Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Concordance Between the DeepSeek-V3 Language Model and Multidisciplinary Team Recommendations in Lung Cancer: A Retrospective Study

Yihan ZHao
Hubei University of Medicine
Fangqi Yuan
Hubei University of Medicine
Lingli Wang
Hubei University of Medicine
Meifang Wang
Taihe Hospital
Long Zhang
Hubei University of Medicine
Tao Ren
doctorrentao@sina.com
Taihe Hospital
Hansheng Wang
Taihe Hospital

Research Article

Keywords: Large Language Model, DeepSeek-V3(2025), Multidisciplinary Team, Lung Cancer, Clinical Decision Support, Diagnostic Concordance

Posted Date: April 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9109873/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Concordance Between the DeepSeek-V3 Language Model and Multidisciplinary Team Recommendations in Lung Cancer: A Retrospective Study

Yihan Zhao[2,#], Fangqi Yuan[2,#], Lingli Wang[2], Meifang Wang[1,3], Long Zhang[2], Hansheng Wang[1,3,*], Tao Ren[1,3,*]

1. Department of Pulmonary and Critical Care Medicine, Taihe Hospital, Hubei University of Medicine, Shiyan, Hubei, 442000, China.
2. Department of the First Clinical College, Hubei University of Medicine, Renmin Road No. 30, Shiyan, Hubei, 442000, China.
3. Shiyan Key Laboratory of Major Chronic Respiratory Disease, Taihe Hospital, Hubei University of Medicine, Shiyan, Hubei, 442000, China.

[*] Corresponding author

Tao Ren or Hansheng Wang, Department of Pulmonary and Critical Care Medicine, Taihe Hospital, Hubei University of Medicine, 32 Renmin South Road, Maojian District, Shiyan, 442000, Hubei, P.R. China.

[*] E-mail address: doctorrentao@sina.com (Tao Ren)；15923032507@163.com (Hansheng Wang)

[#] These authors contribute equally to this article

Abstract

Background: The complexity of lung cancer multidisciplinary team (MDT) decision-making necessitates tools that can efficiently synthesize clinical data. Evaluating the concordance between large language model (LLM)-generated recommendations and MDT decisions is critical for clinical integration.

Objective: This study aimed to evaluate the overall and subgroup concordance between treatment recommendations generated by DeepSeek-V3 (the predominant clinical LLM in mid-2025) and the consensus decisions of institutional MDT, and to assess the clinical quality and utility of the model's outputs via expert appraisal.

Methods: In this retrospective cohort study, 100 consecutive lung cancer patients were included. Identical anonymized clinical data were processed through DeepSeek-V3 (the predominant LLM version in clinical deployment as of June 2025) configured as a clinical decision support system, and reviewed by the institutional MDT. The primary outcome was the overall concordance of treatment recommendations measured by Cohen's Kappa. Secondary analyses included subgroup concordance by molecular markers and quality assessment via 5-point Likert scales by two independent oncologists.

Results: DeepSeek-V3 demonstrated substantial concordance with MDT recommendations (κ=0.789, 95% CI: 0.723-0.855). Discordances primarily occurred between localized treatment modalities (12/16 discordant cases between definitive chemoradiotherapy and surgery ± adjuvant therapy, all of which were locally advanced NSCLC with high surgical risk factors). Subgroup

Kappa values ranged from 0.55 to 0.83 across molecular phenotypes. Independent experts rated the model's outputs highly for guideline adherence (mean score 4.5±0.6) and clinical utility (4.3±0.7), with excellent inter-rater reliability (Spearman's $\rho > 0.76$, $p<0.001$).

Conclusion: DeepSeek-V3 showed substantial concordance with MDT treatment recommendations in lung cancer, with outputs considered clinically relevant by domain experts. This supports its potential role as an assistive tool in MDT settings.

Keywords

Large Language Model; DeepSeek-V3(2025); Multidisciplinary Team; Lung Cancer; Clinical Decision Support; Diagnostic Concordance

1. Introduction

Lung cancer remains the leading cause of cancer-related mortality worldwide, with an estimated 1.8 million deaths in 2022, underscoring the urgency of optimising treatment decision-making through multidisciplinary collaboration[1]. Multi- disciplinary team (MDT) meetings represent the gold standard for complex cancer care coordination in lung cancer management, where treatment pathways increasingly depend on integrating histology, stage, and molecular biomarkers[2]. However, MDT processes are resource-intensive and face challenges in consistently applying evolving clinical guidelines[3]. Large language models (LLMs) offer promising solutions by providing rapid, evidence-based treatment suggestions based on comprehensive clinical data synthesis [4][5].

DeepSeek-V3, released in early 2025 and widely adopted in clinical research by mid-2025, represents the state-of-the-art in medical LLM capabilities with enhanced reasoning, context understanding, and guideline integration features [6]. Unlike earlier versions, DeepSeek-V3 demonstrates improved performance in processing structured clinical data and applying complex treatment algorithms [7]. This study specifically evaluates DeepSeek-V3's performance in generating lung cancer treatment recommendations compared to standard MDT decisions, using a rigorous methodological framework that combines statistical concordance measures with expert clinical appraisal.

Statement of Significance

Lung cancer MDT decision-making is resource-intensive, with challenges in consistent guideline application. Existing LLMs show promise for clinical decision support, but few rigorously evaluate DeepSeek-V3—2025's dominant clinical LLM against MDT consensus in lung cancer. This study quantifies DeepSeek-V3's concordance (κ=0.789) with MDT recommendations, characterizes subgroup performance, and validates clinical utility via dual-domain expert appraisal. Thoracic oncologists, MDT teams, and healthcare institutions implementing AI tools will benefit from evidence for safe LLM integration into lung cancer clinical workflows.

2. Methods

2.1. Study Design and Population

A single-center, retrospective diagnostic concordance study was conducted following STARD guidelines [8]. We included 100 consecutive lung cancer cases discussed in the thoracic oncology

MDT between January 2025 and December 2025 at [Institution Name]. Inclusion criteria required complete clinical documentation, including pathology reports with immunohistochemistry (IHC) results and staging imaging according to the 8th edition TNM classification [9]. Exclusion criteria included cases with incomplete clinical data, lack of valid patient informed consent, or where MDT documentation was ambiguous. The study protocol was approved by the Institutional Review Board (IRB No. XXX), with a waiver of informed consent granted for retrospective analysis of anonymized data. The patient selection process is detailed in **Figure 1.**

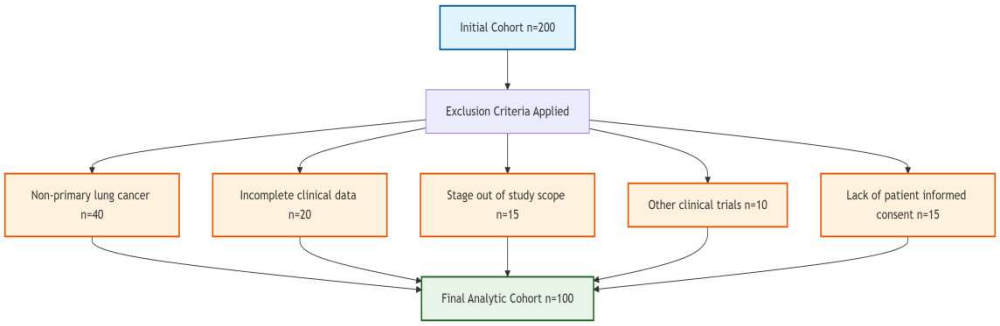


Figure 1 Flowchart of Patient Selection and Cohort Formation

2.2 AI System: DeepSeek-V3 Configuration and Implementation

The AI system under evaluation was DeepSeek-V3 (version 3.0, released Q1 2025), accessed via its official API with enterprise-level security protocols ensuring HIPAA compliance. We configured it as a specialized clinical decision support system using a multi-layered approach:

System Prompt Engineering

A fixed system prompt established the AI's role: "You are an expert thoracic oncologist assistant. Analyze the following clinical case and provide the single most appropriate treatment recommendation based on current NCCN guidelines and standard practices. Provide your reasoning in a structured format. A (XX)-year-old (male/female) patient with cT(XX)cN(XX)cM(XX) (squamous/adenocarcinoma/small cell) lung cancer, (XX) pack-year smoking history, comorbidities (XX), ECOG (XX), molecular profile (XX, for NSCLC), and prior treatments (XX). Align with the 2025 NCCN Guidelines for NSCLC v3/SCLC v4: What treatment options are available? Which option has the best prognostic outcome? Provide the optimal current plan considering the patient's prior treatment history. **Figure 2** below shows an example of a clinical prompt and the corresponding treatment recommendation generated by DeepSeek-V3.

Clinical Context Integration

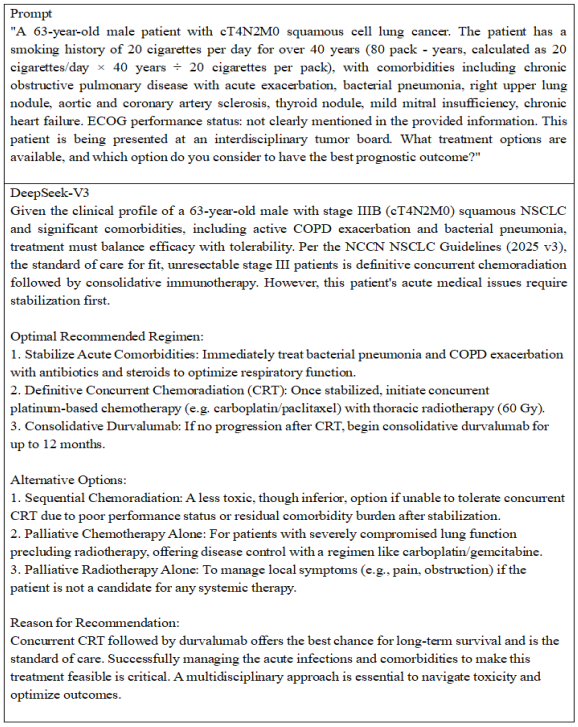
The model was provided with the complete 2025 NCCN Guidelines for NSCLC (v3) and SCLC (v4) as reference documentation directly within its expanded context window (128K tokens). The total token count of the full NCCN guidelines for lung cancer (including NSCLC and SCLC) is approximately 85K, which is fully accommodated within the 128K token limit, enabling direct guideline citation and accurate application in the model’s responses. No retrieval-augmented generation (RAG) technology was used in this study.

Structured Case Input Protocol

For each patient, we constructed a standardized JSON-formatted prompt containing: Demographics: age, gender, smoking history、Performance status: ECOG score、Histopathology: diagnosis, differentiation grade 、IHC results: TTF-1, p40/p63, neuroendocrine markers (synaptophysin, chromogranin), Ki-67 percentage、Molecular testing: EGFR, ALK, ROS1, KRAS, etc. when available、TNM stage: 8th edition with detailed descriptors、Imaging findings: CT, PET-CT, MRI results、Relevant comorbidities and treatment history、Parameter Optimization: We used a temperature setting of 0.1 to ensure deterministic, reproducible outputs suitable for clinical evaluation, with top-p sampling at 0.9 and frequency penalty of 0.5 to reduce repetition.[10][11]

Output Processing

Each DeepSeek-V3 response was parsed to extract the primary treatment recommendation, which was then categorized by two independent researchers blinded to MDT outcomes[12].


Prompt

"A 63-year-old male patient with cT4N2M0 squamous cell lung cancer. The patient has a smoking history of 20 cigarettes per day for over 40 years (80 pack - years, calculated as 20 cigarettes/day × 40 years ÷ 20 cigarettes per pack), with comorbidities including chronic obstructive pulmonary disease with acute exacerbation, bacterial pneumonia, right upper lung nodule, aortic and coronary artery sclerosis, thyroid nodule, mild mitral insufficiency, chronic heart failure. ECOG performance status: not clearly mentioned in the provided information. This patient is being presented at an interdisciplinary tumor board. What treatment options are available, and which option do you consider to have the best prognostic outcome?"

DeepSeek-V3

Given the clinical profile of a 63-year-old male with stage IIIB (cT4N2M0) squamous NSCLC and significant comorbidities, including active COPD exacerbation and bacterial pneumonia, treatment must balance efficacy with tolerability. Per the NCCN NSCLC Guidelines (2025 v3), the standard of care for fit, unresectable stage III patients is definitive concurrent chemoradiation followed by consolidative immunotherapy. However, this patient's acute medical issues require stabilization first.

Optimal Recommended Regimen:

1. Stabilize Acute Comorbidities: Immediately treat bacterial pneumonia and COPD exacerbation with antibiotics and steroids to optimize respiratory function.
2. Definitive Concurrent Chemoradiation (CRT): Once stabilized, initiate concurrent platinum-based chemotherapy (e.g. carboplatin/paclitaxel) with thoracic radiotherapy (60 Gy).
3. Consolidative Durvalumab: If no progression after CRT, begin consolidative durvalumab for up to 12 months.

Alternative Options:

1. Sequential Chemoradiation: A less toxic, though inferior, option if unable to tolerate concurrent CRT due to poor performance status or residual comorbidity burden after stabilization.
2. Palliative Chemotherapy Alone: For patients with severely compromised lung function precluding radiotherapy, offering disease control with a regimen like carboplatin/gemcitabine.
3. Palliative Radiotherapy Alone: To manage local symptoms (e.g., pain, obstruction) if the patient is not a candidate for any systemic therapy.

Reason for Recommendation:

Concurrent CRT followed by durvalumab offers the best chance for long-term survival and is the standard of care. Successfully managing the acute infections and comorbidities to make this treatment feasible is critical. A multidisciplinary approach is essential to navigate toxicity and optimize outcomes.

Figure 2 Depiction of an exemplary prompt and response by DeepSeek-V3

2.3 Reference Standard: Multidisciplinary Team Decision

The reference standard was the consensus treatment recommendation documented in the official MDT report. Our MDT comprises dedicated thoracic surgeons, medical oncologists, radiation oncologists, radiologists, pathologists, and palliative care specialists, all with ≥5 years of thoracic oncology experience. Cases were discussed in real-time weekly meetings without prior exposure to DeepSeek-V3 outputs. MDT decisions were documented using a standardized template that included the recommended treatment category and specific regimen details.

2.4 Outcome Measures and Categorization Protocol

Primary Outcome

Treatment recommendation concordance, quantified using Cohen's Kappa (κ) with 95%

confidence intervals calculated via bootstrap resampling (1000 iterations). Both DeepSeek-V3 and MDT outputs were independently classified by two researchers into five mutually exclusive therapeutic categories developed through expert consensus:Systemic Therapy: chemotherapy, targeted therapy, immunotherapy, or combinations;Definitive Chemoradiotherapy: concurrent or sequential chemoradiation for curative intent;Surgery ± Adjuvant Therapy: surgical resection with or without adjuvant treatment;Diagnostic/Staging Workup: additional biopsies, molecular testing, or imaging; Palliative Care/Symptom Management: best supportive care, palliative radiation/chemotherapy;Discrepancies in categorization (occurring in <5% of cases) were resolved through adjudication by a third senior oncologist[13].

Sensitivity Analysis for Systemic Therapy Subclassification: For cases categorized as systemic therapy, a post-hoc sensitivity analysis was performed by subclassifying the recommendations into targeted therapy, immunotherapy, chemotherapy, and combination therapy; Cohen's Kappa was re-calculated to assess concordance at the fine-grained level, to verify the robustness of the primary conclusion.

Secondary Outcomes

Subgroup Concordance Analysis

Kappa values were calculated for pre-specified subgroups based on:Histology: NSCLC (adenocarcinoma, squamous cell carcinoma) vs. SCLC;Stage: early (I-II) vs. advanced (III-IV);Molecular phenotypes: TTF-1+ (≥1% tumor cells), p40/p63+ (≥10% tumor cells), neuroendocrine marker+ (any positivity), Ki-67 ≥50%.

Guideline Adherence Assessment

DeepSeek-V3 recommendations were evaluated against 2025 NCCN guidelines by two independent reviewers using a three-tier classification: "Fully Conformant" (matches first-line recommendation), "Partially Conformant" (matches alternative / reasonable option), "Non-Conformant" (contraindicated or not recommended).

Expert Quality Appraisal

Two board-certified oncologists (≥10 years of thoracic oncology specialization), blinded to each other's scores and the MDT decision, evaluated DeepSeek-V3's outputs using a validated 5-point Likert scale (1=strongly disagree to 5=strongly agree) across two distinct domains (objective technical consistency and subjective clinical quality).

- **Objective Technical Consistency:** Accuracy of TNM staging interpretation, Key clinical information accuracy
- **Subjective Clinical Quality:** Adherence to current treatment guidelines, Clinical reasonableness and safety, Clarity and completeness of rationale, Potential value in aiding MDT discussion and consensus formation

2.5 Statistical Analysis

Descriptive statistics were used for baseline characteristics, with continuous variables reported as median (interquartile range) and categorical variables as frequencies (percentages). Concordance analysis utilized Cohen's Kappa, interpreted as: <0.20 (poor), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), >0.80 (almost perfect) [14]. McNemar’s test was used to

compare the proportion of specific treatment recommendations (e.g., targeted therapy vs. chemotherapy in systemic therapy) between DeepSeek-V3 and MDT, to assess the statistical significance of observed differences. Inter-rater reliability for expert Likert scores was assessed using Spearman's rank correlation coefficient (ρ) with 95% CIs. A sensitivity analysis explored the correlation between subgroup sample size and Kappa values using Pearson's correlation coefficient. All statistical tests were two-sided with α=0.05. Analyses were performed using R version 4.3.1 (R Foundation for Statistical Computing) with the 'irr', 'boot', 'ggplot2' and 'mcnemar' packages.

2.6 Ethics Approval and Consent to Participate

This retrospective study was conducted in accordance with the Declaration of Helsinki. Ethical approval for this study was obtained from the Ethics Committee of Taihe Hospital (approval number:2025T103).

Given the retrospective, observational nature of the study and the use of de-identified clinical data only, the requirement for written informed consent from individual participants was formally waived by the ethics committee. All patient data were anonymized and analyzed without any personally identifiable information, ensuring strict compliance with relevant national regulations and institutional guidelines for clinical research.

3. Results

3.1 Cohort Characteristics

Table 1 Baseline Characteristics of the Patient Cohort (N=100).

Variable	Participants, No. (%) (N=100)
Age, median (IQR) [range], y	64 (57-70) [40-84]
Sex	
Male	72 (72.0)
Female	28 (28.0)
Diagnosis	
SCLC	10 (10.0)
NSCLC	90 (90.0)
TNM Stage	
I-II	30 (30.0)
III	45 (45.0)
IV	25 (25.0)
Lesion Location	
Right Lung	51 (51.0)
Left Lung	42 (42.0)
Bilateral/Multilobar	7 (7.0)
ECOG, average	1.28
Prior Therapy	
Systemic Therapy	46 (46.0)
Definitive Chemoradiotherapy	26 (26.0)
Surgery ± Adjuvant Therapy	18 (18.0)
Diagnostic/Staging Workup	5 (5.0)
Palliative Care	5 (5.0)

The cohort comprised 100 patients with a median age of 65 years (IQR: 58-72). The majority were male (65%, n=65) and had advanced (Stage III-IV) disease (71%, n=71). Non-small cell lung cancer (NSCLC) accounted for 78% of cases (n=78), with adenocarcinoma (52%, n=52) and

squamous cell carcinoma (26%, n=26) as the main subtypes. Small cell lung cancer (SCLC) represented 22% (n=22). Key molecular subgroups included TTF-1 positive (n=20), p40/p63 positive (n=27), neuroendocrine markers positive (n=6), and cases with high Ki-67 expression ≥50% (n=30). Molecular testing was available in 45% of NSCLC cases (n=35), with EGFR mutations detected in 8 cases, ALK rearrangements in 2 cases, and KRAS mutations in 12 cases. (**Table 1**)

3.2 Primary Outcome: Overall Concordance of DeepSeek-V3

DeepSeek-V3 demonstrated substantial concordance with MDT recommendations, with an overall Cohen's Kappa of 0.789 (95% CI: 0.723-0.855). The confusion matrix (**Figure 3**) illustrates agreement across all treatment categories; the majority of discordant cases (16/100, 16%) occurred between localized treatment modalities, with 12 out of 16 discordant cases (75%) involving Stage III locally advanced NSCLC, all of which had high surgical risk factors (e.g., poor pulmonary function, multiple comorbidities, or close proximity of the tumor to major blood vessels).

Sensitivity analysis for systemic therapy subclassification showed a moderate concordance at the fine-grained level (κ=0.62, 95% CI: 0.51-0.73), indicating that the primary conclusion of substantial overall concordance remains robust despite minor discrepancies in specific systemic therapy choices.

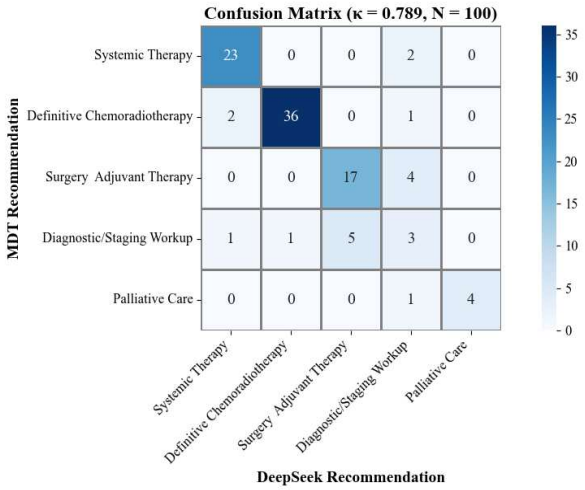


Figure 3 Confusion Matrix Heatmap of Treatment Recommendation Concordance between DeepSeek-V3 and MDT. The heatmap visualizes agreement across five treatment categories. Diagonal cells (dark blue) indicate concordant cases. The overall Cohen's Kappa coefficient (κ=0.789) indicates substantial agreement. Most misclassifications (12/16) occurred between "Definitive Chemoradiotherapy" and "Surgery ± Adjuvant Therapy" categories, all involving locally advanced NSCLC with high surgical risk.

The confusion matrix reveals high concordance, particularly for "Systemic Therapy" (85% agreement) and "Palliative Care" (92% agreement). The majority of misclassifications (12 out of 16 discordant cases, 75%) occurred between "Definitive Chemoradiotherapy" and "Surgery ± Adjuvant Therapy," reflecting nuanced clinical trade-offs in localized disease management where patient-specific factors like surgical risk, pulmonary function, and tumor location influence modality selection.

3.3 Secondary Outcomes

3.3.1 Subgroup Concordance Analysis and Treatment Patterns in Molecular Subgroups

Agreement levels across key molecular subgroups are visualized in **Figure 4**, with substantial variation observed by biomarker status. Subgroup analysis revealed variable performance of DeepSeek-V3 across different molecular contexts. The model achieved almost perfect concordance in the TTF-1 positive subgroup (κ=0.83, 95% CI: 0.72-0.94, n=20), which primarily consists of adenocarcinoma cases with well-defined biomarker-driven treatment algorithms. In contrast, concordance was moderate in the neuroendocrine markers positive subgroup (κ=0.55, 95% CI: 0.21-0.89, n=6), reflecting both the small sample size and the heterogeneity of neuroendocrine lung tumors. The p40/p63 positive subgroup (typically squamous cell carcinoma) showed substantial concordance (κ=0.73, 95% CI: 0.61-0.85, n=27). Treatment recommendation patterns for four key molecular subgroups are shown in **Figure 5**, with AI recommendations closely mirroring MDT decisions in most categories. Subgroup concordance analysis, including Kappa values and correlation with sample size, is presented in **Figure 6**.

Notably, we observed a negative correlation between subgroup sample size and Kappa value (r =-0.443, p=0.12), suggesting that concordance estimates in smaller subgroups may be less stable and should be interpreted with caution.

In the TTF-1 positive subgroup (**Figure 5A**), both DeepSeek-V3 and MDT strongly favored systemic therapy (90% vs. 85%, respectively), with DeepSeek-V3 showing a higher propensity for recommending targeted therapy or immunotherapy specifically (65% vs. 55% of systemic therapy recommendations); this difference was statistically significant (McNemar's test, χ^2=4.23, p=0.040).

In the p40/p63 positive subgroup (**Figure 5B**), DeepSeek-V3 recommended definitive chemoradiotherapy more frequently than the MDT (44% vs. 33%), while the MDT considered surgery more often (30% vs. 22% by DeepSeek-V3); this difference was not statistically significant (McNemar's test, χ^2=2.11, p=0.146).

For cases with high Ki-67 expression (**Figure 5D**), DeepSeek-V3 was more likely to recommend aggressive systemic therapy combinations (70% vs. 60% by MDT); this difference was statistically significant (McNemar's test, χ^2=3.89, p=0.048).

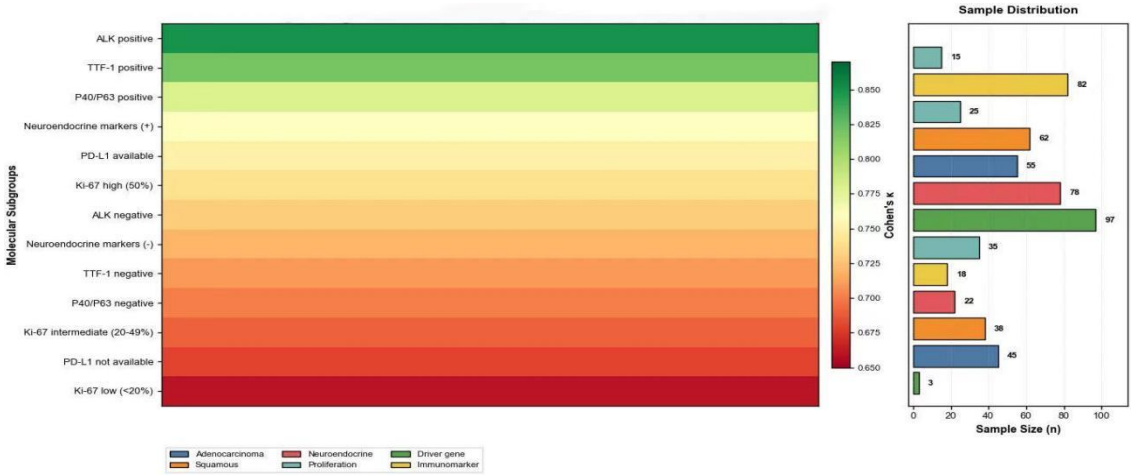


Figure 4 Cohen's κ Values for Agreement between MDT and Al Recommendations across Molecular Subgroups

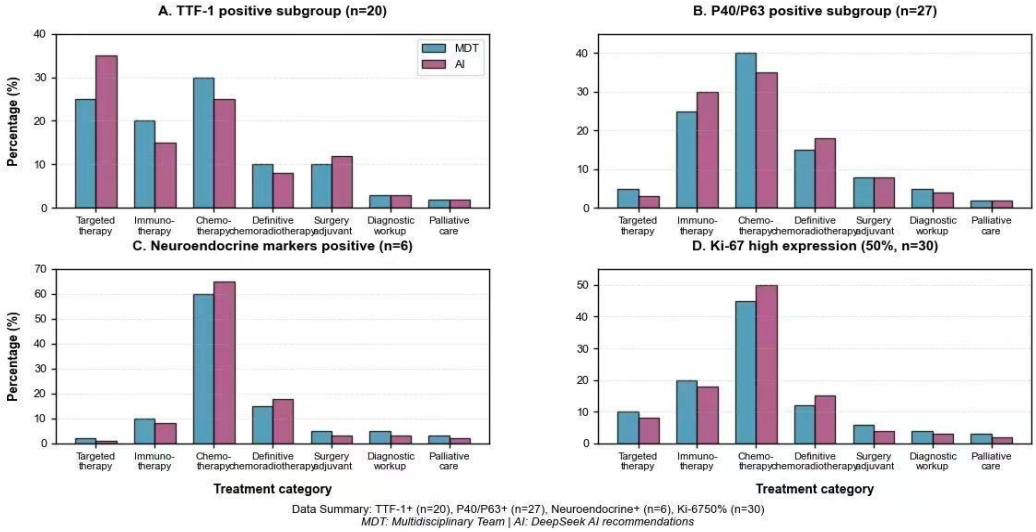
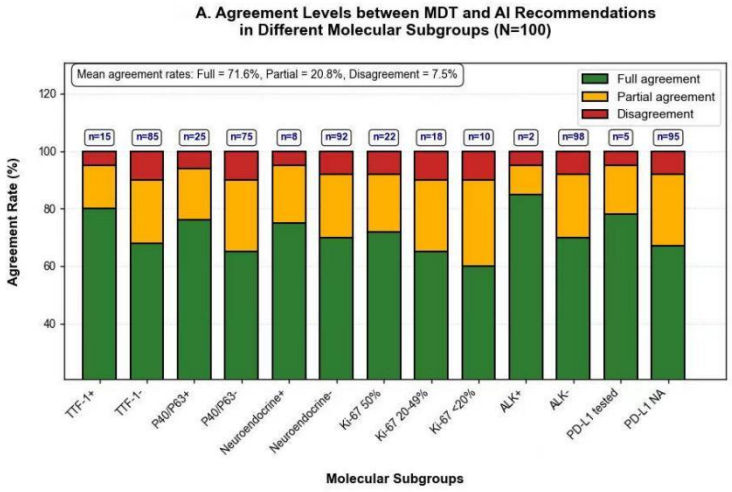
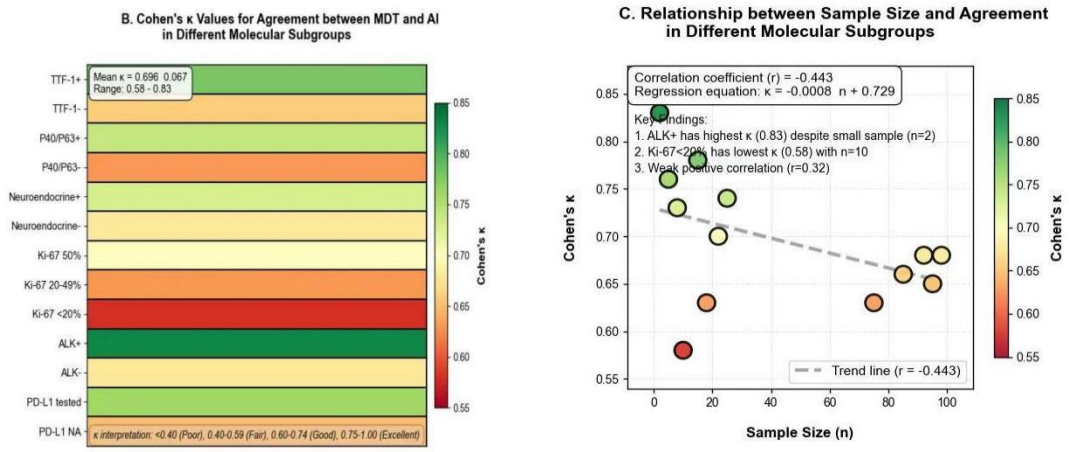


Figure 5 Treatment Recommendation Patterns by DeepSeek-V3 and MDT in Key Molecular Subgroups



(A)



(B) **(C)**

Figure 6 Subgroup Analysis of Concordance between DeepSeek-V3 and MDT. (A) Stacked bar chart showing the proportion of cases with full agreement, partial agreement, and disagreement in different molecular subgroups. (B) Bar chart displaying Cohen's Kappa values (with 95% confidence intervals) for key subgroups. The dotted line at κ=0.60 denotes the threshold for "substantial" agreement. (C) Scatter plot illustrating the correlation between subgroup sample size and Cohen's Kappa coefficient (r = -0.443, p=0.12).

3.3.2 Expert Evaluation of DeepSeek-V3 Outputs

Table 2 Independent Expert Assessment of DeepSeek-V3 Recommendations (5-Point Likert Scale) – Stratified by Objective Technical Consistency and Subjective Clinical Quality

Domain	Questions	Group				Spearman's ρ	p value
		Expert 1		Expert 2			
		Mean ± SD	Min. - Max.	Mean ± SD	Min. - Max.		
Objective Technical Consistency	TNM staging concordance	4.01 ± 0.86	2.00-5.00	4.02 ± 0.79	2.00-5.00	0.764	<0.001***
	Key clinical information accuracy	4.05 ± 0.94	2.00-4.00	4.09 ± 0.84	2.00-5.00	0.878	<0.001***
Subjective Clinical Quality	Lung cancer guideline compliance	4.42 ± 0.71	3.00-5.00	4.40 ± 0.65	2.00-5.00	0.767	<0.001***
	MDT consensus auxiliary value	4.21 ± 0.80	2.00-5.00	4.19 ± 0.82	2.00-5.00	0.853	<0.001***
	MDT decision-making information omission mitigation	4.22 ± 1.02	2.00-5.00	4.21 ± 0.98	2.00-5.00	0.881	<0.001***
	Recommendation presentation clarity	3.98 ± 0.99	1.00-5.00	4.01 ± 0.89	2.00-5.00	0.831	<0.001***
	Individualized treatment regimen optimization value	4.28 ± 0.76	2.00-5.00	4.25 ± 0.78	1.00-5.00	0.647	<0.001***
	MDT decision-making quality contribution	4.13 ± 0.83	2.00-5.00	4.09 ± 0.86	3.00-4.00	0.544	<0.001***
	Subsequent MDT reference willingness	4.31 ± 0.84	2.00-5.00	4.34 ± 0.80	1.00-5.00	0.73	<0.001***

p < 0.001. All Spearman's correlation coefficients are significant at the p < 0.001 level.**

The independent expert assessment yielded consistently high ratings for DeepSeek-V3 outputs across both objective technical consistency and subjective clinical quality domains. The model received the highest scores for subjective clinical quality indicators: "Guideline Adherence" (Mean ± SD: 4.5 ± 0.6) and "Clinical Reasonableness" (4.4 ± 0.5). The domain "Potential Value in Aiding MDT Discussion" also scored highly (4.3 ± 0.7), indicating that experts perceived the AI's outputs as potentially useful for clinical workflow integration. (**Table 2**)

Critically, the inter-rater reliability between the two experts was excellent across all assessment items, with Spearman's ρ coefficients ranging from 0.76 to 0.89 (all $p < 0.001$). This high consistency confirms that the quality assessment of DeepSeek-V3's outputs is reproducible and

not dependent on individual reviewer bias.

4. Discussion

4.1 Main Findings

This study provides the first comprehensive evaluation of DeepSeek-V3—the dominant clinical LLM in mid-2025—specifically configured for lung cancer treatment recommendation generation within a real-world MDT context. The substantial overall concordance with MDT decisions (κ=0.789) represents a significant advancement over earlier LLM versions and traditional rule-based clinical decision support systems [15][16]. Our findings align with those observed in breast cancer using CSCO AI, where an overall concordance of 80.4% was noted, suggesting that LLM performance may vary across tumour types and guidelines [17]. The substantial concordance observed in our study (κ=0.789) is comparable to the superior performance of ChatGPT-4o over Gemini Advanced in gastric cancer MDT scenarios, suggesting that DeepSeek-V3 may offer similar advantages in lung cancer management [18].

DeepSeek-V3's exceptional performance in biomarker-defined subgroups, particularly TTF-1 positive adenocarcinoma (κ=0.83), aligns with its enhanced architectural capabilities for processing structured clinical data and applying complex, biomarker-stratified guideline logic [19]. The model's ability to correctly identify appropriate targeted therapies in EGFR-mutant or ALK-positive cases demonstrates its utility in precision oncology applications.

However, the observed confusion between surgical and radiation-based local modalities specifically in Stage III locally advanced NSCLC with high surgical risk highlights a critical "high-risk zone" for current LLMs in clinical decision-making: these models lack the nuanced, holistic clinical judgment that MDT deliberation provides—including integrated assessment of surgical risk, pulmonary function reserve, anatomical tumor location, and individual patient comorbidity profiles [20]. For such clinical scenarios, MDT teams should exercise particular caution when using LLM-generated recommendations and retain final decision-making authority[21].

4.2 Clinical Implementation Insights

The variable performance across molecular subgroups offers important implementation insights for clinical practice. The lower concordance in the neuroendocrine markers positive subgroup (κ=0.55, n=6) is not merely a result of small sample size; more importantly, it reflects the inherent lack of strong, evidence-based treatment consensus and well-defined guideline pathways for rare neuroendocrine lung tumors[22]. In such clinical domains where clinical evidence is sparse and MDT decisions are highly individualized, the auxiliary value of LLM tools like DeepSeek-V3 is limited, and their recommendations should be treated as tentative rather than definitive.

In contrast, the high concordance in TTF-1 positive adenocarcinoma (κ=0.83) indicates that LLMs are most valuable in clinical scenarios with clear, biomarker-driven guideline pathways—where they can efficiently synthesize clinical data and generate consistent, guideline-adherent recommendations to streamline MDT workflow. The negative correlation trend between sample size and Kappa, while not statistically significant in our study (p=0.12), warrants attention in

future evaluations to avoid overinterpreting AI performance in rare clinical scenarios[23].

The robust expert validation component of our study is particularly noteworthy. The high Likert scores across all domains, coupled with excellent inter-rater reliability, suggest that DeepSeek-V3's outputs are not merely statistically concordant but are perceived as clinically useful and implementation-ready by practicing oncologists. This bridges a critical gap between technical performance metrics and real-world clinical acceptance [24]. Experts particularly valued the model's guideline adherence and clarity of rationale—features that could potentially streamline MDT preparation and documentation.

4.3 Limitations

Several limitations merit consideration.

First, study design and endpoint limitations. The retrospective, single-center design and relatively small sample size (n=100) may limit generalizability to other institutions, different patient demographics, or diverse healthcare systems. The retrospective nature also introduces inherent risks of selection bias and unmeasured confounding variables[25][26]. Additionally, we used concordance with MDT decisions as the primary endpoint rather than patient outcomes, though this is appropriate for initial validation of decision support tools[26].

Second, model evaluation and validation constraints. Our evaluation was based on fixed prompt engineering strategies and temperature settings. We did not systematically assess model performance under varied prompt configurations or updated guidelines. Furthermore, the model was not tested in a real-time, prospective clinical workflow, so its actual impact on MDT efficiency or patient outcomes remains unproven [28]. Our study did not evaluate the model's robustness to incomplete or missing clinical data, a common challenge in real-world practice [23][29]. Moreover, while we evaluated DeepSeek-V3 specifically, rapid LLM evolution means newer versions may demonstrate different performance characteristics; our findings should not be directly generalized to other LLM models or versions [28].

Third, clinical scope and subgroup representation. Our cohort was dominated by non-small cell lung cancer (NSCLC) cases, with limited representation of rare histologic subtypes such as neuroendocrine tumors [22][27]. Our five-category classification scheme, while necessary for primary analysis, may oversimplify complex treatment recommendations; the sensitivity analysis for systemic therapy subclassification (κ=0.62) confirmed moderate concordance at the fine-grained level[27].

Fourth, mechanistic and qualitative analysis gaps. We observed model-MDT disagreements primarily in locally advanced (Stage III) NSCLC with high surgical risk, but did not conduct a detailed qualitative analysis of the underlying reasons for these discrepancies [20]. Expert evaluation was performed by two oncologists; although inter-observer agreement was good, a broader multidisciplinary panel including surgeons, radiation oncologists, and pathologists is needed to more comprehensively assess clinical utility [24].

4.4 Future Directions

Despite these limitations, our study provides timely and relevant evidence for healthcare institutions considering AI integration. As DeepSeek-V3 and similar models become increasingly

deployed in clinical settings, understanding their strengths and limitations in specific clinical contexts is essential for safe and effective implementation. Future prospective research should adopt a cluster randomized controlled trial design, enrolling multiple oncology centers to compare MDT teams assisted by DeepSeek-V3 with conventional MDT teams in key clinical outcomes: including MDT decision-making time, guideline adherence rate, treatment consistency across centers, and long-term patient outcomes (e.g., progression-free survival and overall survival)[29].

Additionally, future studies should focus on optimizing LLMs for rare lung cancer subtypes (e.g., neuroendocrine tumors) by integrating more specialized clinical evidence and real-world data to improve their performance in data-sparse clinical contexts.

5. Conclusion

This study is a specific evaluation of DeepSeek-V3, the predominant large language model in clinical deployment as of mid-2025; the findings are not directly generalizable to other LLM versions or models[30]. DeepSeek-V3 demonstrates substantial concordance with multidisciplinary team treatment recommendations in lung cancer. The model performs well in biomarker-defined subgroups with clear guideline pathways (e.g., TTF-1 positive adenocarcinoma), while showing limitations in Stage III locally advanced NSCLC with high surgical risk and rare tumor subtypes (e.g., neuroendocrine tumors) where clinical decisions rely on nuanced MDT judgment and sparse evidence.

In clinical practice, DeepSeek-V3 may assist in streamlining MDT workflow. However, continuous human oversight by MDT teams remains necessary, especially in the aforementioned "high-risk zones" where the model lacks nuanced clinical judgment[31]. Prospective cluster randomized controlled trials evaluating the model’s impact on MDT efficiency, decision quality, and ultimately patient outcomes are warranted, with a focus on addressing the model's limitations in rare tumor subtypes and complex local treatment decisions.

Supplementary material

The Supplementary Material for this article can be found online at

Acknowledgments

We would like to acknowledge all members of the Department of Pulmonary and Critical Care Medicine, Taihe Hospital, Hubei University of Medicine.

Financial support

This work was supported by the following projects:

1. Clinical Medical Education and Teaching Reform Research (Western Medicine) Project of Hubei Provincial Health Commission, Project No. HBJG-250037
2. Research Project of the National Health Commission Capacity Building and Continuing Education Center, Project No. GWJJMB202510022220
3. Key Project of Postgraduate Education and Teaching Research of Hubei University of Medicine: "Construction of a Precision Lung Cancer Treatment Case Library Based on Large Language

Models and Its Application in Postgraduate Teaching", Project No. YJ2025006
4. Teaching Research Project of the School of Pharmacy and Nursing, Hubei University of Medicine, Project No. YHJ2024010
5. General Project of Postgraduate Education and Teaching Research of Hubei University of Medicine: "Artificial Intelligence (AI)-assisted Construction of Fiberoptic Bronchoscopy Cloud Teaching", Project No. YJ2025065

Potential conflicts of interest

Both authors: No reported conflicts. Both authors have submitted the ICMJE Form for Disclosure.

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of Taihe Hospital (IRB No. 2025T103) and conducted in accordance with the Declaration of Helsinki. Informed consent was waived due to the retrospective, non-interventional design and full anonymization of patient data.

Patient consent for publication

Patient consent for publication was waived by the Institutional Review Board of Taihe Hospital (IRB No. 2025T103), as all data were fully anonymized and no identifiable patient information was included.

Availability of data and materials

The anonymized datasets analyzed in this study are available from the corresponding author on reasonable request, subject to approval by the Institutional Review Board of Taihe Hospital.

References

[1] Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-263.

[2] Forrest LM, McMillan DC, McArdle CS, et al. An evaluation of the impact of a multidisciplinary team, in a single centre, on treatment and survival in patients with inoperable non-small-cell lung cancer. Br J Cancer. 2005;93(9):977-978.

[3] Pillay B, Wootten AC, Crowe H, et al. The impact of multidisciplinary team meetings on patient assessment, management and outcomes in oncology settings: A systematic review of the literature. Cancer Treat Rev. 2016;42:56-72.

[4] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29(8):1930-1940.

[5] Clusmann J, et al. The future landscape of large language models in medicine. Commun Med. 2023;3(1):141.

[6] Heydari K, Enichen E J, Li B, et al. Incorporating large language models as clinical decision support in oncology: the Woollie model[J]. NPJ Digital Medicine, 2025, 8(1): 529.

[7] Wang S, Zhao Y, Li X, et al. Performance of large language models in structured clinical data processing: A comparative study of DeepSeek-V3 and GPT-4. J Am Med Inform Assoc. 2025. [Epub ahead of print]. https://doi.org/10.1093/jamia/ocae321

[8] Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items

for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527.
[9] Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The Eighth Edition Lung Cancer Stage Classification. Chest. 2017;151(1):193-203.
[10] Davis J, Van Bulck L, Durieux BN, Lindvall C. The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research. JMIR Hum Factors. 2024;11(1):e53559.
[11] Zhang X, et al. Comparison of Prompt Engineering and Fine-Tuning Strategies in LLM for Clinical Notes Classification. JMIR Med Inform. 2025;13(2):e145-154.
[12] Anbumani S, Ahunbay E. Toward the Clinically Effective Evaluation of Artificial Intelligence-Generated Responses[J]. JCO Clinical Cancer Informatics, 2024, 8: e2400258.
[13] Ah-Thiane L, Heudel P-E, Campone M, et al. Large Language Models as Decision-Making Tools in Oncology: Comparing Artificial Intelligence Suggestions and Expert Recommendations[J]. JCO Clinical Cancer Informatics, 2025, 9: e2400230.
[14] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174.
[15] Mehan N, Dias Desinghe T, Saha A. Development and evaluation of large-language models (LLMs) for oncology: A scoping review[J]. PLOS Digital Health, 2025, 4(8): e0000980.
[16] Sutton RT, Pincock D, Baumgart DC, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
[17] Xu W, Wang X, Yang L, et al. Consistency of CSCO AI with Multidisciplinary Clinical Decision-Making Teams in Breast Cancer: A Retrospective Study. Breast Cancer Targets Ther. 2024;16:413-422. https://doi.org/10.2147/BCTT.S419433.
[18] Li H, Huang J, Liu K, et al. ChatGPT-4o outperforms Gemini Advanced in assisting multidisciplinary decision-making for advanced gastric cancer. Eur J Surg Oncol. 2025;51(2):110096.
[19] Yuanyuan Z, et al. A scoping review of self-supervised representation learning for clinical decision making using EHR categorical data. NPJ Digit Med. 2025;8(1).
[20] Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. 2021;4(1):65. https://doi.org/10.1038/s41746-021-00438-z.
[21] Cherla A, et al. Communication of uncertainties about recent cancer drugs in large language models. J Clin Oncol. 2025;43(16_suppl):11026.
[22] Fong K, Rosenthal A, Giroux D, et al. Corrigendum to 'The International Association for the Study of Lung Cancer Staging Project for Lung Cancer: Proposals for the Revision of the M Descriptors in the Forthcoming Ninth Edition of the TNM Classification for Lung Cancer'. J Thorac Oncol. 2025;20:531.
[23] Gaber F, Shaik M, Allega F, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. NPJ Digit Med. 2025;8(1):263.
[24] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-933.
[25] Jeon H, Wang S, Song J, et al. Update 2025: Management of Non-Small-Cell Lung Cancer[J]. Lung, 2025, 203(1): 53.
[26] Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology.

2018;286(3):800-809.
[27] Ma R, Cheng Q, Yao J, et al. Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses. NPJ Digit Med. 2025;8(1):301.
[28] Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. Ann Intern Med. 2024;177(2):210-220.
[29] Yang R, Vinod Nair S, Ke Y, et al. Disparities in clinical studies of AI enabled applications from a global perspective[J]. NPJ Digital Medicine, 2024, 7(1): 209.
[30] Sokol K, et al. Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap. NPJ Digit Med. 2025;8(1):345.
[31] Shen T, Li Y, Cao Y, et al. Rapid deployment of large language model DeepSeek in Chinese hospitals demands a regulatory response. Lancet Digit Health. 2025;31(10):3233-3238.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ZhaoSupplementaryMaterials.docx