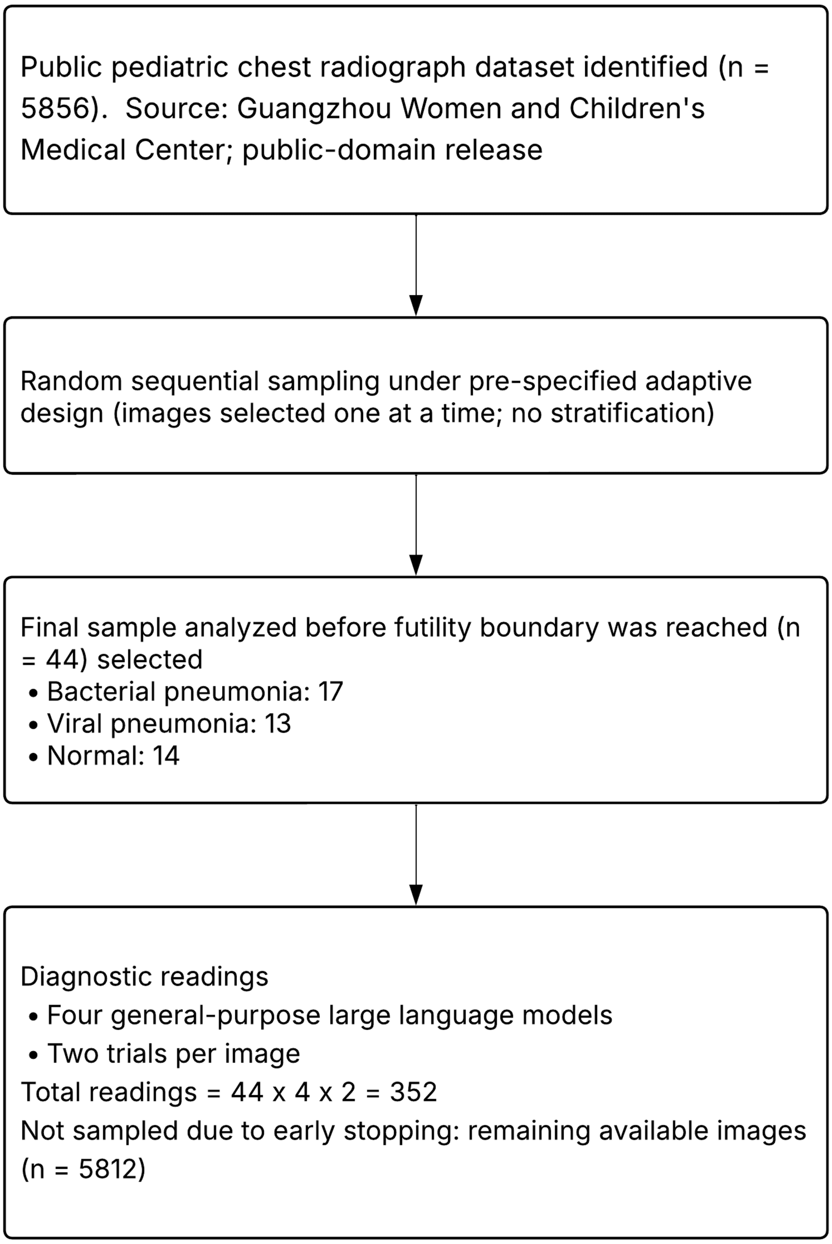
Public pediatric chest radiograph dataset identified (n = 5856). Source: Guangzhou Women and Children's Medical Center; public-domain release
Random sequential sampling under pre-specified adaptive design (images selected one at a time; no stratification)
Final sample analyzed before futility boundary was reached (n = 44) selected
• Bacterial pneumonia: 17
• Viral pneumonia: 13
• Normal: 14
Diagnostic readings
• Four general-purpose large language models
• Two trials per image
Total readings = 44 x 4 x 2 = 352
Not sampled due to early stopping: remaining available images (n = 5812)

FIGURE 1: Standards for reporting diagnostic accuracy studies flow diagram

A public pediatric chest radiograph dataset (n = 5,856) was identified. Images were randomly sampled sequentially under a prespecified adaptive design, and the analysis stopped when the futility boundary was crossed after 44 images (bacterial 17, viral 13, and normal 14). Each analyzed image was evaluated twice by four general-purpose large language models, yielding 352 total readings. The remaining available images (n = 5,812) were not sampled due to early stopping.

Statistical analysis

Sample size determinations were carried out in G*Power, version 3.1.9.7 (Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany) [11]. Three analyses yielded N = 122 for test-retest consistency (Bowker's marginal-homogeneity χ^2; df = 3; w = .30; α = .05; 1-β = .80), N = 48 for AI-human concordance (χ^2 goodness-of-fit; df = 1; w = .408; α = .05; 1-β = .80), and N = 24 for comparing accuracy across the four programs via repeated measures ANOVA (f = .25; α = .05; 1-β = .80; measurements = 4; ρ = .50; ε = 1). The largest required sample size (N = 122) was chosen to ensure adequate power for all planned analyses. An adaptive trial design was employed, with one interim analysis scheduled after approximately one-third of the target sample had been evaluated [12,13]. Prespecified futility criteria were defined such that if the AI models' strict accuracy on the interim cohort remained within 5% of chance (i.e., 33%) and there was a low likelihood of achieving a clinically meaningful accuracy (>50%) with further data, then data collection would be stopped early. This approach minimized unnecessary evaluation of additional CXRs once evidence of futility was established. Although model responses were required to conform to a single-word format (“bacterial,” “viral,” or “normal”), no responses were excluded due to formatting issues; all outputs in the final dataset were interpretable and required no post hoc disambiguation. Although diagnostic categories were not perfectly balanced, case selection was based on image quality, patient age, and confirmed diagnosis rather than predetermined quotas, preserving the real-world nature of the dataset. Confidence intervals for diagnostic accuracy were calculated using the Wilson score interval without continuity correction [14].

Compliance with reporting standards

This study adhered to the Standards for Reporting Diagnostic Accuracy Studies (STARD 2015) guidelines for reporting diagnostic accuracy research involving retrospective imaging data [15]. A flow diagram and structured reporting elements were used to enhance transparency and reproducibility.

Results

A total of 44 pediatric chest radiographs were analyzed between May and June 2025, after which the adaptive trial’s prespecified futility criteria were met, and data collection was stopped early.

Overall diagnostic accuracy

Across all four AI engines, the strict accuracy - defined as both model responses per image matching the human reference standard - averaged 31%, closely matching random-guess performance for a three-category task (chance = 33%). Individual model accuracies were 23% for Grok 3 (95% confidence interval (CI): 15%-33%), 27% for ChatGPT o3 (95% CI: 19%-37%), 45% for Claude 3.7 Sonnet (95% CI: 35%-55%), and 30% for Gemini 2.5 Pro (95% CI: 21%-40%; all CIs calculated using Wilson score method). One-way repeated measures ANOVA revealed no significant difference in accuracy between models (F(3, 129) = 2.178, p = 0.093). A statistical comparison of accuracies revealed significant differences between viral and normal pneumonia (z = 5.52, p < 0.001) and between viral and bacterial pneumonia (z = 4.46, p < 0.001). In contrast, the difference between bacterial and normal pneumonia was not significant (z = 1.39, p = 0.165).

AI-human concordance

Concordance with human radiologists was uniformly poor, with none of the models achieving better than 49% agreement and an overall average of 31%. χ^2 goodness-of-fit testing confirmed that this concordance did not differ significantly from chance performance (χ^2 = 0.819, df = 1, p = 0.366). The overall accuracy was low at 31% with a 95% CI of 25%-38%. For normal CXR findings, the accuracy was 16% (8%-28%); for viral, 54% (40%-67%); and for bacterial, 27% (17%-37%) (Table 1).

CXR type	Accuracy (95% CI)	Interpretation
Normal	16% (8%–28%)	Substantially below chance; suggests unreliable performance in ruling out pathology
Viral	54% (40%–67%)	Slightly above chance; limited discriminative capability
Bacterial	27% (17%–37%)	Below chance; inconsistent differentiation from other etiologies
Overall	31% (25%–38%)	Within the range of random chance

TABLE 1: Accuracy of LLMs by chest radiograph (CXR) type

Accuracies of LLMs are given with 95% CIs. Accuracy refers to strict agreement between model output and human-confirmed reference standard.

Internal consistency

Test-retest consistency was evaluated using Bowker's marginal-homogeneity test (χ^2 = 2.622, df = 3, p = 0.454). While this indicated no systematic bias in the direction of disagreements between first and second readings, overall agreement was only 60.2%, demonstrating substantial inconsistency in model responses. Individual model consistency ranged from 46% to 71%, which is insufficiently reliable for clinical use.

Adaptive stopping

Interim analyses demonstrated that even under overly optimistic performance assumptions, final accuracy could not have risen appreciably above chance levels. Accordingly, the adaptive design’s stopping rule for futility was satisfied at n = 44, obviating the need to reach the initially planned N = 122 evaluations.

These findings indicate that baseline, general-purpose LLMs in their current form perform at chance levels for differentiating bacterial versus viral pneumonia (or normal) on pediatric CXRs and lack sufficient consistency to be considered reliable diagnostic aids.

Discussion

The results of this study indicate a limited concordance between LLMs and human readers when interpreting CXRs. In addition, the LLMs showed low consistency in their readings when presented with the same CXR a second time. In a study by Yao et al., pediatric CXRs with confirmed pneumonia diagnoses were evaluated by four radiologists to assess their diagnostic performance [16]. The four radiologists (two attendings and two residents) in this study achieved a recall (sensitivity) range of 0.81-0.95 for normal CXRs, 0.49-0.71 for bacterial pneumonia, and 0.15-0.21 for viral pneumonia [16]. In contrast, all four LLMs evaluated in this study - general-purpose models without medical fine-tuning or domain-specific training - demonstrated low accuracy, even when identifying normal CXRs.

Their inconsistent responses when asked to reassess the same images suggest a lack of diagnostic confidence and stability. This variability raises concerns about susceptibility to bias, particularly if users repeatedly prompt the model until it produces an answer that aligns with their preconceived diagnosis. Given the models’ poor concordance with human readers and lack of internal consistency, there is a significant risk in relying on them independently, especially by users who may not fully understand the limitations and appropriate use of these tools.

The notably higher accuracy for viral pneumonia (54%) compared with bacterial pneumonia (26%) and normal CXRs (18%) suggests that these general-purpose LLMs may have differential pattern recognition capabilities across diagnostic categories. Viral pneumonia typically presents with bilateral, diffuse, or interstitial patterns that may be more readily identifiable by vision-enabled models compared to the focal consolidations characteristic of bacterial pneumonia. The inferior performance in identifying normal CXRs (18%) is concerning from a clinical perspective, as this represents the models' inability to rule out disease - a critical function in screening and diagnostic workflows. This pattern may reflect training data biases, where viral pneumonia images might be more prevalent in publicly available datasets, or it could indicate that the diffuse, bilateral patterns of viral disease align better with the pattern recognition algorithms underlying these models. The low accuracy for normal CXRs also suggests that these models struggle with the nuanced task of distinguishing subtle normal variations from pathology. This task requires sophisticated clinical judgment developed through extensive training and experience by human radiologists. These findings highlight the complexity of medical image interpretation and underscore the need for purpose-built, medically trained AI systems to achieve reliable diagnostic performance.

Since these publicly available LLMs cannot read as accurately as human radiologists, they would not provide much value if used alone. AI has been proposed as a potential solution to enhance access to medical imaging in low- and middle-income countries [17]. Many physicians ordering imaging may be tempted to use publicly available LLMs because they are easily accessible, have minimal barriers to use, and can quickly yield results. However, the LLMs used in this study are not suitable for practical or clinical use on their own, as their low accuracy significantly limits their reliability and effectiveness. Even if these LLMs are used with radiologist supervision, previous studies have shown that AI predictions with significant errors can lead to adverse treatment effects, with radiologists struggling to differentiate accurate and inaccurate outputs [18]. This highlights the need for accurate AI models that are specifically trained in medical imaging. Further research is needed to determine the most effective way to integrate AI into radiologists' workflows, enhancing diagnostic accuracy and efficiency while minimizing the addition of complexity or burden to their work.

While radiologists will be aware of the limitations of LLMs in interpreting medical imaging, the public will still be interested in using easily accessible LLMs to obtain a second opinion on their diagnoses. Many electronic medical record systems now grant patients immediate access to their imaging and lab results, often before a physician has reviewed them. Increasingly, patients prefer this real-time release of information, even if it has not yet been interpreted by a healthcare practitioner [19]. When patients gain access to their CXRs before a physician interprets them, many may experience anxiety and seek early interpretations from publicly available LLMs. This practice is concerning because LLMs could heighten patient distress with false positives or provide a false sense of reassurance in the case of false negatives. Since many patients are unaware of the limitations and low diagnostic reliability of AI, LLM-generated results may cause them to question or distrust the radiologist’s interpretation that they later receive.

The potential for misdiagnosis, misplaced trust, and unclear accountability demonstrates why these LLMs are not ready for patient care. Overreliance on AI tools for interpreting chest X-rays may give clinicians a false sense of diagnostic certainty, particularly for conditions such as bacterial pneumonia. This could lead to shortcuts in clinical reasoning, such as forgoing additional testing due to the convenience of a rapid AI-generated interpretation. Moreover, the issue of accountability remains unresolved. If an LLM produces an incorrect diagnosis, it is unclear whether responsibility lies with the clinician or the developers of the tool. These concerns highlight the urgent need for clear regulatory standards and rigorous validation before such models are used in patient care. Clinicians must understand both the limitations of these tools and the boundaries of their legal and ethical responsibility. Given the rapid advancement of AI models, evolving ethical concerns will require that regulations and standards be continually reassessed to safeguard patients and promote effective clinical practice [20].

Recent work has emphasized the potential role of agentic AI systems - autonomous agents that can plan, reason, and act within defined clinical boundaries - to improve diagnostic workflows in radiology [21]. While our study focused on general-purpose LLMs operating in passive diagnostic mode, the structured deployment of agentic AI may eventually address many limitations identified here. For instance, such systems could incorporate contextual clinical data, manage uncertainty by deferring to human experts, and engage in multistep diagnostic reasoning. However, realizing this potential will require robust governance, including privacy safeguards, interoperability standards, continuous performance monitoring, and phase-wise clinical integration. The radiology community must proactively evaluate and validate these tools to responsibly harness their benefits while safeguarding patient care.

Limitations

Several limitations should be considered in this study. A key limitation is the lack of transparency regarding the training data used for each LLM. Without knowing the specific images or datasets on which the models were trained, it is difficult to interpret the differences in their performance or understand the underlying factors contributing to their results. Even in authorized medical AI software, most products do not publish information on training data collection and population characteristics [22]. This uncertainty limits our ability to evaluate the training data and raises concerns about how well these models will generalize to different patient populations or clinical settings. LLMs function as black boxes, offering no insight into their decision-making processes [23]. To effectively compare LLMs to human radiologists, it is essential to understand the reasoning processes behind each diagnosis and how these processes may differ between models and humans. Another limitation of this study is that the LLMs were not evaluated in a real clinical environment. In real clinical practice, CXRs are not interpreted in isolation. Radiologists typically incorporate a range of complementary clinical information, such as patient history, physical examination findings, laboratory results, and prior imaging studies, to arrive at a diagnosis. This context is critical because many radiographic findings are nonspecific and require clinical correlation to determine their significance. By contrast, the LLMs in this study analyzed the CXR without access to any supporting clinical data, which does not reflect the way medical imaging is interpreted in real-world settings.

In addition, the dataset used in this study was sourced from a single institution, which introduces more limitations [24]. Imaging protocols, patient demographics, disease prevalence, and equipment settings can vary widely across hospitals, regions, and populations. As a result, the performance of these models on this dataset may not generalize to other clinical environments. This raises concerns about the external validity and real-world applicability of the findings, especially in more diverse or resource-limited settings where imaging conditions and patient profiles may differ substantially.

Conclusions

The results of this study show that publicly available LLMs in their current state should not be used to evaluate pediatric CXRs for pneumonia. Their diagnostic accuracy remains significantly lower than that of human radiologists, and the low concordance between models indicates limited reliability and internal consistency in their diagnostic reasoning. If LLMs are to be used in a clinical setting, there must be extensive oversight by a radiologist in these early stages. Both clinicians and patients must be aware of the current limitations of these models and avoid relying on them as standalone diagnostic tools. Future work should focus on advancing AI tools that are specifically designed and trained for radiologic applications. These models should be developed using diverse and representative datasets that accurately reflect various patient populations to ensure broad generalizability. As LLMs continue to improve, their integration into clinical workflows must be accompanied by rigorous oversight to ensure the safe and effective use of these tools in patient care.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Thomas F. Heston, Justin Gillette

Acquisition, analysis, or interpretation of data: Thomas F. Heston, Justin Gillette, Michelle Lu

Drafting of the manuscript: Thomas F. Heston, Justin Gillette

Critical review of the manuscript for important intellectual content: Thomas F. Heston, Justin Gillette,

Michelle Lu

Supervision: Thomas F. Heston

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Pediatric pneumonia. (2025). Accessed: May 16, 2025: https://emedicine.medscape.com/article/967822-overview.
2. Lipsett SC, Monuteaux MC, Bachur RG, Finn N, Neuman MI: Negative chest radiography and risk of pneumonia. Pediatrics. 2018, 142:e20180236. 10.1542/peds.2018-0236
3. Virkki R, Juven T, Rikalainen H, Svedström E, Mertsola J, Ruuskanen O: Differentiation of bacterial and viral pneumonia in children. Thorax. 2002, 57:438-41. 10.1136/thorax.57.5.438
4. Geanacopoulos AT, Lipsett SC, Hirsch AW, Monuteaux MC, Neuman MI: Impact of viral radiographic features on antibiotic treatment for pediatric pneumonia. J Pediatric Infect Dis Soc. 2022, 11:207-13. 10.1093/jpids/piab132
5. Ostrovsky AM: Evaluating a large language model's accuracy in chest X-ray interpretation for acute thoracic conditions. Am J Emerg Med. 2025, 93:99-102. 10.1016/j.ajem.2025.03.060
6. Lee S, Youn J, Kim H, Kim M, Yoon SH: CXR-LLaVA: a multimodal large language model for interpreting chest X-ray images. Eur Radiol. 2025, 35:4374-86. 10.1007/s00330-024-11339-6
7. Elon Musk invites submission of medical images to xAI's Grok. (2025). Accessed: May 19, 2025: https://www.auntminnie.com/imaging-informatics/artificial-intelligence/article/15707188/elon-musk-invites-submission-....
8. Sarangi PK, Datta S, Panda BB, Panda S, Mondal H: Evaluating ChatGPT-4's performance in identifying radiological anatomy in FRCR part 1 examination questions. Indian J Radiol Imaging. 2025, 35:287-94. 10.1055/s-0044-1792040
9. Kermany D, Zhang K, Goldbaum M: Labeled optical coherence tomography (OCT) and chest X-ray images for classification [IN PRESS]. Mendeley Data. 2018, 10.17632/rscbjbr9sj.2
10. Kermany DS, Goldbaum M, Cai W, et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018, 172:1122-1131. 10.1016/j.cell.2018.02.010
11. Faul F, Erdfelder E, Lang AG, Buchner A: G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007, 39:175-91. 10.3758/bf03193146
12. Cohen J: Statistical Power Analysis for the Behavioral Sciences. Routledge, New York, NY; 1988. 10.4324/9780203771587
13. Maxwell SE, Delaney HD, Kelley K: Designing Experiments and Analyzing Data: A Model Comparison Perspective. Routledge, New York, NY; 2017. 10.4324/9781315642956
14. Newcombe RG: Two-sided confidence intervals for the single proportion: comparison of seven methods . Statist Med. 1998, 17:857-72. 10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E
15. Bossuyt PM, Reitsma JB, Bruns DE, et al.: STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Radiology. 2015, 277:826-32. 10.1148/radiol.2015151516
16. Yao D, Xu Z, Lin Y, Zhan Y: Accurate and intelligent diagnosis of pediatric pneumonia using X-ray images and blood testing data. Front Bioeng Biotechnol. 2023, 11:1058888. 10.3389/fbioe.2023.1058888
17. Frija G, Blažić I, Frush DP, Hierath M, Kawooya M, Donoso-Bach L, Brkljačić B: How to improve access to medical imaging in low- and middle-income countries ?. EClinicalMedicine. 2021, 38:101034. 10.1016/j.eclinm.2021.101034
18. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P: Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med. 2024, 30:837-49. 10.1038/s41591-024-02850-w
19. Steitz BD, Turer RW, Lin CT, et al.: Perspectives of patients about immediate access to test results through an online patient portal. JAMA Netw Open. 2023, 6:e233572. 10.1001/jamanetworkopen.2023.3572
20. Geis JR, Brady AP, Wu CC, et al.: Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Radiology. 2019, 293:436-40. 10.1148/radiol.2019191586
21. Datta S, Sarangi PK: From chatbots to agentic workflows: ensuring responsible deployment of large language models in radiology [IN PRESS]. Indian J Radiol Imaging. 2025, 10.1055/s-0045-1811264
22. Fehr J, Citro B, Malpani R, Lippert C, Madai VI: A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. Front Digit Health. 2024, 6:1267290. 10.3389/fdgth.2024.1267290
23. Ullah E, Parwani A, Baig MM, Singh R: Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. Diagn Pathol. 2024, 19:43. 10.1186/s13000-024-01464-7
24. Rajpurkar P, Irvin J, Ball RL, et al.: Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018, 15:e1002686. 10.1371/journal.pmed.1002686