BMJ Health & Care Informatics

Performance of large language models on advocating the management of meningitis: a comparative qualitative study

Urs Fisch ,[1] Paulina Kliem,[2] Pascale Grzonka,[2] Raoul Sutter[1,2,3]

To cite: Fisch U, Kliem P, Grzonka P, *et al*. Performance of large language models on advocating the management of meningitis: a comparative qualitative study. *BMJ Health Care Inform* 2024;**31**:e100978. doi:10.1136/bmjhci-2023-100978

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100978).

Received 24 November 2023
Accepted 15 January 2024

[1]Department of Neurology, University Hospital Basel, Basel, Switzerland
[2]Clinic for Intensive Care Medicine, University Hospital Basel, Basel, Switzerland
[3]Medical Faculty, University Basel, Basel, Switzerland

Correspondence to
Dr Urs Fisch; urs.fisch@usb.ch

ABSTRACT

Objectives We aimed to examine the adherence of large language models (LLMs) to bacterial meningitis guidelines using a hypothetical medical case, highlighting their utility and limitations in healthcare.

Methods A simulated clinical scenario of a patient with bacterial meningitis secondary to mastoiditis was presented in three independent sessions to seven publicly accessible LLMs (Bard, Bing, Claude-2, GTP-3.5, GTP-4, Llama, PaLM). Responses were evaluated for adherence to good clinical practice and two international meningitis guidelines.

Results A central nervous system infection was identified in 90% of LLM sessions. All recommended imaging, while 81% suggested lumbar puncture. Blood cultures and specific mastoiditis work-up were proposed in only 62% and 38% sessions, respectively. Only 38% of sessions provided the correct empirical antibiotic treatment, while antiviral treatment and dexamethasone were advised in 33% and 24%, respectively. Misleading statements were generated in 52%. No significant correlation was found between LLMs' text length and performance (r=0.29, p=0.20). Among all LLMs, GTP-4 demonstrated the best performance.

Discussion Latest LLMs provide valuable advice on differential diagnosis and diagnostic procedures but significantly vary in treatment-specific information for bacterial meningitis when introduced to a realistic clinical scenario. Misleading statements were common, with performance differences attributed to each LLM's unique algorithm rather than output length.

Conclusions Users must be aware of such limitations and performance variability when considering LLMs as a support tool for medical decision-making. Further research is needed to refine these models' comprehension of complex medical scenarios and their ability to provide reliable information.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Large language models (LLMs) have demonstrated proficiency in responding to medical licensing examination-level queries and shown aptitude in accurate medical triage decision-making. However, performance with knowledge-testing scenarios is not necessarily indicative of effectiveness in real-world medical contexts.

WHAT THIS STUDY ADDS

⇒ This investigation presents a qualitative analysis of the performance of seven publicly accessible LLMs, using a stepwise presentation of a hypothetical bacterial meningitis case reflecting a real-world scenario. While LLMs generally offered helpful triage and diagnostic advice, there were significant discrepancies in their recommendations for treatment and specific diagnostic work-ups. Moreover, the generation of misleading statements and variability in performances between different sessions were observed among individual LLMs.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study highlights the current capabilities of LLMs in handling real-world medical emergency situations and identifies areas of future research, such as enhancing LLMs' understanding of complex medical scenarios and their capacity for delivering reliable and deterministic information.

INTRODUCTION

Large language models (LLMs) are powerful artificial intelligence (AI) models trained on extensive text data to generate human-like text. They can interpret user-generated textual instructions (prompts) and respond immediately with the contextually most appropriate response based on probabilistic computations learnt during their training. Lately, several LLMs were released to the public, attracting substantial attention for their chat-like interfaces requiring no technical prerequisites.

Recently, both trained and untrained LLMs have shown proficiency in handling medical licensing examination-level questions and demonstrated the ability to make rapid and accurate judgments in medical triage and diagnosing or provide helpful information to patients, underscoring their potential applicability in the healthcare sector.[1–6] However, the ability to perform well in knowledge-testing

vignettes does not fully reflect the needs of real-world medical settings which demand parallel work-up and nuanced decision-making on the basis of sometimes incomplete information. Considering that physicians already frequently use internet resources for diagnostic decisions and treatment options and that not all hospitals may have free access to the medical literature, it is likely that LLMs will be increasingly used as potential aids in clinical practice.[7–9] However, a deeper understanding of their potential and limitations is essential for an appropriate use.[10–12]

This study explored the potentials and limitations of current LLMs by presenting these models with a predefined hypothetical but typical scenario of a patient with acute bacterial meningitis. The aim was to analyse their performance and alignment with good clinical practice and established medical guidelines regarding suggested diagnostic and treatment measures. Bacterial meningitis was chosen for its life-threatening nature, urgency required in diagnosis and treatment and the range of differential diagnoses it involves, making it ideal for assessing the performance of LLMs in a realistic and high stakes medical scenario.

METHODS

Seven publicly accessible LLMs were evaluated between 5 and 8 August 2023: Bard by Google, Bing by Microsoft, generative pre-trained transformer (GTP)-3.5 by OpenAI, GTP-4 by OpenAI (accessed via Poe (Quora)), Claude-2 by Anthropic PBC (accessed via Poe), pathways language model (PaLM) 2 chat-bison-001 by Google (accessed via Poe) and Llama-2-70b by Meta Platforms (accessed via Poe).

Each LLM was presented with the same hypothetical scenario of a patient presenting with symptoms of acute bacterial meningitis (as outlined below) three times within 3 days. The actual diagnosis was not provided. For the LLM Bard, the settings were chosen to inhibit inter-session information storage. All other LLMs claimed that they are incapable of storing user information between sessions. Each session was initiated with a context clearance of previous conversations.

Hypothetical scenario of a patient with acute bacterial meningitis

The patient vignette described a clinical scenario of a patient with acute symptoms due to pneumococcal meningitis secondary to mastoiditis without providing definite diagnosis. The text of the inputted case vignette and the subsequent follow-up queries consisted of five text blocks that were predefined and presented unchanged to each LLM in every session (online supplemental table 1). Given that the performance of LLMs is heavily influenced by prompting,[13] the initial question began with a contextualisation wherein the LLM was asked to act as an 'experienced medical assistant' and the user was identified as a 'junior medical doctor' seeking advice for a 52-year-old female patient suffering from severe headache and confusion, followed by an open-ended question about the next steps. This prompt engaged all LLMs in a conversation about the hypothetical case. Second, a detailed vignette was presented, depicting the medical history (notably acute headache and confusion, a history of diabetes type 2 and migraine), vital signs (tachycardia and fever) and prominent abnormal clinical findings (ie, a Glasgow Coma Scale (GCS) of 12 with lethargy, disorientation, fast downward drift of extremities, absence of stiff neck, signs of inflammatory skin of the right mastoid), followed by the open-ended request for a detailed step-by-step recommendation of how to proceed. Third, two closed-ended questions were asked: (1) if a computer tomography (CT) scan of the head needs to be awaited before lumbar puncture (LP) and (2) if administration of antibiotics should be delayed until LP has been performed. Fourth, the exact dosages of antibiotics were asked. Fifth, an open-ended question was asked about any other considerations regarding the treatment or work-up.

The case was created to reflect clinical reality and not a medical license examination question, meaning that information was presented stepwise and reflected a realistic clinical case where not all typical signs and symptoms are necessarily present from the beginning. For example, neck stiffness has shown to have a low sensitivity and as such, its absence cannot rule out meningitis.[14] A search for an infectious focus is crucial and patients should be examined for otitis media or mastoiditis.[15] By this design we aimed to challenge the LLMs in multiple aspects, including good clinical practice, possible differential diagnoses and consideration of risk factors and comorbidities, such as age, diabetes and migraine, for diagnosis and treatment.

Evaluation of LLM performance

The Infectious Diseases Society of America (IDSA) and the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) guidelines were chosen as references as they have previously both been shown in a systematic review to be excellent clinical management guidelines for bacterial meningitis with multinational validity (online supplemental table 2, right column).[14 16 17]

Individual responses from the LLMs underwent two temporally separated qualitative assessments (accomplished vs unaccomplished) of predefined tasks (online supplemental table 2, middle column) in adherence with good clinical practice and the reference guidelines.[14–18] Accomplished tasks were summarised to a qualitative performance summary. Response consistency was defined as the percentage of responded tasks that were assessed identically (regardless of accomplished or unaccomplished) across all sessions of an individual LLM. In cases where an LLM declined to respond to a question, the corresponding tasks were excluded from the assessment.

As the two reference guidelines differently define criteria for imaging before LP (ie, according to the IDSA guideline, a scan of the brain would be required as

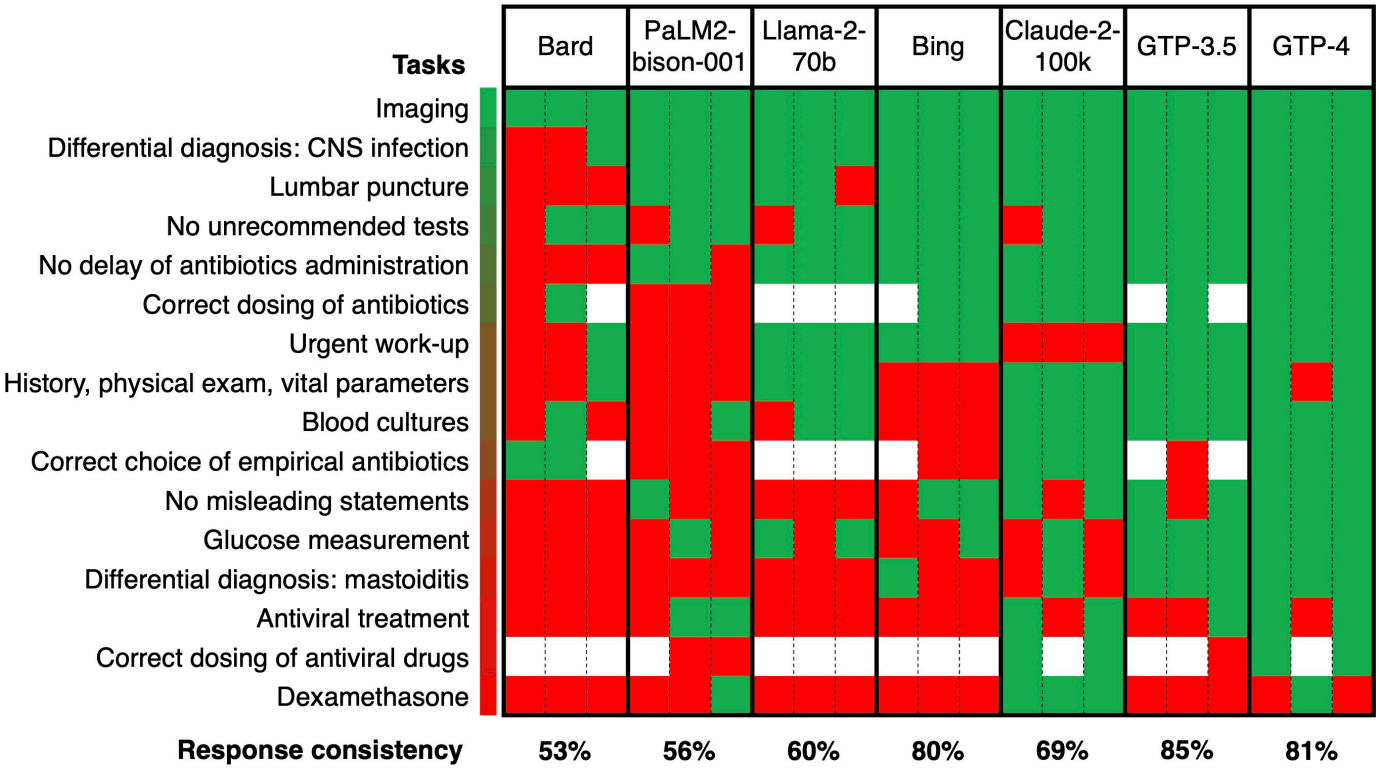

Figure 1 Qualitative assessment of large language models (LLMs) performance on a case of bacterial meningitis. Each LLM was tested three times with a standardised case vignette (individual sessions separated by dashed lines). Accomplished tasks are marked in green in decreasing order of agreement among all LLMs, while unaccomplished tasks are highlighted in red. White boxes represent tasks where the model either declined to respond or no additional information could be provided due to gaps in previous responses. Response consistency was defined as identically assessed responded tasks across different sessions of a single LLM. CNS, central nervous system.

the patient expresses any altered mental status and has downward drift of extremities, whereas according to the ESCMID guideline, a scan of the brain is not mandatory with a GCS>10) and maximal allowed delay to start antibiotics, these aspects were not included in the qualitative performance summary.[14 16]

Statistics

Descriptive statistics with numbers and percentages and the two-sided Pearson correlation coefficient were used where appropriate (R, V.4.3.1). Due to the principally qualitative aim of this study, a statistical comparison between the LLMs was not intended.

RESULTS

The individual responses of all 21 sessions of the seven LLMs are summarised in figure 1. We noticed marked differences in the qualitative performance summary between different LLMs and to a lesser extent also between different sessions of individual LLMs. The response consistency ranged from 53% to 85%. LLMs with low numbers of accomplished tasks also had low response consistency. Among all the LLMs evaluated, GPT-4 demonstrated the most consistent performance, effectively addressing almost all tasks and having a high response consistency across all tasks and responses. Exemplary transcripts of the first conversations with Bard and GTP-4 are shown in online supplemental material.

The word count of individual LLMs sessions varied significantly, ranging from 325 (PaLM 2 chat-bison-001) to 2045 (GTP-3.5), with an average of 1270 words (standard deviation 477). There was no significant correlation (r=0.29, p=0.20) between the total length of individual LLM responses and the summative performance of accomplished tasks, indicating that simply generating more text output does not necessarily lead to improved performance.

Suggested differential diagnoses and recommended diagnostic work-up

In 62% of the sessions, LLMs suggested an urgent work-up without direct prompting. In 57% of sessions, they recommended measuring vital parameters, taking the patient's history and performing a physical examination as initial steps. Furthermore, in 90% of the sessions, the LLMs accurately suspected a central nervous system (CNS) infection as a possible cause of the patient's symptoms. However, only 38% of the responses mentioned mastoiditis as a potential underlying cause or suggested correspondent diagnostic procedures (imaging with purpose of investigating mastoiditis, otoscopy, ear–nose–throat consultation). The most frequently mentioned differential diagnoses

were stroke (86%), followed by intracranial/subarachnoid haemorrhage and brain tumour (both 48%). Other proposed differential diagnoses were migraine (19%), metabolic/endocrine disbalances (19%), medication side effects (10%), non-CNS infections (10%), severe hypertension (5%), drug intoxication (5%) and neurodegenerative disorders (5%).

Regarding diagnostic work-up, cranial imaging was recommended in 100% of sessions, LP in 81% and blood cultures in 62%. Blood glucose measurement in the diabetic patient with altered mental status was suggested in 53%. Unrecommended tests by the IDSA and ESCMID guidelines (eg, electroencephalogram, electrocardiogram, chest radiography) were proposed in 19% of sessions as an initial work-up.

In 43% of responses, LLMs stated that a cranial CT scan is necessary before LP, while 14% suggested to perform an LP without CT scan and another 43% gave unclear answers. Only three LLMs (GTP-3.5, Claude-2, GTP-4) provided a case-specific rationale for their recommendation (92% responses suggested CT scan before LP). Due to different definitions of criteria for cranial imaging before LP in the reference guidelines and maximal allowed delay to start antibiotics,[14 16] these aspects were not included in the qualitative performance summary displayed in figure 1.

Recommended treatment

Regarding treatment, 81% of responses stated that rapid administration of antibiotics is necessary. The correct choice of empirical antibiotic treatment, consisting of a third-generation cephalosporin with ampicillin (alternatives: amoxicillin, penicillin G) with or without vancomycin, was provided in 38%, and of those, almost 90% with correct dosing.[14 16] Another 29% provided an incomplete choice of antibiotic treatment and 33% declined to comment on any choice of antibiotics. In 33% of the sessions, antiviral treatment was considered with approximately half of them providing correct dosing. Dexamethasone administration was recommended in 24% of all responses.

Misleading statements

Misleading statements were identified in 52% of the sessions, such as performing an LP to relieve intracranial pressure or carrying it out prior to imaging in order to facilitate image interpretation; administering prophylactic antiseizure medication or giving benzodiazepines for sedation; adjusting ceftriaxone dosage based on age, weight and kidney function or administering dexamethasone for meningococcal meningitis; the presence of a stiff neck and Kernig's sign (while the vignette stated that these were absent); or the misinterpretation of mastoiditis as herpes zoster ophthalmicus.

DISCUSSION

This study investigated qualitative performance characteristics of different LLMs when challenged with a hypothetical clinical case of an adult patient with bacterial meningitis and revealed marked discrepancies between the LLMs. This reflects both the potentials and limitations of these models when used as a guidance for medical work-up and treatment.[9] The differences in qualitative performances observed among the LLMs did not demonstrate a correlation with the length of their respective outputs. This suggests that the performance variations can be attributed to the unique algorithmic designs of each LLM rather than their quantitative output.

CNS infection was identified as a probable cause among other differential diagnosis in the majority of cases and almost all LLMs succeeded in identifying and recommending appropriate investigations, including cranial imaging and LP. A fair proportion underscored the need for urgent diagnostics and antibiotic treatment. These results align with previous findings demonstrating a satisfactory performance of GTP-3 (the predecessor of GTP-3.5) in terms of triage and reasoning on differential diagnoses and the high performance of GTP-4 in diagnostic case challenges.[4 19–21] Our study expands on these findings by examining an additional five LLMs which were not available at the time of the previous studies.

Our investigation also highlights limitations of most LLMs regarding their understanding of case complexity and their ability to link different disease entities. For instance, the identification of mastoiditis as an underlying cause was mentioned infrequently, as were blood glucose measurements, drawing blood cultures, considerations of empirical antiviral treatment and the administration of dexamethasone. The considerable heterogeneity in the responses of individual LLMs, despite standardised prompts, raises further concerns about their reliability and consistency. The presentation of misleading statements in more than half of the LLM sessions underscores the potential risk that comes along with their use for critical medical decision-making, especially in complex, life-threatening and time-sensitive situations, such as with bacterial meningitis. Such challenges must be addressed in future research when developing tools on the basis of LLMs for medical purposes.[10–12]

Most LLMs' inability to provide definitive guidance on whether to conduct a cranial CT scan before an LP might be due to the differences in the guidelines.[14 16] However, the lack of clear direction in many LLM responses could also suggest an insufficiency in handling complex clinical situations where there is a need for reasoned decision-making. This finding may be viewed in the context of the research gap between healthcare AI development and the challenge of its validation and implementation in real-world clinical settings.[22–24]

Limitations

Our study has several limitations. Most importantly, none of the LLMs was designed to assist in medical diagnostics and treatment and most correctly included respective disclaimers. However, as LLMs are powerful, new and easily accessible AI tools, it is highly probable that they will find increasing use in the health sector, which

justifies studying their reliability and applicability.[1–6] Further, prompting has significant influence on the result.[13] While our study did not explore the impact of different prompting strategies, we used standardised prompts, which included contextualisation and step-by-step reasoning, to ensure comparability between LLMs. Although we evaluated the LLMs' intuitive assessment of the scenario's urgency, we did not directly inquire this in the prompts. In addition, the selection of tasks for the qualitative assessment was unweighted and focused on important initial management steps, while other aspects, such as laboratory testing procedures or duration of antimicrobial treatment, were not investigated. Lastly, the study was limited to a single case scenario, and the results may not be generalisable to other clinical scenarios. Thus, we refrained from an absolute ranking of the LLMs.

CONCLUSIONS

The latest versions of LLMs show potential in helping healthcare professionals. Our study underscores the need for cautious and informed use of most of these models as demonstrated by the limitations in providing specific information and potentially misleading information for diagnostic work-up and treatment of adult patients with bacterial meningitis. Users should be aware of the variability in their performance.

Further research is needed to refine these models and enhance their understanding of complex medical scenarios and their ability to provide deterministic, reliable information regardless of prompt nuances. Concurrently, efforts are necessary to mitigate the potential for disseminating erroneous content.

Contributors UF and RS planned and designed the study. UF acquired the data and wrote the first draft of the manuscript and is responsible for the overall content as guarantor. All authors interpreted the data, revised the manuscript for important intellectual content and approved the final submitted version.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval As no real patients were involved in this study, ethical approval was not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Urs Fisch http://orcid.org/0000-0003-1557-9062

REFERENCES

1 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature* 2023;620:172–80.
2 Kung TH, Cheatham M, Medenilla A, *et al*. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198.
3 Gilson A, Safranek CW, Huang T, *et al*. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.
4 Levine DM, Tuwani R, Kompa B, *et al*. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. *Health Informatics* [Preprint] 2023.
5 Ayers JW, Poliak A, Dredze M, *et al*. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183:589–96.
6 Haver HL, Ambinder EB, Bahl M, *et al*. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* 2023;307:e230424.
7 Tang H, Ng JHK. Googling for a diagnosis--use of Google as a diagnostic aid: internet based study. *BMJ* 2006;333:1143–5.
8 Russell-Rose T, Chamberlain J. Expert search strategies: the information retrieval practices of healthcare information professionals. *JMIR Med Inform* 2017;5:e33.
9 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023;388:1233–9.
10 Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? *Lancet Infect Dis* 2023;23:405–6.
11 Liévin V, Hother CE, Motzfeldt AG, *et al*. Can large language models reason about medical questions? *arXiv:220708143* 2023. Available: https://doi.org/10.48550/arXiv.2207.08143
12 Norori N, Hu Q, Aellen FM, *et al*. Addressing bias in big data and AI for health care: a call for open science. *Patterns (N Y)* 2021;2:100347.
13 Wang J, Shi E, Yu S, *et al*. Prompt engineering for healthcare: methodologies and applications. 2023. Available: https://doi.org/10.48550/arXiv.2304.14670
14 van de Beek D, Cabellos C, Dzupova O, *et al*. ESCMID guideline: diagnosis and treatment of acute bacterial meningitis. *Clin Microbiol Infect* 2016;22 Suppl 3:S37–62.
15 Dyckhoff-Shen S, Koedel U, Pfister H-W, *et al*. SOP: emergency workup in patients with suspected acute bacterial meningitis. *Neurol Res Pract* 2021;3:2.
16 Tunkel AR, Hartman BJ, Kaplan SL, *et al*. Practice guidelines for the management of bacterial meningitis. *Clin Infect Dis* 2004;39:1267–84.
17 Sigfrid L, Perfect C, Rojek A, *et al*. A systematic review of clinical guidelines on the management of acute, community-acquired CNS infections. *BMC Med* 2019;17:170.
18 Steiner I, Budka H, Chaudhuri A, *et al*. Viral meningoencephalitis: a review of diagnostic methods and guidelines for management. *Eur J Neurol* 2010;17:999–e57.
19 Nori H, King N, McKinney SM, *et al*. Capabilities of GPT-4 on medical challenge problems. 2023. Available: https://doi.org/10.48550/arXiv.2303.13375
20 Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI* 2023;1.
21 Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023;330:78–80.
22 Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res* 2021;23:e25759.
23 Susanto AP, Lyell D, Widyantoro B, *et al*. Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. *J Am Med Inform Assoc* 2023;30:2050–63.
24 Gama F, Tyskbo D, Nygren J, *et al*. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. *J Med Internet Res* 2022;24:e32215.