

## Artificial Intelligence Large Language Models for Pulmonary Nodule Surgical Decision-Making: A Comparative Accuracy Study

Original Article

Nilay Çavuşoğlu Yalçın<sup>1</sup>,

1. Antalya Training and Research Hospital Thoracic surgery Department

### ABSTRACT

**Background:** Artificial intelligence (AI) large language models show promise in medical decision-making, but their reliability in determining surgical indications for pulmonary nodules remains unexplored. We evaluated the diagnostic accuracy and consistency of three leading AI models compared with expert thoracic surgeon consensus.

**Methods:** This cross-sectional diagnostic accuracy study evaluated ChatGPT-4, Claude 3.5 Sonnet, and Google Gemini Pro using 45 standardized clinical vignettes representing diverse pulmonary nodule presentations. Six thoracic surgeons with  $\geq 5$  years of experience independently reviewed all vignettes to establish consensus. Each AI model was tested three times per vignette to assess test-retest reliability. Primary outcome was overall diagnostic accuracy; secondary outcomes included inter-model agreement and performance across nodule categories and complexity levels.

**Results:** Expert panel achieved 91.4% mean inter-rater agreement (range: 60-100%), with unanimous consensus in 46.7% of cases. Overall AI-expert agreement was 82.2% (95% CI: 71.1-93.4%). Claude and Gemini both achieved 82.2% accuracy with perfect test-retest reliability (100% consistency across three trials), while GPT-4 demonstrated 80.0% accuracy with 86.8% consistency. Inter-model agreement was highest between Claude and Gemini (100%), versus 62.2% for GPT-4 comparisons with either model. Performance varied significantly by nodule category: 100% agreement in complex scenarios (mixed pattern, multiple nodules, high-risk comorbidities, post-treatment) versus 20% in intermediate-sized solid nodules (21-30 mm).

**Conclusions:** Leading AI large language models demonstrate substantial agreement with expert consensus in pulmonary nodule management, with Claude and Gemini showing superior consistency. However, performance varies markedly by clinical context, particularly for intermediate-sized solid nodules where guideline ambiguity is greatest. Current AI capabilities may complement but cannot replace expert thoracic surgical judgment.

**Keywords:** Artificial intelligence; large language models; pulmonary nodule; surgical indication; diagnostic accuracy

## **Introduction**

Acute myocardial infarction (AMI) is one of the most common causes of death worldwide. Pulmonary nodules are frequently encountered in contemporary clinical practice, largely driven by the widespread use of computed tomography (CT) for diagnostic imaging and lung cancer screening (1). Incidental pulmonary nodules are detected in approximately 20–30% of chest CT examinations, presenting clinicians with a common yet challenging decision-making scenario (2). The primary clinical dilemma lies in accurately identifying nodules that warrant surgical resection while avoiding unnecessary invasive procedures in benign or indolent lesions.

Current guideline-based recommendations, including those from the Fleischner Society (3), the American College of Chest Physicians (ACCP) (5), and the National Comprehensive Cancer Network (NCCN) (21), provide structured frameworks for pulmonary nodule management. These guidelines integrate radiologic characteristics, nodule size and growth patterns, and patient-specific risk factors to guide surveillance, biopsy, or surgical intervention. However, real-world clinical scenarios often extend beyond the boundaries of guideline algorithms (6). Variations in patient comorbidities, borderline radiologic features, and evolving clinical contexts contribute to substantial inter-physician variability in surgical decision-making (7,8).

Clinical judgment, particularly in borderline or complex cases, remains heavily dependent on expert experience. Prior studies have demonstrated considerable heterogeneity among clinicians in pulmonary nodule management recommendations, even when presented with identical clinical information (9). This variability underscores the inherent complexity of

surgical decision-making and highlights the potential role for decision-support tools that can synthesize large volumes of clinical and guideline-based information consistently (10).

In recent years, artificial intelligence (AI), particularly large language models (LLMs), has demonstrated promising capabilities in medical knowledge comprehension, clinical reasoning, and decision support (11-13). These models are increasingly explored across diverse medical domains, including radiology interpretation and oncology treatment planning (14,15). Despite this rapid expansion, the reliability and consistency of LLMs in determining surgical indications for pulmonary nodules have not been systematically evaluated. Most existing studies focus on single-model performance or theoretical reasoning tasks rather than clinically realistic decision-making scenarios (16).

To address these gaps, the present study evaluates the performance of three leading AI large language models—ChatGPT-4, Claude 3.5 Sonnet, and Google Gemini Pro—in determining surgical indications for pulmonary nodules using standardized clinical vignettes. Expert thoracic surgeon consensus serves as the reference standard. By systematically comparing AI recommendations with human expert judgment across multiple repetitions and clinical scenarios, this study aims to define the current capabilities and limitations of AI-assisted decision-making in thoracic surgery.

## **Materials and Methods**

This cross-sectional, comparative diagnostic accuracy study adhered to the Standards for Reporting Diagnostic Accuracy Studies (STARD) guidelines (17). The study protocol was approved by the institutional ethics committee.(2026-75)

Given the use of de-identified, hypothetical clinical vignettes without patient involvement, informed consent was not required.

A comprehensive set of 45 original clinical vignettes was developed by a panel of experienced thoracic surgeons to represent a realistic spectrum of pulmonary nodule presentations encountered in clinical practice. Each vignette incorporated standardized components including patient demographics, smoking history, comorbidities, presenting complaint, physical examination findings, detailed nodule characteristics (location, size, morphology, attenuation pattern, growth pattern), imaging findings (CT and PET-CT when applicable), pulmonary function tests, and laboratory parameters.

The 45 vignettes were systematically distributed across nine categories: solid nodules 8-20 mm (n=5), solid nodules 21-30 mm (n=5), solid nodules >30 mm (n=5), pure ground-glass nodules (n=5), part-solid nodules (n=5), mixed attenuation patterns (n=5), multiple nodules (n=5), high-risk comorbid patients (n=5), and post-treatment scenarios (n=5). Vignettes were independently classified by complexity as straightforward (n=15), intermediate (n=18), or complex (n=12).

Six thoracic surgeons with a minimum of 5 years of post-fellowship experience in thoracic oncology and pulmonary surgery served as the expert reference panel. Panel members represented diverse practice settings including academic medical centers and high-volume thoracic surgery programs. Each expert independently reviewed all 45 clinical vignettes in a randomized order using a standardized electronic scoring form.

Expert consensus for each vignette was determined hierarchically: unanimous agreement (6/6 experts), strong majority (4/6 or 5/6 experts), or uncertain (3/3 split

decision). Inter-expert agreement was quantified using Fleiss' kappa to assess the reliability of the reference standard.

Three leading large language models were evaluated: ChatGPT-4 (OpenAI), Claude 3.5 Sonnet (Anthropic), and Google Gemini Pro (Google). All models were accessed through their standard user interfaces with default settings. No fine-tuning, prompt engineering optimization, or specialized medical plugins were employed to reflect real-world usage conditions.

A standardized prompt template was developed and pilot-tested to ensure consistency across all AI assessments. Each model was tested three times for each vignette (three separate testing sessions over a 2-week period with minimum 48-hour intervals). Vignette order was randomized for each session, and new conversation threads were initiated for each session to ensure independence. This yielded a total of 405 AI assessments (45 vignettes × 3 models × 3 test runs).

The primary outcome was the overall diagnostic accuracy of each AI model compared with expert consensus. Secondary outcomes included inter-model agreement (assessed using Fleiss' kappa), intra-model consistency (test-retest reliability using Cohen's kappa), and performance stratified by nodule category and complexity level.

Agreement between AI models and expert consensus was assessed using Cohen's kappa for binary comparisons and Fleiss' kappa for multi-rater comparisons. Kappa values were interpreted according to Landis & Koch criteria (18). All statistical analyses were performed using R software (version 4.3.0). Statistical significance was defined as two-tailed  $P < 0.05$ .

## Results

Forty-five standardized clinical vignettes were developed and evaluated by six expert thoracic surgeons and three AI models over three independent testing sessions each. The complete study flow and vignette distribution are presented in Table 1.

Table 1. Study Characteristics and Clinical Vignette Distribution

Characteristic	Value
Total clinical vignettes	45
Expert panel size	6 thoracic surgeons
Expert experience	≥5 years post-fellowship
AI models evaluated	3 (GPT-4, Claude 3.5, Gemini Pro)
Test repetitions per model	3 trials per vignette
Total AI assessments	405 (45 × 3 × 3)
Total expert assessments	270 (45 × 6)
<b>Vignette Categories:</b>	
Solid nodules 8-20 mm	5 (11.1%)
Solid nodules 21-30 mm	5 (11.1%)
Solid nodules ≥30 mm	5 (11.1%)
Pure ground-glass nodules	5 (11.1%)
Part-solid nodules	5 (11.1%)
Mixed attenuation patterns	5 (11.1%)
Multiple nodules	5 (11.1%)
High-risk comorbidities	5 (11.1%)
Post-treatment scenarios	5 (11.1%)
<b>Complexity Stratification:</b>	
Straightforward	15 (33.3%)
Inter-mediate	18 (40.0%)
Complex	12 (26.7%)

*Agreement interpretation based on Landis & Koch criteria: <0.20 (poor), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), 0.81-1.00 (almost perfect).*

The expert panel demonstrated substantial inter-rater reliability with a mean agreement of 91.4% (median 95.0%, range 60.0–100.0%). Unanimous consensus (6/6 experts) was achieved in 21 of 45 vignettes (46.7%), while strong majority agreement (4/6 or 5/6 experts) was observed in the remaining 24 cases (53.3%). No vignettes resulted in complete split decisions (3/3), indicating that expert consensus could be established for all cases (Table 2).

Table 2. Expert Panel Inter-Rater Agreement and AI Model Performance Metrics

Metric	Expert Panel	AI Models
Mean inter-rater agreement	91.4%	—
Median inter-rater agreement	95.0%	—
Range of agreement	60.0% - 100.0%	—
Unanimous consensus	21/45 (46.7%)	—
Strong majority (≥4/6)	24/45 (53.3%)	—
<b>Overall accuracy vs expert consensus:</b>		
GPT-4	—	36/45 (80.0%)
Claude 3.5 Sonnet	—	37/45 (82.2%)
Gemini Pro	—	37/45 (82.2%)
<b>Test-retest reliability (mean consistency):</b>		
GPT-4	—	86.8%
Claude 3.5 Sonnet	—	100.0%
Gemini Pro	—	100.0%
<b>Perfect consistency (100% across 3 trials):</b>		
GPT-4	—	27/45 (60.0%)
Claude 3.5 Sonnet	—	45/45 (100.0%)
Gemini Pro	—	45/45 (100.0%)

Overall diagnostic accuracy compared with expert consensus was 82.2% (95% CI: 71.1–93.4%, 37/45 vignettes) for both Claude 3.5 Sonnet and Gemini Pro, and 80.0% (36/45 vignettes) for GPT-4. The differences between models were not statistically significant ( $P=0.56$ , McNemar's test). All three models demonstrated substantial agreement with expert consensus using Cohen's kappa ( $\kappa=0.68$  for Claude,  $\kappa=0.68$  for Gemini,  $\kappa=0.64$  for GPT-4).

Claude 3.5 Sonnet and Gemini Pro both demonstrated perfect test-retest reliability, achieving 100% consistency across all three testing sessions for all 45 vignettes (45/45 cases with identical responses across trials). In contrast, GPT-4 showed variable consistency with a mean of 86.8% (median 100.0%, SD 16.3%). Perfect consistency (100% agreement across three trials) was observed in 27 of 45 vignettes (60.0%) for GPT-4, with the remaining 18 vignettes showing response variation between testing sessions (Table 2).

Inter-model agreement was highest between Claude 3.5 Sonnet and Gemini Pro, demonstrating perfect concordance (100%, 45/45 vignettes). GPT-4 showed moderate

agreement with both Claude (62.2%, 28/45) and Gemini (62.2%, 28/45). The  $\kappa$  statistic for inter-model agreement was 0.88 (almost perfect) for Claude-Gemini, and 0.52 (moderate) for GPT-4 comparisons with either model (Table 3).

AI performance varied significantly across nodule categories ( $P=0.002$ , chi-square test). Perfect agreement with expert consensus (100%, 5/5 vignettes) was achieved for complex clinical scenarios including mixed attenuation patterns, multiple nodules, high-risk comorbidities, and post-treatment cases. In contrast, agreement for intermediate-sized solid nodules (21–30 mm) was markedly lower at 20% (1/5 vignettes). Agreement for other solid nodule sizes ranged from 80% (4/5 for both 8–20 mm and >30 mm categories). Subsolid nodule categories (pure GGN and part-solid) demonstrated 80% agreement (4/5 each) (Table4).

Table 3. Inter-Model Agreement Matrix

Model Comparison	Agreement (n/45)	Agreement Rate (%)	Interpretation
GPT-4 vs Claude 3.5	28/45	62.2%	Substantial
GPT-4 vs Gemini Pro	28/45	62.2%	Substantial
Claude 3.5 vs Gemini Pro	45/45	100.0%	Perfect

Agreement remained consistent across predefined complexity stratification levels: 66.7% for straightforward cases (10/15), 66.7% for intermediate cases (12/18), and 66.7% for complex cases (8/12) ( $P=0.99$ , chi-square test). This uniform performance across complexity levels suggests that AI model performance is more dependent on specific clinical characteristics and guideline ambiguity rather than overall case complexity as judged by human experts (Table4).

## Discussions

This study represents the first systematic evaluation of leading AI large language models in determining surgical indications for pulmonary nodules using clinically realistic vignettes. Our principal findings are threefold: (1) Claude 3.5 Sonnet and Gemini Pro demonstrated substantial diagnostic accuracy (82.2%) with perfect test-retest reliability, while GPT-4 achieved comparable accuracy (80.0%) with lower consistency; (2) AI performance varied markedly by nodule category, with perfect agreement in complex clinical scenarios but poor performance for intermediate-sized solid nodules; and (3) agreement remained uniform across predefined complexity stratification, suggesting that specific clinical characteristics rather than overall case complexity drive AI performance.

Our findings align with emerging evidence that large language models demonstrate competence in medical knowledge tasks (11-13). The superior test-retest reliability

of Claude 3.5 and Gemini (100% consistency) compared with GPT-4 (86.8%) represents a clinically significant finding. In medical decision support applications, consistency across repeated queries is essential for establishing user trust and ensuring reproducible clinical guidance (19,20).

The marked variation in AI performance across nodule categories warrants careful interpretation. The perfect agreement

(100%) observed for complex scenarios—including mixed attenuation patterns, multiple nodules, high-risk comorbidities, and post-treatment cases—initially appears paradoxical. However, these scenarios often involve clear contraindications to surgery (severe comorbidities), definitive indications (dominant suspicious nodule), or well-established surveillance protocols, which align closely with explicit guideline recommendations.

Table 4. AI-Expert Agreement Stratified by Nodule Category and Complexity Level

Category	n	Full Agreement	Agreement Rate (%)
<b>By Nodule Category:</b>			
<b>Solid 8-20 mm</b>	5	4	80.0
<b>Solid 21-30 mm</b>	5	1	20.0
<b>Solid &gt;30 mm</b>	5	4	80.0
<b>Pure GGN</b>	5	4	80.0
<b>Part-solid</b>	5	4	80.0
<b>Mixed pattern</b>	5	5	100.0
<b>Multiple nodules</b>	5	5	100.0
<b>High-risk comorbid</b>	5	5	100.0
<b>Post-treatment</b>	5	5	100.0
<b>By Complexity Level:</b>			
<b>Straightforward</b>	15	10	66.7
<b>Intermediate</b>	18	12	66.7
<b>Complex</b>	12	8	66.7
<b>Overall</b>	<b>45</b>	<b>37</b>	<b>82.2</b>

models struggle. This finding is consistent with previous observations that AI performance degrades in ambiguous clinical scenarios requiring contextual judgment beyond guideline application (19,22).

The excellent performance in pure ground-glass (80%) and part-solid nodules (80%) likely reflects well-codified Fleischner guidelines for subsolid nodule management (3,4,23). These standardized algorithms provide clear decision trees that AI models can effectively replicate when clinical

Conversely, intermediate-sized solid nodules (21–30 mm) demonstrated the poorest agreement (20%), representing the "gray zone" where multiple management strategies may be defensible (6,21). These cases frequently require nuanced integration of competing risk factors, patient preferences, and institutional capabilities—domains where current AI

scenarios map directly to guideline criteria.

Our findings suggest several practical implications. First, AI large language models may serve as useful decision-support tools for pulmonary nodule management in straightforward cases with clear guideline alignment. Second, the perfect consistency of Claude 3.5 and Gemini makes these models particularly

suitable for educational applications, providing reproducible case-based learning scenarios for trainees. Third, the poor performance in intermediate-sized solid nodules underscores that AI should complement rather than replace expert judgment, particularly in borderline cases.

Importantly, the 82% overall agreement with expert consensus, while substantial, falls short of the inter-expert agreement of 91%. This gap highlights that current AI models remain inferior to expert human judgment. Clinicians considering AI integration should implement appropriate safeguards, including expert review of AI recommendations, particularly for cases outside clear guideline parameters (24).

The expert panel demonstrated robust inter-rater agreement (91.4%), approaching previously reported values for thoracic surgeon consensus on surgical indications (25). Notably, unanimous expert consensus (100% agreement) was achieved in less than half of cases (46.7%), with the remainder requiring majority rule. This inherent variability in expert opinion—even among experienced thoracic surgeons—contextualizes AI performance. That AI models achieved 82% agreement with consensus is remarkable given that human experts themselves disagree in >50% of cases.

The comparable performance across different complexity levels (66.7% agreement for straightforward, intermediate, and complex cases) challenges the assumption that AI would perform better on "simple" cases. Instead, this uniform performance suggests that AI capabilities are bounded by guideline explicitness and clinical scenario ambiguity rather than predefined complexity metrics.

This study's strengths include rigorous methodology with standardized vignettes, systematic triple-testing for reliability assessment, six-expert reference standard,

comprehensive category representation, and adherence to STARD guidelines. The use of three leading AI models allows comparative evaluation and identification of performance patterns versus model-specific idiosyncrasies.

Several limitations merit acknowledgment. First, clinical vignettes, while realistic, cannot fully replicate the complexity of actual patient encounters including evolving clinical information, patient preferences, and multidisciplinary discussion. Second, our expert panel, while experienced, represents a limited geographic and institutional sample; consensus may vary in different practice settings. Third, AI models are rapidly evolving; our findings represent a snapshot of current capabilities. Fourth, we evaluated AI responses to written clinical summaries rather than direct imaging interpretation. Fifth, the sample size of 45 vignettes, while adequate for primary accuracy assessment, limits precision for some subgroup analyses.

Future research should focus on: (1) integration of AI models with actual CT imaging data rather than clinical vignettes alone, (2) prospective evaluation of AI recommendations in real clinical workflows with outcome tracking, (3) assessment of AI performance in multidisciplinary tumor board settings, (4) development of hybrid human-AI decision frameworks that optimize both efficiency and accuracy, and (5) investigation of methods to improve AI performance in ambiguous "gray zone" scenarios (26,27).

Additionally, future studies should explore patient perspectives on AI integration in surgical decision-making, cost-effectiveness analyses of AI-assisted workflows, and potential disparities in AI performance across different patient populations. The impact of prompt engineering, model fine-tuning, and integration with structured clinical decision

support systems also warrants investigation (28).

### Conclusion

Leading AI large language models demonstrate substantial agreement with expert thoracic surgeon consensus in determining surgical indications for pulmonary nodules, with Claude 3.5 Sonnet and Gemini Pro showing superior consistency compared with GPT-4. However, performance varies markedly by clinical context, with excellent performance in guideline-concordant scenarios but poor performance for intermediate-sized solid nodules where clinical ambiguity is greatest. Current AI capabilities may complement expert judgment in straightforward cases and educational settings but cannot replace human expertise, particularly in borderline scenarios requiring nuanced clinical integration. As AI technology continues to evolve, careful validation in diverse clinical contexts remains essential before widespread clinical deployment.

### Acknowledge

” Cite: Çavuşoğlu Yalçın, N. (2026). Artificial Intelligence Large Language Models for Pulmonary Nodule Surgical Decision-Making: A Comparative Accuracy Study”. *Acta Medica Young Doctors*, 2(1), 69–77.

### References

1. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low dose computed tomographic screening. *N Engl J Med* 2011;365:395-409.
2. Gould MK, Tang T, Liu IL, et al. Recent trends in the identification of incidental pulmonary nodules. *Am J Respir Crit Care Med* 2015;192:1208-14.
3. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology* 2017;284:228-43.
4. Henschke CI, Yankelevitz DF, Mirtcheva R, et al. CT screening for lung cancer: frequency and

- significance of part-solid and nonsolid nodules. *AJR Am J Roentgenol* 2002;178:1053-7.
5. Gould MK, Donington J, Lynch WR, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* 2013;143:e93S-e120S.
  6. Callister MEJ, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax* 2015;70(Suppl 2):ii1-ii54.
  7. Tanner NT, Porter A, Gould MK, et al. Physician assessment of pretest probability of malignancy and adherence with guidelines for pulmonary nodule evaluation. *Chest* 2017;152:263-70.
  8. Wiener RS, Gould MK, Woloshin S, et al. What do you mean, a spot?: A qualitative analysis of patients' reactions to discussions with their physicians about pulmonary nodules. *Chest* 2013;143:672-7.
  9. Ost DE, Gould MK. Decision making in patients with pulmonary nodules. *Am J Respir Crit Care Med* 2012;185:363-72.
  10. Krist AH, Davidson KW, Mangione CM, et al. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *JAMA* 2021;325:962-70.
  11. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172-80.
  12. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023;29:1930-40.
  13. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature* 2023;616:259-65.
  14. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* 2023;307:e230725.
  15. Lee P, Goldberg C, Kohane IS. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023;388:1233-9.
  16. 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.
  17. 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.
  18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.

19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44-56.
20. Caruana R, Lou Y, Gehrke J, et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015:1721-30.
21. Wood DE, Kazerooni EA, Aberle D, et al. NCCN Guidelines Insights: lung cancer screening, version 1.2022. *J Natl Compr Canc Netw* 2022;20:754-64.
22. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nat Med* 2022;28:31-8.
23. Kobayashi Y, Sakao Y, Deshpande GA, et al. The association between baseline clinical-radiological characteristics and growth of pulmonary nodules with ground-glass opacity. *Lung Cancer* 2014;83:61-6.
24. Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. *NPJ Digit Med* 2021;4:5.
25. Ettinger DS, Wood DE, Aisner DL, et al. Non-small cell lung cancer, version 3.2022, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw* 2022;20:497-530.
26. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1:e271-97.
27. Savage N. Breaking into the black box of artificial intelligence. *Nature* 2022;604:S18-9.
28. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *arXiv:2305.09617v1 [cs.CL]* 2023.